Arguing AIs: A Proposal

Kieran M. Dziallo KDZIALL1@JH.EDU
Johns Hopkins University, Baltimore, MD

Peter Rasmussen PRASMU3@JH.EDU
Johns Hopkins University, Balitmore, MD

Adam Walker LWALKE66@JH.EDU
Johns Hopkins University, Baltimore, MD

1. Introduction & Background

Recent work in the field of Large Language Models (LLMs) have been in the direction of using multiple agents to debate. Our proposal rests on similar research that has been conducted recently that uses multiple agents to debate and improve response quality. Ultimately, our goal is to demonstrate that a system made up of multiple agents can make persuasive arguments.

Liang et al. (2023) proposed the Multi-Agent Debate (MAD) framework to address the problem of LLMssuffering from the Degeneracy-of-Thought (DoT) problem.[1] Through this, multiple agents express their arguments in the state of a "tit for tat" with a "judge" managing the interactions to finally arrive a "final solution" and allowing for more deep thought and "contemplation." The MAD framework consists of N debaters and a single judge. The debaters are instantiated by using a meta-prompt to introduce the topic at hand and other requirements, such as number of debaters, rounds, and more. The judge moderates and can determine when a valid solution is found. Similarly, Du et al. (2023) used debate among agents to improve LLM responses and to reduce factual inaccuracies. Specifically, it was found that mathematical and strategic reasoning improved when agents were placed into a debate setting because the response of the first agent would inform the second, and so on, to iteratively improve response quality (Du et al., 2023).

Another debate framework that has been proposed is the Formal Debate Framework (FORD), which features a "three-stage debate aligned with real-world scenarios: fair debate, mismatched debate, and roundtable debate" (Xiong et al., 2023). Xiong et al. (2023) found that a strong LLM agent tended to dominate a debate by adhering to their initial views, whereas an agent which changed its point of view tended to have weaker performance, and due to this, it was determined that a "competent judge" was necessary for moderation. Integral to debate performance was having different personas through diverse role prompts. If using the same role description in the initial prompt, there is likely to be a degradation in performance (Chan et al., 2023).

1. Although not central to the proposal, this problem was the impetus of the MAD Framework. DoT is the phenomenon of when an LLM cannot create novel thoughts once grounded in an initial, erroneous idea" (Liang et al., 2023).

2. Methods

Due to most publicly available LLMs having guard-rails to disallow polarized political discourse, the decision was made to use historical debates to demonstrate our methods. Three debates were chosen: Roman tyranny, Greek democracy, and American independence. All debates will be moderated by a judge agent.

Utilizing OpenAI's ChatGPT and the AgentVerse framework for scripting multiple agents, we propose creating a debate framework where two personas debate a given issue while being moderated by a "judge" agent that will determine which case is stronger. In addition, we propose using a live, human control group to vote on the persuasive abilities of the debating agents. In so doing, we look to compare the decisions of the judge against the consensus decision of a human group.

2.1 Personas

Based on the work of Chan et al. (2023), it was determined that strong personas would need to be used to zealously advocate for the different views. To provide historical context and draw on the vast library of ingested information, each side of these debates will take the persona of a factual historical figure.

2.1.1 Roman Tyranny

Roman Republicanism is represented by Marcus Porcius Cato. Cato was a Roman politician and staunch opponent of Caesar. He championed traditional Roman values, including and especially representative government (Astin, 1978). Through his career, Cato defended the authority of the Roman Senate against Caesar and his popular policies because he felt that these reforms were destabilizing the bedrock of Roman society and political order (Chisholm, 1911). When Caesar triumphed in the Roman Civil War, Cato killed himself rather than submit to Caesar's autocracy. In general, Cato held staunch opposition to the very idea of autocracy and tyranny, no matter how well intentioned (Chisholm, 1911; Astin, 1978).

On the other side, Roman Imperialism is represented by Julius Caesar. A patrician by birth, Caesar became a Roman general and statesman. Through his military career, Caesar had several successes, including the conquest of Gaul and the invasion of Britain leading to gaining power and prestige within the Roman Republic. He began to consolidate power in the First Triumvirate, but this alliance broke down. At this point, Caesar crossed the Rubicon and began the Roman Civil War. Caesar emerged victorious and declared himself dictator. He led many political reforms which allowed him to consolidate more power to become dictator for life, the Emperor (Morstein-Marx, 2021).

2.1.2 Greek Democracy

Ephialtes, the persona representing direct democracy, was an ancient Athenian politician who is primarily known for his role in the political reforms of Athens during the early stages of the Peloponnesian War. Ephialtes was a proponent of radical democratic changes and sought to reduce the power of the aristocracy and strengthen the democratic system by shifting more power towards the Assembly, which was comprised of all male citizens, regardless of wealth or social status. Ephialtes initiated measures that paved the way for

a more direct form of democracy. This included reducing the powers of the Council of the Areopagus, thereby enabling the Assembly to have a greater say in the decision-making process. Ephialtes' reforms ultimately contributed to the development of a more inclusive and egalitarian democratic system in Athens (Wallace, 1974).

Plato, who is championning the anti-democratic ideals, was an ancient Greek philosopher and student of Socrates. He founded the Academy in Athens, one of the earliest known institutions of higher learning in the Western world. One of his most notable ideas was the concept of the Philosopher King, which is in direct opposition to the prevailing Athenian belief in democracy. Plato believed that the ideal state should be ruled by philosopher-kings, individuals with a deep love for wisdom and a profound understanding of truth and justice. He argued that these enlightened rulers, possessing a unique form of knowledge, would make decisions based on reason and the pursuit of the highest good, rather than personal gain or popular opinion, thus creating a just and harmonious society. Conversely, Plato believed that direct democracy was a dangerous form of government in which the basest of personalities would triumph (Field, 1969).

2.1.3 American Independence

Representing American patriots is John Adams, one of the Founding Fathers of the United States of America and an early advocate for American independence from Britain. As a delegate to both Continental Congresses, Adams was a prime mover for the Declaration of Independence in 1776. He believed strongly that the American colonies should break free of Britain's unfair taxation policies and other restrictions on their liberty, and he felt the King and Parliament were tyrannical and that the colonies deserved their own self-government. Adams was also instrumental in persuading Congress to declare independence and assisted Thomas Jefferson in drafting the Declaration of Independence (McCullough, 2008).

William Franklin is the Loyalist champion. Franklin was the son of American Founding Father Benjamin Franklin and the governor of New Jersey after being appointed by King George III. As tensions rose between Britain and the American colonies, Franklin opposed independence and defended the authority of the crown. He believed Britain had the right to tax the colonies and thought American rights could be protected while remaining part of the Empire. Franklin tried to keep New Jersey neutral in the Revolutionary War but ultimately was arrested for supporting Loyalists. Even after the war, Franklin continued advocating the Loyalist position while living in Britain (Skemp, 1998).

3. Conclusion

Overall, this proposal is to look into and develop a system whereby multiple agents debate while being moderated by a third. This will involve prompt engineering and using historical personas to enable such a debate. In the end, we hope to evaluate the judge's rulings based on a consensus of a human group.

References

Astin, A. E. (1978). *Cato the censor*. Oxford University Press.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2023). Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Chisholm, H. (1911). Marcus porcius cato. In *Encyclopedia Britannica* (11 edition)., Vol. 5, pp. 535–536.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Field, G. C. (1969). *Philosophy of Plato* (2nd edition). Opus Books. Oxford University Press, London, England.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., & Shi, S. (2023). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

McCullough, D. (2008). *John Adams*. Simon & Schuster, New York, NY.

Morstein-Marx, R. (2021). *Julius Caesar and the Roman People*. Cambridge University Press.

Skemp, S. L. (1998). Benjamin franklin, patriot, and william franklin, loyalist. *Pennsylvania History: A Journal of Mid-Atlantic Studies*, *65*(1), 35–45.

Wallace, R. W. (1974). Ephialtes and the areopagos. *Greek, Roman, and Byzantine Studies*, *15*(3), 259–269.

Xiong, K., Ding, X., Cao, Y., Liu, T., & Qin, B. (2023). Examining the inter-consistency of large language models: An in-depth analysis via debate. *arXiv e-prints*, arXiv–2305.