Language Agents: Foundations, Prospects, and Risks

Yu Su[1] **Diyi Yang**[2] **Shunyu Yao**[3] **Tao Yu**[4]
[1]The Ohio State University, [2]Stanford University, [3]Princeton University, [4]University of Hong Kong
su.809@osu.edu, diyiy@cs.stanford.edu, shunyuy@princeton.edu, tyu@cs.hku.hk

1 Introduction

A heated discussion thread in AI and NLP is *autonomous agents*, usually powered by large language models (LLMs), that can follow language instructions to carry out diverse and complex tasks in real-world or simulated environments. There are numerous proof-of-concept efforts on such agents recently, including ChatGPT Plugins,[1] AutoGPT,[2] generative agents (Park et al., 2023), just to name a few. The public is also showing an unprecedentedly high level of excitement. For example, AutoGPT has received 147K stars in just 4 months, making it the fastest growing repository in the Github history, despite its experimental nature with many known and sometimes serious limitations.

However, the concept of agent has been introduced into AI since its dawn. So what has changed recently? We argue that the most fundamental change is the capability of using language. Contemporary AI agents *use language as a vehicle for both thought and communication*, a trait that was unique to humans. This dramatically expands the breadth and depth of the problems these agents can possibly tackle, autonomously. The capability of using language, bestowed by their LLM foundations, allows these agents to 1) use a wide range of tools and reconcile their heterogeneous syntax and semantics (Parisi et al., 2022; Schick et al., 2023; Qin et al., 2023a; Patil et al., 2023; Qin et al., 2023b; Mialon et al., 2023), 2) operate in complex environments and ground to environment-specific semantics (Brohan et al., 2023b; Yao et al., 2022a; Gu et al., 2023; Wang et al., 2023a; Deng et al., 2023; Zhou et al., 2023), 3) conduct complex language-driven reasoning (Wei et al., 2022; Shinn et al., 2023; Chen et al., 2023), and 4) form spontaneous multi-agent systems (Park et al., 2023; Liu et al., 2023b). Therefore, to distinguish from the earlier AI agents, we suggest that these AI agents capable of using language for thought and communication should be called "*language agents*," for language being their most salient trait.

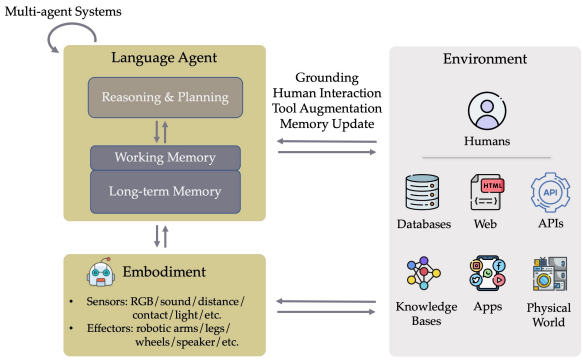

Figure 1: A conceptual framework for language agents.

Language played a critical role in the evolution of biological intelligence, and now artificial intelligence may be following a similar evolutionary path. This is remarkable and concerning at the same time. Despite the rapid progress, there has been a significant lack of systematic discussions regarding the conceptual definition, theoretical foundation, promising directions, and risks associated with language agents. This proposed tutorial endeavors to fill this gap by giving a comprehensive account of language agents based on both contemporary and classic AI research while drawing connections to cognitive science, neuroscience, and linguistics when appropriate.

2 Outline of Tutorial Content

This **cutting-edge** tutorial will be **half-day** and cover a conceptual framework for language agents as well as important topic areas including tool augmentation, grounding, reasoning and planning, multi-agent systems, and risks and societal impact.

2.1 Overview [30mins]

What are language agents and how they differ from the previous generations of AI agents? We

[1]https://openai.com/blog/chatgpt-plugins
[2]https://github.com/Significant-Gravitas/Auto-GPT

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 17–24
November 12–16, 2024.

will start by discussing why the capability of using language for thought and communication empowered by LLMs is the defining trait of the contemporary agents, drawing connections to the role language played in the evolution of biological intelligence (Dennett, 2013). We will then discuss a potential conceptual framework for language agents (Figure 1) and how each component (agent/embodiment/environment) differs from previous agents. One foundational construct is *memory*. We will discuss the resemblances and differences between a language agent/LLM's memory and human memory, including the storage mechanism (Kandel, 2007), long-term memory (LLM's parametric memory/vector databases), and working memory (in-context learning), and how such memory may support general-purpose language-driven reasoning. We will wrap up this section by outlining the key technical and societal aspects that will be discussed in the rest of the tutorial.

2.2 Tool Augmentation [30mins]

Tool augmentation or tool use (Schick et al., 2023; Mialon et al., 2023) is a natural extension of language agents due to their capability of using language for thought and communication. Language agents start to demonstrate a possibility of autonomously understanding and reconciling the heterogeneous syntax and semantics (e.g., XML vs. JSON) of different tools (i.e., using language for communication), and orchestrating the tool execution results into a coherent reasoning process (i.e., using language for thought). At present, tool augmentation mainly serves three purposes:

- Provide up-to-date and/or domain-specific information (Nakano et al., 2021; Lazaridou et al., 2022; Guu et al., 2020).
- Provide specialized capabilities (e.g., high-precision calculation) that a language agent may not have or be best at (Schick et al., 2023; Shen et al., 2023; Cheng et al., 2023; Gao et al., 2022).
- Enable a language agent to act in external environments (Liang et al., 2022; Wang et al., 2023a).

Two metrics are essential for practical tool augmentation: robustness, i.e., accuracy in using tools, and flexibility, i.e., ease of integrating a new tool. While existing efforts, e.g., ChatGPT Plugins, have made meaningful progress on flexibility, robustness still presents a significant challenge. This is particularly problematic for tools that produce side effects in the world (e.g., a tool for sending emails). We will discuss the challenges and opportunities around tool augmentation.

2.3 Grounding [30mins]

Most of the transformative applications of language agents involve connecting an agent to some real-world environments (e.g., through tools or embodiment), be it databases (Cheng et al., 2023), knowledge bases (Gu et al., 2023), the web (Deng et al., 2023; Zhou et al., 2023), or the physical world (Brohan et al., 2023a). Each environment is a unique context that provides possibly different interpretations of natural language. Grounding, i.e., the linking of (natural language) concepts to contexts (Chandu et al., 2021), thus becomes a central and pervasive challenge. There are two types of grounding related to language agents:

- Grounding natural language to an environment (Gu et al., 2023). This is also closely related to the *meaning* of natural language, which, as Bender and Koller (2020) put it, is the mapping from an utterance to its *communicative intent*.
- Grounding an agent's decisions in its own context (i.e., working memory), which includes external information from tools (Liu et al., 2023a; Yue et al., 2023; Gao et al., 2023; Cheng et al., 2023).

We will discuss the current work on both types of grounding, the remaining challenges, and promising future directions.

2.4 Reasoning and Planning [30mins]

The simplest way for language agents to interact with external worlds is to generate the next action via the LLM (Nakano et al., 2021; Schick et al., 2023), but the mapping from context to action is often non-trivial and such approaches often require fine-tuning to learn the mapping. Inspired by prior work that leverages intermediate reasoning to improve LLM performance (Nye et al., 2021; Wei et al., 2022), approaches such as ReAct (Yao et al., 2022b) start to leverage intermediate reasoning for better acting by flexibly analyzing environmental observations, making plans, tracking task status, recovering from exceptions, etc. Subsequent studies (Shinn et al., 2023; Chen et al., 2023) further leverage LLM reasoning for explicit self evaluation,

critic, or reflection, to further improve agent performance. On the other hand, the simplest way for language agents to plan multiple steps of actions is to generate an action plan (Huang et al., 2022), but the token-by-token autoregressive decoding makes it hard to forecast planned future, backtrack from error, or maintain a global exploration structure for planning. To this end, recent works have begun to enhance LMs with re-planning (Song et al., 2022) or tree search algorithms (Yao et al., 2023; Hao et al., 2023) to systematically explore and make decisions in the planning space, analogous to planning-based agents such as AlphaGo (Silver et al., 2016). We will also discuss the recent trend that blurs the boundary between reasoning and acting, which leads to a more unified methodology between reasoning and planning (e.g., Monte-Carlo tree search applied for both reasoning (Hao et al., 2023) and action planning (Silver et al., 2016)).

2.5 Multi-Agent Systems [30mins]

When AI agents are equipped with the capability of using language for thought and communication, it starts to enable multi-agent systems quite different from the conventional ones (Ferber and Weiss, 1999)—agents can now act and communicate with each other in a more autonomous fashion. On the one hand, agents may now be generated with minimal specification instead of preprogrammed and can continually evolve through use and communication to produce complex social behaviors (Park et al., 2023), collaborate for task solving (Wu et al., 2023; Qian et al., 2023; Hong et al., 2023), or debate for more divergent and faithful reasoning (Chan et al., 2023; Liang et al., 2023; Du et al., 2023). On the other hand, human users are also agents, and these artificial language agents can interact with human agents in much richer and more flexible ways than before. There are numerous emerging opportunities, such as providing guardrails and alignment for language agents (Bai et al., 2022) and resolving uncertainties (Yao et al., 2020). We will discuss the opportunities and challenges in this new generation of multi-agent and human-AI collaborative systems.

2.6 Risks and Societal Impact [30mins]

Despite being powerful in a wide range of tasks, language agents are very likely to suffer from key risks and societal harms (Wang et al., 2023b). The first aspect is towards hallucination. The aforementioned memory module, retrieval, or even tool augmentation can largely increase faithfulness of model output, but hallucination issues might still exist and could lead to misleading, unsecure, and even harmful output especially when it comes to high-stake scenarios, raising key concerns towards privacy and truthfulness of the resulting interaction. Bias and fairness remain another primary risk, as language agents might inherit biases from the training corpus. The simulated AI agents might perpetuate stereotypes or discriminate against certain groups of people (Schramowski et al., 2022). Other potential risks include: the lack of transparency in why AI agents behave in their decision-making process, the robustness in AI agents in terms of being manipulated by malicious actors (Zou et al., 2023), and the ethics in terms of what AI agents can and cannot do, etc. Our tutorial will provide a detailed walkthrough of these potential risks in AI agents (Aher et al., 2023), using a few representative case studies to demonstrate how such risks might affect downstream applications, and how human-in-the-loop (Wu et al., 2022) or mixed initiative agents can be leveraged to build more responsible language agents. More importantly, we will briefly discuss the multifaceted impact of language agents, when it comes to user trust (Hancock et al., 2020; Liu et al., 2022), and cultural and societal implications. We will also discuss efforts on evaluating and benchmarking language agents (Liu et al., 2023c,d).

3 Other Required Information

The proposed tutorial is considered a **cutting-edge** tutorial that gives a systematic account of the emerging topic of language agents. There is no prior tutorial at *CL conferences that has covered this topic. There are a few recent tutorials *covering some related aspects* of language agents, such as "ACL'23: Tutorial on Complex Reasoning over Natural Language" on reasoning, "ACL'23: Retrieval-based Language Models and Applications" on retrieval augmentation, and "EMNLP'23: Mitigating Societal Harms in Large Language Models" on societal considerations of LLMs. However, there lacks a comprehensive coverage on the foundations, prospects, and risks of language agents, a void this proposed tutorial aspires to fill.

3.1 Target Audience and Prerequisites

This tutorial is targeted at a broad audience who are interested in language agents. There are no strict prerequisites for the audience's background, but

having 1) basic knowledge of machine learning and deep learning and 2) basic knowledge of language models will help deeper understanding.

3.2 Diversity and Inclusion

We deeply value diversity and strongly believe it can greatly help realize the tutorial's goal and will ensure diversity in the following aspects:

Diversity of instructors. The instructor team has a diverse background including faculty members and graduate students from four institutes spanning two continents and from different gender groups.

Diversity of participants. Language agents are an emerging multi-disciplinary research topic with a very high level of interests in both academia and industry, so we expect a diverse audience. To further promote the awareness of the tutorial in underrepresented communities, we will work with affinity groups such as Black in AI, WiNLP, and LatinX in AI to broadcast the tutorial as well as solicit suggestions on the tutorial content.

Diversity of topics. Given the multi-disciplinary nature of language agents, the materials of this tutorial will cover both contemporary and classic AI/NLP research as well as related discussions from reinforcement learning, cognitive science, neuroscience, linguistics, human-computer interaction, and social science.

3.3 Tutorial Logistics

Estimated audience size. Based on prior tutorials and workshops we organized on related topics, we expect **100-150 attendees** including researchers and practitioners in related fields.

Open access. All materials will be released online on a dedicated website for the tutorial.

Preferred venue. We prefer to have the tutorial co-located with **ACL 2024** or **EMNLP 2024**.

3.4 Breadth

At least 60% of the tutorial will center around work done by researchers other than the instructors. This tutorial categorizes promising approaches for language agents into several groups, and each of these groups includes a significant amount of other researchers' works.

4 Tutorial Instructors

Yu Su is a distinguished assistant professor of engineering at the Ohio State University. His research investigates the role of language as a vehicle for thought and communication in artificial intelligence. His work at Microsoft has been deployed as the official conversational interface for Microsoft Outlook. His work on language agents has won awards such as Outstanding Paper Award at ACL'23 and COLING'22 and from the Amazon Alexa Prize Challenge. He has given 30+ invited talks internationally. Homepage: https://ysu1989.github.io/.

Diyi Yang is an assistant professor in the Computer Science Department at Stanford University. Her research focuses on human-centered natural language processing and computational social science. Diyi has organized four workshops at NLP conferences: Widening NLP Workshops at NAACL 2018 and ACL 2019, Causal Inference workshop at EMNLP 2021, NLG Evaluation workshop at EMNLP 2021, and Shared Stories and Lessons Learned workshop at EMNLP 2022. She gave a tutorial at ACL 2022 on Learning with Limited Data, and a tutorial at EACL 2023 on Summarizing Conversations at Scale. Homepage: https://cs.stanford.edu/~diyiy/.

Shunyu Yao is a PhD student at Princeton NLP Group, advised by Karthik Narasimhan and supported by Harold W. Dodds Fellowship. His research focuses on various facets of developing language agents, such as reasoning, acting, learning, and benchmarking. Homepage: https://ysymyth.github.io.

Tao Yu is an assistant professor of computer science at The University of Hong Kong. He completed his Ph.D. at Yale University and was a postdoctoral fellow at the University of Washington. His research aims to build language model agents that ground language instructions into code or actions executable in real-world environments. Tao is the recipient of an Amazon Research Award and Google Scholar Research Award. He has co-organized multiple workshops and a tutorial related to language agents at ACL, EMNLP, and NAACL. Homepage: https://taoyds.github.io/.

5 Ethics Statement

Language agents, with the ability of autonomously acting in the real world, pose significant potential ethical and safety risks. A main purpose of this proposed tutorial is to systematically define and analyze the unique capabilities and associated risks of language agents. We have a dedicated section on risks and societal impact, and we also cover related discussion in every other section when appropriate.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023a. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023b. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318. PMLR.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding language models in symbolic languages. *ICLR*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*.

Daniel C Dennett. 2013. The role of language in intelligence. *Sprache und Denken/Language and Thought*, page 42.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Jacques Ferber and Gerhard Weiss. 1999. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-wesley Reading.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *ArXiv*, abs/2211.10435.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Yu Gu, Xiang Deng, and Yu Su. 2023. Don't generate, discriminate: A proposal for grounding language models to real-world environments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4928–4949, Toronto, Canada. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Eric R Kandel. 2007. *In search of memory: The emergence of a new science of mind*. WW Norton & Company.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *ArXiv*.

Jacky Liang, Wenlong Huang, F. Xia, Peng Xu, Karol Hausman, Brian Ichter, Peter R. Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023b. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023c. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will ai console me when i lose my pet? understanding perceptions of ai-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–13.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, et al. 2023d. Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents. *arXiv preprint arXiv:2308.05960*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-Assisted Question-Answering with Human Feedback. *arXiv preprint arXiv:2112.09332*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *ArXiv*, abs/2205.12255.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2023a. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023b. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Yongliang Shen, Kaitao Song, Xu Tan, Dong Sheng Li, Weiming Lu, and Yue Ting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *ArXiv*, abs/2303.17580.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2022. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *arXiv preprint arXiv:2212.04088*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023b. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Ziyu Yao, Yiqi Tang, Wen-tau Yih, Huan Sun, and Yu Su. 2020. An imitation game for learning semantic parsers from user interaction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6883–6902, Online. Association for Computational Linguistics.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Appendix

A Past Tutorials/Workshops by the Instructors

The instructors of the proposed tutorial have given tutorials or co-organized workshops at leading international conferences as follows:

Yu Su:

- ACL'21: Workshop on Natural Language Processing for Programming
- ACL'20: Workshop on Natural Language Interfaces
- WWW'18: Tutorial on Scalable Construction and Querying of Massive Knowledge Bases
- CIKM'17: Tutorial on Construction and Querying of Large-scale Knowledge Bases

Diyi Yang:

- EACL'23: Tutorial on Summarizing Conversations at Scale
- ACL'22: Tutorial on Learning with Limited Data
- EMNLP'21: Workshop on Causal Inference & NLP
- NAACL'18 & ACL'19: Widening NLP Workshop

Tao Yu:

- ACL'23: Tutorial on Complex Reasoning over Natural Language
- NAACL'22: Structured and Unstructured Knowledge Integration Workshop
- EMNLP'20: Interactive and Executable Semantic Parsing Workshop

B Recommended Reading List

The audience is recommended (but not required) to read the following papers before the tutorial to facilitate more engagement during the tutorial:

- Daniel C Dennett. The role of language in intelligence. (Dennett, 2013)

- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. (Schick et al., 2023)

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. (Wei et al., 2022)

- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. (Yao et al., 2022b)

- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. (Aher et al., 2023)

- 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. (Wang et al., 2023b)

- Yu Gu, Xiang Deng, and Yu Su. Don't generate, discriminate: A proposal for grounding language models to real-world environments. (Gu et al., 2023)

- Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Binding language models in symbolic languages. (Cheng et al., 2023)

- Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. (Park et al., 2023)

- Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. (Schramowski et al., 2022)

- Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. (Bender and Koller, 2020)