# Evaluating Decision-Making Generalization in RAG Agent Architectures

Mehar Shienh
*University of Waterloo*
msshienh@uwaterloo.ca

Evan Dennison
*University of Waterloo*
edennison@uwaterloo.ca

Jordan Leis
*University of Waterloo*
j2leis@uwaterloo.ca

Devon Kisob
*University of Waterloo*
dkisob@uwaterloo.ca

Jennifer Yu
*University of Waterloo*
j545yu@uwaterloo.ca

Yalda Nikookar
*University of Waterloo*
ynikooka@uwaterloo.ca

Madhav Malhotra
*University of Waterloo*
madhav.malhotra@uwaterloo.ca

***Abstract*—This paper explores LLMs as generalized decision-making assistants. We propose an assessment framework where retrieval-augmented generation (RAG) architectures are compared in simulated environments. By comparing objective win rates in games like Monopoly and Werewolf, we assess the efficacy of architectural options like reflection or multi-agent roles. This allows us to then apply the best performing architectures to the real-life context of political analysis. With this method, we find that the RAG architectures explored do not show generalization across decision-making contexts.**

## I. Introduction

Political decision-making is inherently complex, requiring the ability to navigate conflicting interests, ethical considerations, and long-term policy consequences. Unlike structured tasks such as Go, where AI has achieved superhuman performance through reinforcement learning [1], political decisions involve subjective judgments. Research has shown that AI struggles with strategic reasoning in multi-agent settings where human behaviour is unpredictable, as seen in attempts to apply AI to judicial decisions [2]. Moreover, political decision-making is constrained by legal frameworks and ethical concerns, making it difficult to define optimal strategies solely through data-driven approaches [3].

Despite advancements in AI applications for law, most existing models focus on legal text analysis, compliance automation, and case law retrieval rather than autonomous decision-making [4]. To develop AI capable of making informed political choices, a training environment must simulate the strategic negotiation and decision-making pressures inherent in politics. Simulated environments like Monopoly [5] and Werewolf [6] have been used in behavioural studies to model economic and social decision-making, making them useful for training AI in competitive and cooperative strategies. By engaging in these controlled simulations, AI agents can develop decision-making frameworks that incorporate long-term strategy, weighing uncertainties, and adaptability, all of which are key skills necessary for legislative reasoning.

Still, the generalization of learned strategies from games to real-world contexts remains challenging. Research in transfer learning has demonstrated that AI systems often struggle to apply strategies across domains with different structures and reward functions [7]. Political decisions rarely have objective 'win conditions' like games do, complicating the transfer of game-derived strategies to legislative contexts. However, recent advances in meta-learning approaches have shown promising results in enabling AI to adapt learned strategies to novel tasks with limited additional training [8].

LLMs also present unique advantages for this generalization challenge. Unlike traditional reinforcement learning systems, LLMs trained on diverse corpora already possess broad knowledge about political systems, historical precedents, and ethical frameworks [9]. This background knowledge potentially enables them to contextualize strategies learned in simulated environments within appropriate political frameworks. Studies examining zero-shot and few-shot learning capabilities of LLMs suggest they can rapidly adapt to new decision contexts with minimal domain-specific examples [10]. This raises the question of whether LLMs can be effective decision-making assistants across generalized environments, from structured games to unstructured political analysis.

## II. Related Works

Recent advancements in LLM-driven agent-based modelling have demonstrated the potential for simulating complex decision-making systems across social, economic, and legal domains. Prior research has explored the use of LLMs as autonomous agents, capable of interacting with dynamic environments, learning from experience, and optimizing decision strategies. For instance, several studies have explored the use of LLMs in economic simulations. [11] studied LLM-driven economic forecasting, demonstrating that GPT-4-based agents could simulate macroeconomic trends and follow real-world principles like the Phillips Curve. However, the study noted that LLM agents struggled with long-term reasoning capabilities. Likewise, [5] applied LLMs to negotiation games, noting their tendencies to not make optimal decisions from a game theoretic perspective, Still, structured prompting techniques showed improvements in rational decision-making strategy.

Furthermore, many works show the promise of improving decision-making through multi-agent systems. [12] applies multi-agent simulations to examine how LLMs can simulate social media discourse on contentious topics like nuclear energy policy and gender discrimination. Its findings highlight how LLMs can replicate real-world sentiments, but also risk amplifying biases and polarization. Similarly, [13] developed COLA, a multi-agent stance detection system, where agents acted as linguistic, domain-specific, and social media experts to analyze public discourse. Works like [14] applied adversarial multi-agent legal reasoning, while [13] showed that structured debates lead to more robust decision-making. Similarly, [15] used two debating LLM agents to generate and refine arguments in cooperative problem-solving scenarios. More generally, [16] proposes a hierarchical language architecture, finding improved decision-making via delegating complex decisions to multiple sub-agents. These studies show that multi-agent systems reduce logical inconsistencies and lead to more structured decision-making than single-agent approaches.

Lastly, studies also explore the use of vector stores and other memory implementations to improve the performance of decision-making systems over time. [17] introduced self-reflection prompting, enabling models to review past decisions and self-correct over time, improving logical consistency and strategy formation. [18] carries out ablation studies to assess the importance of memory-based and reflection-based RAG agents in logical coherence. It finds that both components play a significant role in improving subjective impressions of believability and coherence. As seen, diverse memory implementations can contribute to decision-making assistants in subjective and objective environments.

Building on these findings, we evaluate multiple components of RAG agents like multi-agent systems, memories, and reflection. We evaluate these agents in objective and subjective environments, comparing win rates in games of Monopoly and Werewolf before applying the agents to generate arguments for a political bill. This bridges the gap between abstract game mechanics and political decision-making. By sequentially increasing the complexity and realism of these simulations, we investigate whether strategic reasoning skills transfer effectively from game environments to political contexts.

## III. Methodology

In brief, we compare four RAG architectures by their win rates in two structured game environments: Monopoly and Werewolf. For illustration, the two best-performing architectures then analyze legislative texts, demonstrating subjective impressions on the quality of the analysis produced. This applies the architectures to increasingly complex decision-making contexts; Monopoly involves independent decision-making with structured rules. Werewolf introduces conversational decision-making with structured rules. Political legislative analysis presents the most complex scenario with no structured rules to guide decision-making.

The four RAG architectures investigate combinations of two agent environments with two memory approaches. Each architecture runs 50 games of Monopoly and 50 games of Werewolf to determine average win rates. A random seed was used to ensure all games involved different initial conditions where necessary, like when determining dice rolls in Monopoly. Meta-prompting techniques, which have been found to induce more reasoned responses in multi-agent systems, standardize prompts across games where possible to promote a fair comparison [19]. This factorial design systematically evaluates the contribution of each component to decision-making performance.

TABLE I: Architectures Compared

| Architecture | Agent roles | Memory Contains |
|---|---|---|
| Courtroom; Raw | Two lawyers; judge | Past raw outputs |
| Courtroom; Reflection | Two lawyers; judge | Past output summaries |
| Advisory; Raw | Fast mind; slow mind | Past raw outputs |
| Advisory; Reflection | Fast mind; slow mind | Past output summaries |

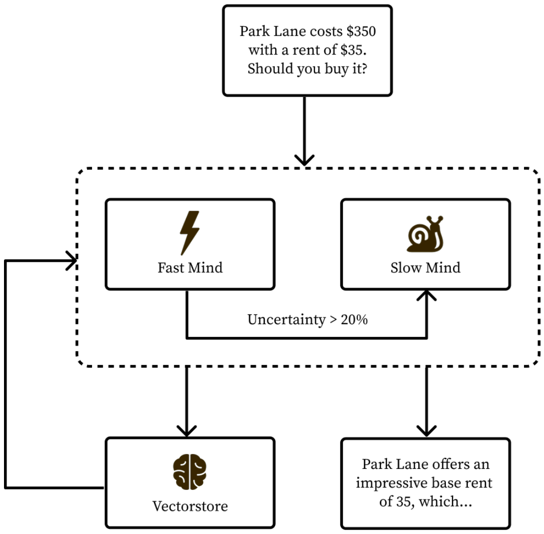
Park Lane costs $350 with a rent of $35. Should you buy it?
Fast Mind
Slow Mind
Uncertainty > 20%
Vectorstore
Park Lane offers an impressive base rent of 35, which...

Fig. 1: Courtroom Architecture

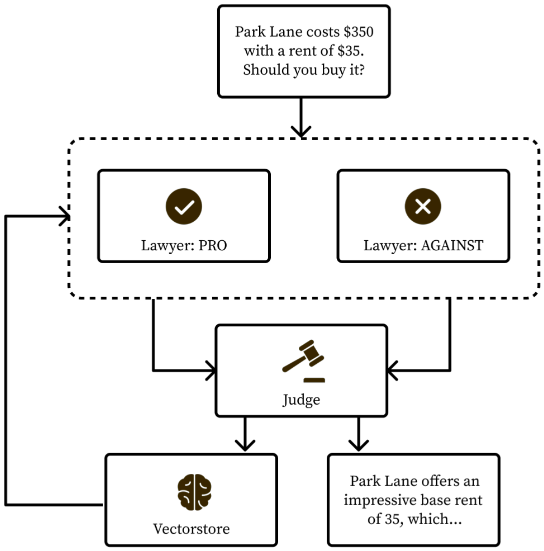
Park Lane costs $350 with a rent of $35. Should you buy it?
Lawyer: PRO
Lawyer: AGAINST
Judge
Vectorstore
Park Lane offers an impressive base rent of 35, which...

Fig. 2: Advisory Architecture

The courtroom multi-agent environment has two lawyer agents argue for different decisions, with a judge agent determining which argument is stronger. All agents use GPT-4o Mini. The second multi-agent environment follows an advisory model where a small, fast model (GPT-4o Mini; 'fast mind') makes most decisions and outputs an uncertainty score. This agent falls back to consulting a slower, larger model (GPT-4o; 'slow mind') when uncertainty exceeds a predefined threshold. Detailed prompts and hyperparameters are available in the supplementary material.

The raw memory vector store simply records past decisions and retrieves the two most similar decisions as examples for current decision-making. The reflective memory vector store takes a more sophisticated approach by storing and retrieving reflective summaries of multiple past decisions. Summaries are generated dynamically throughout the games by GPT-4o after a given number of turns. The embedding model used is Ada 02. Hyperparameters and prompts are detailed in the supplementary materials.

In Monopoly simulations, the custom agent implementing our architectures plays against a default player that follows a simple strategy of buying property whenever funds are available. Games conclude when one player depletes their funds or after 200 turns. The custom agent wins if its combined cash and mortgageable property value exceeds that of the default player. For Werewolf scenarios, the custom agent assumes the role of the werewolf and competes against default chatbots emulating the seer, witch, and villager roles. Detailed configurations for these default agents are provided in the supplementary materials. Games continue until either the werewolf is eliminated through voting or is the sole remaining player.

To evaluate performance in real-world contexts, we selected a random bill from the first session of the 44th Canadian Parliament for the subjective legislative analysis. Bill data was webscraped from openparliament.ca [20] and is available in supplementary materials. The full text of each bill was divided into 500-character chunks. The two best-performing RAG architectures analyzed each chunk's implications on the overall decision to support or oppose the bill, mimicking the turn-based structure of the games. After analyzing all chunks, each architecture produced a final argument either supporting or opposing the bill.

These arguments were presented to 53 survey participants from a convenience sample. Participants ranked which argument they found more structured, balanced, compelling, and professional. The complete set of survey questions is available in the supplementary materials.

## IV. Results

We ran 50 games for each architecture, totalling 200 games of Monopoly and 200 games of Werewolf. This had a cost of approximately $50 CAD in API credits, including testing runs before carrying out the final experiments. The average win rates from these are presented below.

TABLE II: Win Rates From Monopoly

| Architecture | Memory | Win Rate |
|---|---|---|
| Courtroom | Raw | 32 % |
| Courtroom | Reflection | 26 % |
| Advisory | Raw | 46 % |
| Advisory | Reflection | 48 % |

TABLE III: Win Rates From Werewolf

| Architecture | Memory | Win Rate |
|---|---|---|
| Courtroom | Raw | 18 % |
| Courtroom | Reflection | 20 % |
| Advisory | Raw | 20 % |
| Advisory | Reflection | 20 % |

The Advisory architectures outperformed the Courtroom counterparts, so we used them to provide arguments on Bill C-242. 53 undergraduate students at the University of Waterloo were surveyed on their subjective preferences regarding the results. Participants were asked to rank the architectures across four question categories:

1) Structure: "Which response did the best in presenting a structured argument for or against the bill?"
2) Balance: "Which response provides the most balanced discussion of multiple perspectives?"
3) Persuasion: "Which response is the most compelling? Select the one that would be most likely to sway your opinion."
4) Decorum: "Which response is the most appropriate for parliament? Select the response that maintains the best formal and professional tone."

These aggregated preferences are reported below.

TABLE IV: Survey Rankings

| Architecture | Question Category & Preference | | | |
|---|---|---|---|---|
| | Structure | Balance | Persuasion | Decorum |
| Advisory: Raw | 53% | 38% | 49% | 51% |
| Advisory: Reflection | 47% | 62% | 51% | 49% |

## V. Discussion

### A. Limitations

There are numerous limitations in our findings. First, we consider flaws in our experiments in the objective decision-making environments. Assessing two games alone is not sufficient to claim generalization in decision-making ability, but was necessary to control costs in our study. Future work can improve upon this limitation by adding other objective decision-making tasks. For instance, games like Risk or Diplomacy with military or political themes may be viable candidates. In addition, forecasting tasks in political or stock prediction datasets may be suitable. Moreover, comparisons between different games are currently not always balanced. For instance, our default Monopoly player uses hard-coded rules to decide to always buy a property in Monopoly if funds are available, while the default player in Werewolf is a chatbot with non-deterministic behaviour. This decision was made since Werewolf required natural language conversation between agents which made hard-coded default players seem unsuitable. Future work may wish to empirically evaluate these assumptions by using hard-coded or chatbot-based default players for both games. Lastly, there was only one opponent in Monopoly while there were three other default players in Werewolf due to the varying roles in the games. Future work may wish to balance the number of opponents across the games to ensure a similar level of difficulty, though this is likely to increase costs.

Carrying on, we consider flaws in our survey in the subjective decision-making arguments. The largest flaw in our survey is that convenience sampling was used to control study costs. However, this creates a very uniform demographic of participants with a similar background as undergraduates at the same university. With greater funding for marketing and participation compensation, a more diverse sample of participants across demographic strata would be feasible. It would be especially valuable in collecting responses from participants of varying ages, political affiliations, and familiarity with AI technology. In a similar vein, future works could be improved by a greater survey sample size backed by tests for statistical significance.

### *B. Conclusions*

Our results indicate that the Advisory model consistently outperformed its Courtroom counterpart in both Monopoly, and to a lesser extent Werewolf. This outcome implies that a hierarchical structure, which leverages multiple levels of decision-making appears more suited for tasks involving clearly defined strategic goals. While the two-layered approach in Advisory agents may not fully capture all nuances—particularly in conversational environments like Werewolf—it nonetheless demonstrated greater adaptability in managing complex turn-by-turn decisions compared to the adversarial debate format used in Courtroom.

When extended to political scenarios, participants did not exhibit a strong preference for responses generated with raw versus reflective vector stores as preferences within the set categories for the two responses were relatively evenly split. This suggests that summarizing retrieved content may offer limited marginal benefit under these conditions.

Overall, the RAG architectures tested here did not show compelling evidence of generalizable decision-making competence across the diverse settings examined. Our study underscores the importance of domain-aware architectural choices and the need to develop more robust strategies for bridging the gap between controlled simulation environments and real policy discourse. Future work could explore more specialized architectural variants, such as introducing different foundational models or increasing test-time compute to allow sampling over multiple candidate responses. It may also be promising to explore fine-tuned LLMs or LLMs with reinforcement-learning techniques like reward modelling. In all, we present this work as a rough first demonstration to bridge models from objective tasks to subjective tasks, hoping that it may spark future work to improve our methods and experimental results in this interdisciplinary topic.

### *C. Supplementary Materials*

Supplementary materials are available online at: https://github.com/Madhav-Malhotra/political-chatbot

# References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016. [Online]. Available: https://www.nature.com/articles/nature16961

[2] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human Decisions and Machine Predictions*," *The Quarterly Journal of Economics*, Aug. 2017. [Online]. Available: http://academic.oup.com/qje/article/doi/10.1093/qje/qjx032/4095198/Human-Decisions-and-Machine-Predictions

[3] J. J. Bryson, "The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation," in *The Oxford Handbook of Ethics of AI*, M. D. Dubber, F. Pasquale, and S. Das, Eds. Oxford University Press, Jul. 2020, pp. 1–25. [Online]. Available: https://academic.oup.com/edited-volume/34287/chapter/290654580

[4] A. Deroy, K. Ghosh, and S. Ghosh, "Applicability of large language models and generative models for legal case judgement summarization," *Artificial Intelligence and Law*, Jul. 2024. [Online]. Available: https://link.springer.com/10.1007/s10506-024-09411-z

[5] W. Hua, O. Liu, L. Li, A. Amayuelas, J. Chen, L. Jiang, M. Jin, L. Fan, F. Sun, W. Wang, X. Wang, and Y. Zhang, "Game-theoretic LLM: Agent Workflow for Negotiation Games," 2024. [Online]. Available: https://arxiv.org/abs/2411.05990

[6] Y. Xu, S. Wang, P. Li, F. Luo, X. Wang, W. Liu, and Y. Liu, "Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf," 2023. [Online]. Available: https://arxiv.org/abs/2309.04658

[7] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9134370/

[8] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick, "Learning to reinforcement learn," 2016. [Online]. Available: https://arxiv.org/abs/1611.05763

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[10] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent Abilities of Large Language Models," 2022. [Online]. Available: https://arxiv.org/abs/2206.07682

[11] N. Li, C. Gao, M. Li, Y. Li, and Q. Liao, "EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities," 2023. [Online]. Available: https://arxiv.org/abs/2310.10436

[12] C. Gao, X. Lan, Z. Lu, J. Mao, J. Piao, H. Wang, D. Jin, and Y. Li, "S3: Social-network Simulation System with Large Language Model-Empowered Agents," 2023. [Online]. Available: https://arxiv.org/abs/2307.14984

[13] X. Lan, C. Gao, D. Jin, and Y. Li, "Stance Detection with Collaborative Role-Infused LLM-Based Agents," 2023. [Online]. Available: https://arxiv.org/abs/2310.10467

[14] G. Chen, L. Fan, Z. Gong, N. Xie, Z. Li, Z. Liu, C. Li, Q. Qu, S. Ni, and M. Yang, "AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents," 2024. [Online]. Available: https://arxiv.org/abs/2408.08089

[15] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society," 2023. [Online]. Available: https://arxiv.org/abs/2303.17760

[16] J. Liu, C. Yu, J. Gao, Y. Xie, Q. Liao, Y. Wu, and Y. Wang, "LLM-Powered Hierarchical Language Agent for Real-time Human-AI Coordination," 2023. [Online]. Available: https://arxiv.org/abs/2312.15224

[17] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language Agents with Verbal Reinforcement Learning," 2023. [Online]. Available: https://arxiv.org/abs/2303.11366

[18] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," 2023. [Online]. Available: https://arxiv.org/abs/2304.03442

[19] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber, "MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework," 2023. [Online]. Available: https://arxiv.org/abs/2308.00352

[20] M. Mulley, "Open Parliament API." [Online]. Available: https://openparliament.ca/api/