The Multi-agent System based on LLM for Online Discussions

Doctoral Consortium

Yihan Dong
Kyoto University
Kyoto, Japan
dong.yihan.77v@st.kyoto-u.ac.jp

ABSTRACT

The considerable improvement on the Internet and the corresponding applications leads to the result of online discussions becoming far more popular and significant than any other method for people to communicate with each other and reach a consensus. Meanwhile, the incredible improvement in Large Language Models (LLM) has promoted the performance of LLM-based agents in text understanding and content generation capabilities. The research objective of the PhD thesis is to build democratic discussion environments, with three main issues existing right now: 1) Large-scale discussions tend to be complicated, 2) Rumours and misinformation bring negative effects to the discussions, and 3) Direct democratic discussions are complex and time-consuming. This extended abstract introduces the efforts that have been made to address those issues, with the introduction of the potential directions in the future.

KEYWORDS

Multi-agent System: Human Interaction; Trustworthy AI

ACM Reference Format:
Yihan Dong. 2024. The Multi-agent System based on LLM for Online Discussions: Doctoral Consortium. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

The considerable improvement in the Internet and the corresponding applications leads to the result of discussions online becoming far more popular and significant than other methods for people to communicate with each other. Online discussions are also posed to be the next-generation solution for democratic citizen involvement [12]. For example, Twitter has demonstrated its influence during the American presidential election and COVID-19 periods [15, 18].

Some attempts have been made such as online discussion platforms like D-agree [13], with an automated facilitation agent [12] is also developed to encourage participants to express themselves.

On the other hand, the incredible improvement of the Large Language Model (LLM) has proven that it has high performance in text understanding and generation capabilities [17]. Generally, the LLM-based agents have two interaction paradigms with humans [20]: 1) Instructor-Executor Paradigm, which means that humans give commands to the agents to let them finish tasks, and 2) Equal Partnership Paradigm, which means that agents reach the same level as humans and participate in events equally.

The main research objective is to build democratic discussion environments, and three main issues and obstacles exist for this objective: 1) Large-scale discussions tend to be complicated, due to the various characteristics of participants, 2) Rumours and misinformation bring negative effects to the discussions by influencing the final consensus reached by participants, and 3) Direct democratic discussions are time-consuming and complex, because of the diversity of preferences and positions and participants' ignorance.

Thus, to solve the aforementioned issues and obstacles, three research topics are raised as follows: 1) An automated LLM-based facilitation agent framework to further encourage participants in online discussions, 2) A multi-agent fact-checking framework to measure the credibility level of a piece of text, and 3) A personal agent framework that can represent humans with their preferences and positions in online discussions.

2 RESEARCH RESULTS

2.1 An Automated Multi-phase Facilitation Agent Framework based on LLM

Automated facilitation artificial intelligence (AI) agents have been realized since they can efficiently facilitate large-scale discussions. For example, D-Agree [13] is a large-scale discussion support system where an automated facilitation AI agent facilitates discussion among people. Since the current automated facilitation agent was designed following the structure of the issue-based information system (IBIS) and the IBIS-based agent has been proven to have superior performance [8–12]. In the IBIS structure, every post is classified as issue, idea, pro or con to describe its contents, and those posts should be in strict sequence (issue-idea-pro/con). The IBIS-based facilitation agent identifies the type of posts first, then encourages participants to raise the expected contents following the sequence of IBIS. Thus, it is clear that the IBIS-based agent can only respond to existing posts and ignore the detailed contents of the posts raised by participants.

Based on the description above, in this work, we focused on the following three main problems: 1) The IBIS-based agent was designed to only promote other participants' posts by replying to existing posts accordingly, lacking the consideration of different behaviours taken by participants with diverse characteristics, leading to a result that sometimes the discussion is not sufficient. 2)The facilitation messages generated by the IBIS-based agent were not natural enough, leading to consequences that the participants were not sufficiently promoted and did not follow the flow to discuss a topic. 3)Since responding to participants according to detailed post

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

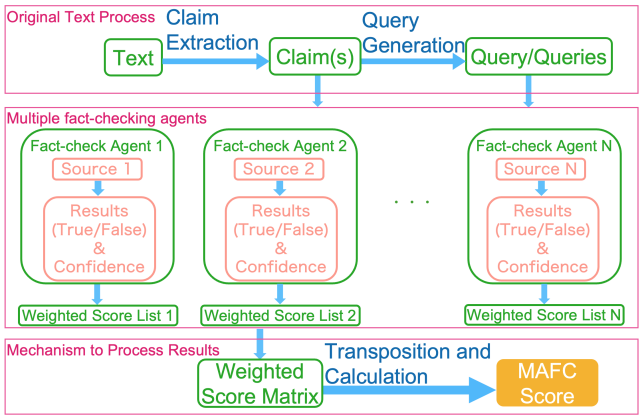

Figure 1: The General Design of Multi-agent Fact-checking Framework

contents could further promote discussions, designing the control of LLM is necessary.

Thus, to address the aforementioned issues and build an inclusive discussion environment [5, 16], an automated multi-phase facilitation agent framework is designed and implemented. Discussions are divided into multiple phases manually to encourage participants to join discussions better under the guidance of the facilitation agent. Multiple groups of discussion experiments were held using the framework, and the naturalness and diversity of the new facilitation agent were far better than the IBIS-based facilitation agent no matter due to human feedback or evaluation using distinct-1 score and PLL score [3].

2.2 Multi-agent Fact-checking based on Large Language Models

Multiple previous research points out that rumours and misinformation spread faster than truth through online social media [7]. Since online discussions can also be affected by rumours and misinformation, fact-checking is significant for online discussion scenarios. The previous machine-learning-based fact-checking methods aimed to detect particular patterns of lies, rumours and misinformation to verify them. However, the patterns of rumours and misinformation also changed in these years, for example, Trump used Twitter to spread misinformation to influence America's election in 2016 and 2020. Thus, multiple related research tried to address the fact-checking problem by covering claims detection, evidence retrieval and claims verification with the use of the large language model (LLM) since the incredible development of LLM [1, 2, 6, 14].

However, three main issues hinder further applications of the fact-checking system in the online discussion field: 1) Most of the fact-checking works are based on a single source which is assumed to be authoritative. 2) The judgement results made by large language models (LLM) with provided information are always considered overconfident. 3) Only the binary label classification task is insufficient.

Therefore, to concur with the obstacles mentioned above, a multi-agent fact-checking framework combined with LLM is proposed to measure the degree of credibility of the text. Specifically, it contains two main parts: 1) multiple fact-checking agents driven by LLM, with independent and unique information sources provided to judge the truthfulness of claims extracted from the original text with the confidence of the judgement result, and 2) a scoring mechanism to convert the judgement results and the confidence to express the credibility of the original text. The general design of the framework is illustrated in Figure 1

Multiple comparative experiments are conducted to measure the performance of the proposed method. By comparing the proposed method with single agents and a multi-agent system with majority rule, the proposed method driven by gpt-3.5-turbo only has a slight performance improvement (8% in average f1-score) in binary fact-checking tasks is revealed. Comparing the proposed method with a single LLM, demonstrated that the proposed method has better performance in multi-label fact-checking especially for the text between pure true and pure false. Finally, since the definition of credibility labels and the scope of different credibility levels have not been clearly defined yet, this work could be a potential contribution to the fact-checking field.

3 DISCUSSIONS AND FUTURE WORK

This extended abstract outlines two frameworks that are the combination and applications of the multi-agent system with agents driven by LLM, and to address the aforementioned research issues, I will focus on the following research topic for the rest of PhD period:

LLM-based Personal Agents that Represent Users in Online Discussions. As mentioned in the first section, direct democratic discussions are time-consuming and complex due to diverse preferences and positions, with the participants' ignorance of the costs and benefits of policies [4].

With the incredible development in the performance of LLM, using LLM to estimate people's political ideologies has been proven feasible [19]. Thus, the reversed method, using LLM to generate messages based on particular political ideologies, should also be feasible. Based on this idea, a personal agent framework representing humans to express their preferences and positions, while raising high-quality claims with the help of LLM, can be designed and implemented.

However, there are two main obstacles to this research topic: 1) How to use LLM-based agents to simulate humans with a proper method, and 2) How to define and monitor the discussions among agents and humans. From my perspective, I will use a four-dimensional political spectrum including social, economic, cultural and international policies to estimate the possible ideas that a specific person may raise or support, to address obstacle 1; For obstacle 2, defining different states in discussions and applying Markov Decision Process (MDP) to LLM-based agents to simulate the consensus-making process is my initial idea.

After solving this particular research topic, the fields of *general multi-agent system driven by LLM*, *LLM-based agent reinforcement learning* and *improving democracy with LLM-based multi-agent system* are what I am interested in right now.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR20D1, Japan and JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123.

REFERENCES

[1] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528* (2023).

[2] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495* (2023).

[3] Yihan DONG, Shiyao DING, and Takayuki ITO. 2023. An Implementation of An Automated Facilitation Agent Promoting Inclusive Discussion. *Proceedings of the Annual Conference of JSAI* JSAI2023 (2023), 1P5OS16b03–1P5OS16b03. https://doi.org/10.11517/pjsai.JSAI2023.0_1P5OS16b03

[4] Anthony Downs. 1960. Why the Government Budget is Too Small in a Democracy. *World Politics* 12, 4 (1960), 541–563. http://www.jstor.org/stable/2009337

[5] Vanessa Grubbs. 2020. Diversity, equity, and inclusion that matter. *New England Journal of Medicine* 383, 4 (2020), e25.

[6] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[7] Manjul Gupta, Denis Dennehy, Carlos M. Parra, Matti Mäntymäki, and Yogesh K Dwivedi. 2023. Fake news believability: The effects of political beliefs and espoused cultural values. *Information Management* 60, 2 (2023), 103745. https://doi.org/10.1016/j.im.2022.103745

[8] Rafik Hadfi, Jawad Haqbeen, Sofia Sahab, and Takayuki Ito. 2021. Argumentative conversational agents for online discussions. *Journal of Systems Science and Systems Engineering* 30 (2021), 450–464.

[9] RAFIK HADFI and TAKAYUKI ITO. [n.d.]. Exploring Interaction Hierarchies in Collaborative Editing using Integrated Information. ([n. d.]).

[10] Jawad Haqbeen, Takayuki Ito, Rafik Hadfi, Tomohiro Nishida, Zoia Sahab, Sofia Sahab, Shafiq Roghmal, and Ramin Amiryar. 2020. Promoting discussion with AI-based facilitation: Urban dialogue with Kabul city. In *Proceedings of the 8th ACM Collective Intelligence, ACM Collective Intelligence Conference Series, Boston (Virtual Conference), South Padre Island, TX, USA*, Vol. 18.

[11] Jawad Haqbeen, Takayuki Ito, Sofia Sahab, Rafik Hadfi, Shun Okuhara, Nasim Saba, Murataza Hofaini, and Umar Baregzai. 2021. A contribution to covid-19 prevention through crowd collaboration using conversational AI & social platforms. *arXiv preprint arXiv:2106.11023* (2021).

[12] Takayuki Ito, Rafik Hadfi, and Shota Suzuki. 2022. An Agent that Facilitates Crowd Discussion: A Crowd Discussion Support System based on an Automated Facilitation Agent. *Group Decision and Negotiation* (2022), 1–27.

[13] Takayuki Ito, Shota Suzuki, Naoko Yamaguchi, Tomohiro Nishida, Kentaro Hiraishi, and Kai Yoshino. 2020. D-agree: crowd discussion support system based on automated facilitation agent. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13614–13615.

[14] Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896* (2023).

[15] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 1–9.

[16] Lynn M Shore, Amy E Randel, Beth G Chung, Michelle A Dean, Karen Holcombe Ehrhart, and Gangaram Singh. 2011. Inclusion and diversity in work groups: A review and model for future research. *Journal of management* 37, 4 (2011), 1262–1289.

[17] Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. *arXiv preprint arXiv:2306.03314* (2023).

[18] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[19] Patrick Y Wu, Joshua A Tucker, Jonathan Nagler, and Solomon Messing. 2023. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057* (2023).

[20] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).