APOLLO: An Open Platform for LLM-based Multi-Agent Interaction Research

Abel Johny[1 [0009−0006−9293−0185]], Eike Schneiders[2 [0000−0002−8372−1684]], and Jeremie Clos[1 [0000−0003−4280−5993]]

[1] University of Nottingham, Nottingham, UK
{psxaj3, jeremie.clos}@nottingham.ac.uk
[2] University of Southampton, Southampton, UK
eike.schneiders@soton.ac.uk

Abstract. Traditional decision-making processes often struggle to capture diverse stakeholder perspectives and anticipate potential outcomes. Complex decisions and persuasions might rely on insights and perspectives which might not be available. In this paper, we leverage recent advances in large language models and retrieval-augmented generation to introduce APOLLO—an Architecture and oPen-source system that Orchestrates Large Language mOdels. APOLLO coordinates multiple LLMs by engaging them in collaborative discourse to reach a consensus on user-defined prompts. This system enables HCI and AI researchers and practitioners, and allows them to explore and experiment with LLM-based multi-agents systems in a user-configurable and customisable manner. By providing this flexible platform, APOLLO enables new avenues for studying and designing human-AI interactions, investigating the impact of multi-agent interaction on human behaviour, and ultimately facilitates a deeper understanding of how AI-driven collaboration can enhance human-AI interaction and decision making.

Keywords: Human-AI Interaction · Multi-Agent Systems · Large Language Models · LLM · Research Software · Decision-Making · Generative AI

1 Introduction

Since the release of ChatGPT in November 2022 the HCI community has demonstrated an interest in investigating the implications on human interactions, behaviours, and reliance using Large Language Models (LLM) [12, 10]. However, as a community, we still lack flexible open-access tools allowing us to investigate multi-agent interaction with LLMs. To address this, we present an Architecture and oPen-source platform that Orchestrates Large Language mOdels (APOLLO). We developed APOLLO to enable researchers and practitioners to customise the behaviour of multiple large language models, and subsequently study user interaction with these systems. This is increasingly relevant for tasks such as AI supported medical [1] and legal [9] decision making, problem solving [7], and persuasion or interaction with opposing viewpoints online [13].

In this paper, we present APOLLO, an Architecture and oPen-source system that Orchestrates Large Language mOdels (LLMs) for autonomous decision making[3]. The APOLLO system provides a platform that enables LLMs to function autonomously, effectively transforming them into autonomous AI agents, capable of pursuing complex user-specified goals with or without human interaction. These agents are orchestrated by APOLLO to engage in collaborative discourse aimed at reaching consensus. This agent-based deliberation process mirrors the complexity of real-world decision-making environments but allows for a more thorough exploration of potential solutions and their implications by providing options to model complex alternative scenarios. With this platform we provide HCI and AI researchers and practitioners the means to investigate interaction with LLM-based multi-agent systems in a user configurable way.

2 The APOLLO System

In this key part of this demonstrator paper, we briefly introduce the Architecture and oPen-source system that Orchestrates Large Language mOdels, in short: the APOLLO system. We designed and implemented APOLLO to facilitate researchers and practitioners in investigating decision-making, problem solving, persuasion strategies, and human-LLM interaction by orchestrating multiple Large Language Models (LLMs) in collaborative scenarios. By enabling multiple AI agents to engage in structured discourse, for instance in an effort to reach consensus or persuade, APOLLO bridges a critical gap between single-agent AI systems and the multi-faceted nature of real-world AI-interactions. The system integrates diverse perspectives and reasoning approaches, similar to human group deliberations[4].

2.1 System Personalisation through Configuration

A key feature of APOLLO is its ability to manipulate the behaviour of the LLMs without the need for technical expertise. APOLLO facilitates this through the configuration of key system behaviours and characteristics. The System Parameter Configuration component is the primary interface for engaging with AI agents in the system. It allows researchers to configure the system. Amongst others, APOLLO users have the option to enter custom textual prompts, define the behaviour of the LLM agents, as well as supplement their knowledge base by uploading domain-specific PDFs and text files. The interaction back-end provides a number of parameters for tuning and personalisation of the interaction environment. Specifically, the parameters can be configured:

1. **System Prompt:** All LLM agents share a common protocol through a centrally maintained system prompt. This shared foundation serves as the "common vocabulary", as described by Allan et al. [3]. The system prompt is exposed to the

[3] APOLLO as open-source on github: https://github.com/abeljohny/apollo
[4] A brief APOLLO demo: https://www.youtube.com/watch?v=TqA7yZXAPBo

APOLLO user to allow personalisation of the agents through customisation of LLM agents behaviour during the discourse.

2. **Maximum Number of Turns:** A turn is defined as a complete cycle of responses from all participating agents. To ensure a balance in representation of all agents [11], we ensured that each agent contributed during each turn. Providing the option to limit the amount of turns prevents indefinite conversations while providing users with a predictable timeframe for discourse completion. Should the LLM agents reach consensus before the user-defined maximum number of turns, the conversation will conclude early.
3. **Harmfulness Detection:** To ensure the safety and appropriateness of LLM responses,a label can be displayed after every response classifying it as *Harmful* or *Harmless*. Harmful responses include a toxicity metric and a percentage indicating their level of toxicity [5].
4. **Domain-Specific Mode (e.g., Lawyer Mode):** APOLLO repurposes LLMs for specific domains by modifying the system prompt. In this mode, agents act as virtual legal strategists, analysing documents and formulating legal strategies. Users can upload court cases to enhance the LLMs knowledge. Agents collaborate to extract relevant details and generate responses, showcasing APOLLO's adaptability[5].
5. **Select Model:** APOLLO detects locally available LLMs and allows users to add models via the Ollama [8] API. Users select models from a drop-down menu; if none is chosen, APOLLO defaults to Gemma 2 (9B) and Llama 3.1 (8B). If unavailable, the system selects two random models or duplicates one.
6. **Agent Behaviour:** APOLLO supports two response formats: *Round-Robin Discussion* and *Summarised Discussion*. In the former, LLMs provide responses based on a sequence. Each response is displayed through a chat-based GUI. In the latter, a concise turn-level summary is generated by the first agent in the sequence, starting from the second turn. Subsequent turns use only these summaries instead of the individual agent responses to continue the discourse.
7. **Only Show Final Consensus:** This option configures APOLLO to present only the final consensus. Each LLM will respond individually, however, these responses will not be shown and only the final consensus will be summarised.

2.2 Auxiliary Utilities

Alongside the core components that define the behaviour of the LLM agents, which can be personalised by the user (see Section 2.1), APOLLO incorporates several auxiliary utilities. These utilities are coordinated by the orchestrator which acts as link between the user interface, the system configuration, and the auxiliary utilities. The system's orchestrator follows a game loop model [4] to simulate the continuous, turn-based interactions among LLMs. It manages critical state variables such as the discussion topic, references to persistence mechanisms (database and file system), and a list of participating LLMs, as well as auxiliary functions such as prompt formatting and context-aware responses via Retrieval-Augmented Generation (RAG) [6, 2].

[5] LLM agents should not be relied upon for legal advice; this context serves as an example of APOLLO's capabilities only.

1. **Retrieval-Augmented Generation (RAG):** The open source Haystack framework is integrated to allow LLMs to query external knowledge sources using natural language. This utility extends the model's memory limits, allowing the agents to work with information from external documents without requiring the entire text to fit within their limited context capacity. The RAG subsystem thus serves as a bridge between the extensive knowledge contained in uploaded documents and the processing capabilities of lightweight LLMs.
2. **Harmfulness Classifier:** To ensure the safety and appropriateness of agent responses, we integrated the Detoxify classifier [5] to identify potentially harmful content in text. Each agent's response is analysed across multiple toxicity categories, including insults, threats, and sexually explicit comments, with toxicity thresholds set to a default value (50%) for each category. This threshold is customisable. When an agent's response exceeds the threshold in any category, the system flags the response as potentially harmful and displays to the user both the specific toxicity metric that triggered the flag and the percentage at which the classifier assessed the harmfulness.
3. **Persistence (Database & File System):** Upon completion of the discussion, when all LLM agents have reached a consensus, the entire conversation chain is maintained in a database, preserving the reasoning chain and ensuring accountability and traceability.

2.3 Example Cases for Research with APOLLO

1. **Case 1: Can LLM-based multi-agent interaction persuade to healthier life choices?** Imagine a smoker considering quitting but lacking a supportive network to reinforce their decision. The APOLLO system allows researchers to explore whether LLM-based multi-agent systems can persuade users to make healthier choices and provide the support needed to follow through on their decision to quit.
2. **Case 2: Nudge towards Friendlier Online Discourse.** Imagine an online forum where discussions frequently become hostile, discouraging constructive dialogue. The APOLLO system allows researchers to explore whether LLM-based multi-agent systems can guide conversations toward more positive and respectful interactions, fostering a friendlier and more inclusive online environment.

3 Conclusion

In this demonstration paper, we introduced APOLLO, a configurable multi-agents system which enables HCI and AI researchers and practitioners to manipulate the behaviour of multiple large language models, and subsequently study user interaction. The APOLLO system thereby facilitates future research investigating human-AI interaction within various use cases such as, AI-supported decision making, persuasions using agent based discussions, and problem-solving. The system overcomes the limitations of a model's memory limits by allowing it to work with information from external documents without requiring the entire text to fit within its limited context capacity.

Acknowledgement. *This project was supported by the Engineering and Physical Sciences Research Council Responsible AI UK [grant number EP/Y009800/1].*

References

1. Barabucci, G., Shia, V., Chu, E., Harack, B., Laskowski, K., Fu, N.: Combining multiple large language models improves diagnostic accuracy. NEJM AI **1**(11), AIcs2400502 (2024)
2. Chang, C.C., Chang, H.P., Lee, H.S.: Leveraging retrieval-augmented generation for culturally inclusive hakka chatbots: Design insights and user perceptions. In: 2024 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE). pp. 1–6 (2024). https://doi.org/10.1109/RASSE64357.2024.10773731
3. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepel, T.: Cooperative AI: machines must learn to find common ground. Nature **593**(7857), 33–36 (May 2021). https://doi.org/10.1038/d41586-021-01170-
4. Gregory, J.: Game Engine Architecture, Second Edition. A. K. Peters, Ltd. (2014)
5. Hanu, L., Unitary, t.: Detoxify (Nov 2020), https://github.com/unitaryai/detoxify
6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
7. Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B.: Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems **36**, 51991–52008 (2023)
8. Morgan, J., Chiang, M.: Ollama (2023), https://ollama.com/
9. Schneiders, E., Seabrooke, T., Krook, J., Hyde, R., Leesakul, N., Clos, J., Fischer, J.E.: Objection overruled! lay people can distinguish large language models from lawyers, but still favour advice from an llm. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25, ACM, New York, NY, USA (2025). https://doi.org/10.1145/3706598.3713470
10. Seabrooke, T., Schneiders, E., Dowthwaite, L., Krook, J., Leesakul, N., Clos, J., Maior, H., Fischer, J.: A survey of lay people's willingness to generate legal advice using large language models (llms). In: Proceedings of the Second International Symposium on Trustworthy Autonomous Systems. TAS '24, ACM, New York, NY, USA (2024). https://doi.org/10.1145/3686038.3686043
11. Tennent, H., Shen, S., Jung, M.: Micbot: A peripheral robotic object to shape conversational dynamics and team performance. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 133–142 (2019). https://doi.org/10.1109/HRI.2019.8673013
12. Wang, C., Hasler, S., Tanneberg, D., Ocker, F., Joublin, F., Ceravola, A., Deigmoeller, J., Gienger, M.: Lami: Large language models for multi-modal human-robot interaction. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. CHI EA '24, ACM, New York, NY, USA (2024). https://doi.org/10.1145/3613905.3651029
13. Zhang, Y., Sun, J., Feng, L., Yao, C., Fan, M., Zhang, L., Wang, Q., Geng, X., Rui, Y.: See widely, think wisely: Toward designing a generative multi-agent system to burst filter bubbles. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, ACM, New York, NY, USA (2024). https://doi.org/10.1145/3613904.3642545