Simulating Market Equilibrium with Large Language Models

Enric Junqué de Fortuny
IESE Business School
ejunque@iese.edu

Abstract

Large Language Models (LLMs) have the potential to simulate complex human decision-making and economic behavior, making them well-suited for training simulations. This study explores LLMs' ability to simulate a market equilibrium game commonly used in MBA classrooms to train future business leaders. We test three simulation architectures: prompt-based, Retrieval Augmented Generation-based, and controller-based. The prompt-based approach struggled with limited context, while the RAG-based system improved information retrieval but occasionally veered off course. To address these challenges, we developed a controller-based simulation integrating multiple LLM agents and custom tools, which enhanced both control and accuracy. Our results show that this approach not only improves the fidelity of economic simulation, but also enriches the learning experience for students by providing more accurate, engaging, and realistic business case environments.

Keywords: LLM, RAG, RPG, controller, education

1. Introduction

In an era where technology increasingly mirrors human cognition, Large Language Models (LLMs) are well on their way to redefining how we simulate and understand complex economic and social behaviors. State-of-the-art LLMs such as Llama, ChatGPT, and Mistral have attained the ability to accurately interpret human language and intentions, generating responses that closely mimic human communication (Hewitt et al., 2024; van Duijn et al., 2023). Beyond text generation, they are now being explored for their potential to simulate complex human behaviors in economic and social contexts (Aher et al., 2023; Horton, 2023; Park et al., 2022). These studies suggest that LLMs, with their embedded social knowledge and computational power, are well-suited for economic simulations. However, most of these simulations treat the LLM as a static entity (the "homo silicus"), responding to a single query without interaction (Horton, 2023). The true potential of LLMs in dynamic, interactive environments — such as those found in business education through case studies — remains largely unexplored. More research is needed to understand how LLMs can engage in continuous, real-time interactions with users rather than one-off responses.

In the case method, students tackle real-world scenarios faced by professionals, adopting roles like managers or lawyers to analyze business cases, make decisions, and present their conclusions. This method involves role-play and mirrors real-world decision-making processes (Barnes et al., 1994). Case studies have long been central to business education, especially in MBA programs. They are recognized for cultivating critical skills such as decision-making, problem-solving, and communication (Blumenthal, 1991; Correia & Mayall, 2012), while also fostering deep learning and critical thinking (Swanson & Morrison, 2010). With the rise of LLMs like ChatGPT, this traditional approach is under pressure. On one hand, students increasingly depend on generative AI tools for case preparation, potentially weakening the learning process and resulting in more uniform and shallow analyses. On the other hand, when appropriately integrated, LLMs can offer adaptive learning experiences tailored to individual student needs, ultimately enriching the learning process (E. R. Mollick & Mollick, 2023).

This paper examines how LLMs can extend case-based learning beyond the classroom. By engaging with complex, real-world scenarios in non-classroom environments, LLMs could offer students more diverse learning opportunities and present increasingly sophisticated challenges. This approach may improve student engagement and foster a more level playing field when they return to class, ultimately enhancing long-term educational outcomes (Chiu & Hsieh, 2017; Tarng & Tsai, 2010).

2. Background & Literature

Our pilot study aims to extract the role-playing elements inherent in the case method and simulate them using Large Language Models (LLMs). To ground this approach, we review two key areas of literature: first, the capabilities and limitations of LLMs, particularly in interactive, educational settings; and second, the role of Role-Playing Games (RPGs) in fostering engagement and complex skill development in educational contexts. By bridging these areas, we aim to explore how LLMs can be used in the case method in business education.

2.1. Large Language Models

Large Language Models (LLMs) are advanced machine learning systems capable of processing and generating text based on vast datasets. Though sometimes considered stochastic parrots, certain cognitive abilities such as theory of mind seem to be emerging from them (van Duijn et al., 2023). Moreover, when asked to simulate a human being, they are able to mimick complex human behaviors in various contexts, including supply-demand dynamics (Brand et al., 2023), representation of diverse subpopulations (Beck et al., 2024; Moon et al., 2024), social science simulations (Argyle et al., 2023; Hewitt et al., 2024; Park et al., 2022).

While these models offer significant benefits for creating dynamic and engaging simulations, they also present several challenges. LLMs are known to produce "hallucinations" — instances where the models generate misleading or entirely fabricated information (Huang et al., 2023; Li et al., 2023; Xu et al., 2024). This tendency can enhance a simulation by adding creative elements, but it risks introducing inaccuracies regarding crucial learning outcomes, which could confuse students. The issue is particularly pronounced in back-and-forward interactions like in question-answering systems, where hallucinations are more frequent. Mitigating these inaccuracies often requires human oversight and remains a challenge (Duan et al., 2024).

Another significant limitation is the lack of explainability in LLMs. Despite advances in explainable AI, these systems often remain opaque due to their complex, multi-layered architectures (Boyko et al., 2023; Zhao et al., 2023). This opacity can hinder their integration into scientific treatments of them where understanding how conclusions are reached is essential.

Lastly, bias in LLMs is an additional concern. These models can perpetuate existing social and cognitive stereotypes, reflecting biases related to gender, race, age, and other socio-demographic factors (Acerbi & Stubbersfield, 2023; Kotek et al., 2023). They may also introduce biases related to beauty and age (Kamruzzaman et al., 2023), along with cultural biases that impact broader societal interactions (Talboy & Fuller, 2023; Tao et al., 2023). Predominantly trained on data from WEIRD (Western, Educated, Industrialized, Rich, and Democratic) societies, LLM outputs can exclude diverse global perspectives (Atari et al., 2023; Santurkar et al., 2023), often drawing on internet-based content that does not represent a universal view (Crockett & Messeri, 2023). While some argue that these biases merely reflect prevailing societal views and evoke ”naturalistic” responses (Dillion et al., 2023), the case method thrives on diversity and tension of opinions. It is therefore crucial to choose models that do not inadvertently resort to one world-view or steer them in other directions when they do.

2.2. Role Playing Games

A Role Playing Game (RPG) is a system with rules for game mechanics, context-providing stories, and social interaction frameworks for co-creating narratives (Cheville, 2016). In its most traditional form, participants act out their characters in historical or fictional scenarios. Table-top RPGs (TRPGs), also known as pen-and-paper role-playing games, involve participants describing their characters' actions through speech and sometimes movements; Dungeons & Dragons is perhaps the most well-known example.

RPGs are highly regarded for their educational potential, as they engage students on tactical, social, moral, and strategic levels, focusing on complex skill development rather than mere rote memorization (An & Cao, 2017; Bergström, 2012; Prager, 2019). Live Action Role Playing (LARP) games provide immersive experiences that significantly increase motivation and interest (Bowman & Standiford, 2015), while multiplayer RPGs promote collaboration and community building (Chen & Hwang, 2017; Snow, 2008). RPGs foster active involvement, engagement, and greater intrinsic motivation (Daniau, 2016).

Although there is limited research on the intersection of the case method and RPGs, the case method often parallels TRPGs by offering immersive, socially engaging experiences that similarly boost interest and motivation (Bowman & Standiford, 2015). Essential elements like reading case files and active student involvement set the stage for effective classroom interaction. However, implementing meaningful interactions for every participant in large classrooms can be challenging, particularly when scenarios demand numerous decisions or direct interactions between many players in parallel. Previous research has suggested that one potential solution is to designate the instructor as a player while students navigate the scenarios (Kadakia, 2005). Although this method addresses logistical hurdles, the authors also note that instructor-led RPG adaptations may reduce student engagement, underscoring the need for innovative approaches to sustain student interest and involvement.

3. Methodology

3.1. Negotiation Game

In our study, we investigate how to support RPG elements of the case method with LLM-based simulations. We selected a case commonly used in the Decision Analysis core course of our MBA program (de Santiago & Montgomery, 2018), which involves a scenario of a price war between an incumbent company and a new entrant with lower pricing. Typically, this case is role-played by two volunteers from the audience, each representing one of the companies with distinct strategies: the incumbent aims to maximize profitability, while the new entrant has a choice to focus on either increasing market share or profitability. Throughout the simulation, participants confront strategic trade-offs, such as the impact of consumer price sensitivity and the value of market share versus profit.

This game is an instance of a non-cooperative game where choosing to prioritize market share leads to a distributive outcome, while focusing on contribution margin maximization allows for potential cooperation. The main utility function for player i is the contribution margin (M in USD):

$$f_i(P_0, P_1) = f_i^E + \sigma \times \frac{\Delta P}{100} \tag{1}$$

$$M_i(P_0, P_1) = (P_i - C_i) \times \text{TAM} \times f_i(P_0, P_1), \tag{2}$$

where TAM is the Total Addressable Market, σ the sensitivity, f_i^E the MarketShare at equal prices, and P_i the price (ΔP the price difference with the other player), and C_i the variable cost. In this scenario, we assume that the demand function is known to both players, enabling them to calculate contribution margins and make informed pricing decisions. Only the price is under the direct control of the player and the players take turns. A representative example is shown in Figure 3 which shows the optimal strategy for both players (Rubrum and OceanTrans). The blue "Start" dot shows an initial situation where the price levels are $P_O = \$1400$ and $P_R = \$1300$. The best strategy for OceanTrans is to reduce its price to \$1169.5. For pedagogical purposes, we round numbers to the nearest fifty to $P'_O = \$1150$. As the simulation continues, the price war converges to a price of $P^*_O = \$950$ and $P^*_R = \$900$, close to the theoretical optimum at the intersection of the blue and green lines.

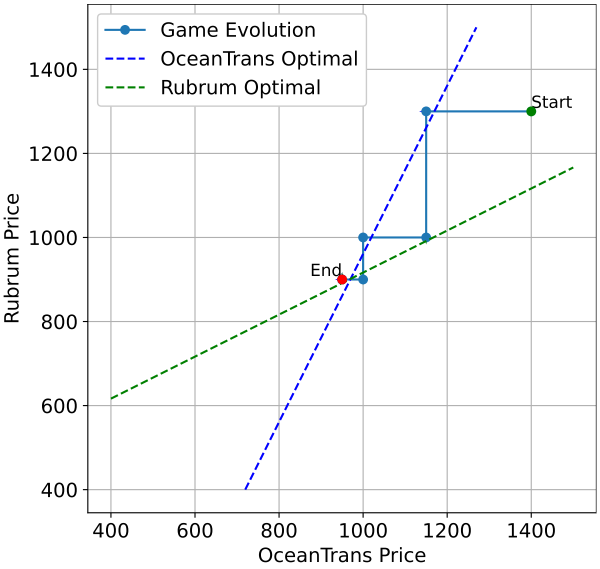

Figure 1. The profit-optimal strategy for each player (dotted lines), and its equivalent approximation through in-class role-play through discrete moves.

3.2. Simulations

We explored three simulation methods: prompt-based, RAG-based, and controller-based.

Prompt-based simulation is the most frequently used simulation method and also the simplest. This approach involves providing a commercial LLM, such as ChatGPT, with a detailed prompt designed to elicit responses similar to those from a role-playing game (RPG) simulator. Following the approach of E. Mollick et al., 2024, we develop an 8000-character prompt that defines roles and objectives for participants, step-by-step instructions, constraints for interaction, and a pedagogical context to ensure learning outcomes.

Retrieval Augmented Generation (RAG) uses a vector database to store case details. During inference, the user's query is embedded and matched with relevant case details, which are then integrated into the context to enrich the model's responses (Lewis et al., 2021). This approach overcomes the key limitation of the prompt-based simulation, which suffers from a restricted context window. For our RAG simulations, we used the Llama3:8b model (AI@Meta, 2024) for the chat module, alongside a Chroma vector database and Nomic embeddings ("Chroma DB", 2024). These are both open-source, improving the reproducibility of our findings compared to commercial systems like ChatGPT.

The Control System (CS) approach further refines the interaction dynamics between the user and the system by providing dedicate task-specific LLMs (Llama3:8b). The CS coordinates communication across various stages of the simulation, ensuring a smoother experience. Our implementation includes three main components.

First, the *set-up module* gathers information about the user's experience with the subject matter and their preferred role within the simulation, allowing the system to custom-tailor the interaction based on user proficiency and interest.

Second, the *simulation component* drives the actual RPG, creating a dynamic environment where decisions and interactions evolve in response to user inputs. Unlike conventional implementations, this LLM was not fine-tuned but operated using one-shot instruction learning similar to our initial prompt-based simulation. Additionally, the LLM was equipped with function calls ("tools") to perform calculations and manage case data on contribution margins, using an agentic approach to avoid generating incorrect information (Yao et al., 2023).

Lastly, the *debriefing module* helps users reflect on the game, reviewing decisions and clarifying the business principles applied during the simulation.

4. Findings and Implications

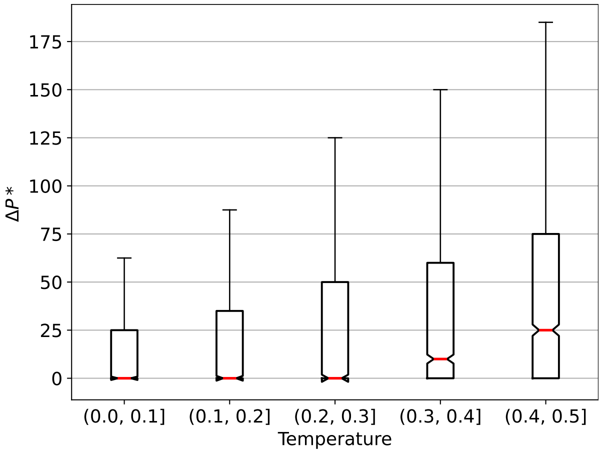

Figure 2. As we increase the temperature (x-axis), the difference between the final converged-upon price and the theoretical optimum becomes higher (y-axis).

As we fine-tune our three approaches, we asked a focus group of students to evaluate the three models on qualitative dimensions (e.g., engagement, accuracy, and learning outcomes). Although the sample size was too small for statistical analysis, the qualitative findings were consistent (Table 1). At higher temperature settings, all methods excelled at creating engaging scenarios (e.g., "*You are sitting in your office, pondering the next move. The clock is ticking, and you need to announce your freight rate for the following week.*"). However, higher temperature settings often caused the LLMs to deviate from the intended scripts, reducing accuracy despite efforts to craft prompts that managed this behavior. This was particularly noticeable in the prompt-based approach. We believe this stems from the challenge of balancing detailed and sparse information in the prompt, making it difficult for the LLM to stay focused and accurate.

Table 1. Qualitative findings of the pilot ($N = 4$).

Method	Engaging	Accurate	Pedagogy
Prompt	Excellent	Poor	Low
RAG	Excellent	Moderate	Moderate
CS	Excellent	Good	High

The RAG-based approach, while avoiding hallucinations, sometimes failed to stick to the script, triggering debriefing stages early and struggling with retrieving accurate data — such as contribution margins at specific price points. It also occasionally made basic calculation errors related to company cost structures, which hurt the simulation's credibility and realism.

The Control System (CS) approach was explicitly designed to address these shortcomings. We tested its performance by pitting it against itself in the game. In an ideal world, these simulations should end with both players choosing the theoretical optimum. In 8,000 such simulations, we observed deviations from this optimum due to hallucinatory behavior at higher temperature settings (Figure 2)[1]. For instance, at $t \approx .45$, the simulation did not converge to the optimal solution in most cases. Additionally, 25% of simulations ended \$75 off the intended results, even though the LLM had access to an oracle tool calculating the best action. At very low temperature settings ($t \approx 0$), the environment produced repetitive outcomes, despite instructions to avoid doing so. We settled on a temperature setting of $t = 0.1$ for the final simulation. All in all, we find this setting to provide a reliable simulation environment despite the fact that we are "only" using an 3x8 billion parameter model.

5. Conclusion

In this study, we explored how integrating the case method with LLMs using RPG environments can enhance student engagement. We evaluated three approaches: prompt-based, Retrieval Augmented Generation (RAG)-based, and controller-based simulations. Our analysis showed that while prompt-based simulations are easy to implement, they often lack accuracy due to limited context. The RAG-based method improves context management, but can still result in the model deviating and making occasional errors in data retrieval. However, the controller-based approach effectively addresses these issues, providing a more robust and engaging educational experience when the temperatures are controlled appropriately.

Looking ahead, we aim to further refine our methods by using bigger models and fine-tuning LLMs to better handle specific tasks within the simulations. We also plan to explore advanced RAG systems, such as graph RAG (Edge et al., 2024), to improve the accuracy and effectiveness of our LLMs. These efforts will keep the current focus on integration in educational settings to maximize learning outcomes. In a world where learning environments continuously evolve, the power of LLMs will not merely lie in their ability to explain concepts to us — but in their capacity to transform how we engage, learn, and grow in the face of complex challenges.

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, *120*(44), e2313790120. https://doi.org/10.1073/pnas.2313790120

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies.

AI@Meta. (2024). Llama 3 Model Card.

An, Y.-J., & Cao, L. (2017). The Effects of Game Design Experience on Teachers' Attitudes and Perceptions regarding the Use of Digital Games in the Classroom. *TechTrends*, *61*(2), 162–170. https://doi.org/10.1007/s11528-016-0122-8

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which Humans? https://doi.org/10.31234/osf.io/5b26t

Barnes, L. B., Christensen, C. R., & Hansen, A. J. (1994). *Teaching and the Case Method: Text, Cases, and Readings* (Third edition). Harvard Business Review Press.

Beck, T., Schuff, H., Lauscher, A., & Gurevych, I. (2024). Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2589–2615). Association for Computational Linguistics.

Bergström, K. (2012). Creativity Rules: How Rules Impact Player Creativity in Three Tabletop Role-playing Games. *International Journal of Role-Playing*, (3), 4–17. https://doi.org/10.33063/ijrp.vi3.221

Blumenthal, J. (1991). Use of the Case Method in MBA Education. *Performance Improvement Quarterly*, *4*(1), 5–13. https://doi.org/10.1111/j.1937-8327.1991.tb00486.x

Bowman, S., & Standiford, A. (2015). Educational Larp in the Middle School Classroom: A Mixed Method Case Study. *International Journal of Role-Playing*, 4–25. https://doi.org/10.33063/ijrp.vi5.233

Boyko, J., Cohen, J., Fox, N., Veiga, M. H., Li, J. I.-H., Liu, J., Modenesi, B., Rauch, A. H., Reid, K. N., Tribedi, S., Visheratina, A., & Xie, X. (2023). An Interdisciplinary Outlook on Large Language Models for Scientific Research. https://doi.org/10.48550/ARXIV.2311.04929

[1] We measure accuracy as the average distance between the final price set by the simulation and the theoretical optimum $\Delta P^* = \frac{1}{N}\sum_i (P_i - P_i^*)$

Brand, J., Israeli, A., & Ngwe, D. (2023). Using LLMs for Market Research.

Chen, C.-H., & Hwang, G.-J. (2017). Effects of the Team Competition-Based Ubiquitous Gaming Approach on Students' Interactive Patterns, Collective Efficacy and Awareness of Collaboration and Communication. *Educational Technology & Society*, *20*(1), 87–98.

Cheville, R. A. (2016). Linking capabilities to functionings: Adapting narrative forms from role-playing games to education. *Higher Education*, *71*(6), 805–818.

Chiu, F.-Y., & Hsieh, M.-L. (2017). Role-Playing Game Based Assessment to Fractional Concept in Second Grade Mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(4), 1075–1083. https://doi.org/10.12973/eurasia.2017.00659a

Chroma DB. (2024).

Correia, C., & Mayall, P. (2012). The Use of the Case Method in Teaching Corporate Finance: An Evaluation.

Crockett, M., & Messeri, L. (2023). Should large language models replace human participants? https://doi.org/10.31234/osf.io/4zdx9

Daniau, S. (2016). The Transformative Potential of Role-Playing Games-. *Simulation and Gaming*, *47*(4), 423–444. https://doi.org/10.1177/1046878116650765

de Santiago, R., & Montgomery, T. (2018). OceanTrans: Shipping in the Red Sea.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Duan, H., Yang, Y., & Tam, K. Y. (2024). Do LLMs Know about Hallucination? An Empirical Investigation of LLM's Hidden States.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From Local to Global: A Graph RAG Approach to Query-Focused Summarization.

Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R. (2024). Predicting Results of Social Science Experiments Using Large Language Models.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.

Kadakia, M. (2005). Increasing student engagement by usingMorrowind to analyze choices and consequences. *TechTrends*, *49*(5), 29–32. https://doi.org/10.1007/BF02763687

Kamruzzaman, M., Shovon, M. M. I., & Kim, G. L. (2023). Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models. https://doi.org/10.48550/ARXIV.2309.08902

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. *Proceedings of The ACM Collective Intelligence Conference*, 12–24. https://doi.org/10.1145/3582269.3615599

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. https://doi.org/10.48550/arXiv.2005.11401

Li, M., Shi, T., Ziems, C., Kan, M.-Y., Chen, N. F., Liu, Z., & Yang, D. (2023). CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1487–1505. https://doi.org/10.18653/v1/2023.emnlp-main.92

Mollick, E., Mollick, L., Bach, N., Ciccarelli, L., & Przystanski, B. (2024). AI Agents and Education: Simulated Practice at Scale.

Mollick, E. R., & Mollick, L. (2023). Assigning AI: Seven Approaches for Students, with Prompts. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4475995

Moon, S., Abdulhai, M., Kang, M., Suh, J., Soedarmadji, W., Behar, E. K., & Chan, D. M. (2024). Virtual Personas for Language Models via an Anthology of Backstories.

Park, J. S., Popowski, L., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2022). Social Simulacra: Creating Populated Prototypes for Social Computing Systems.

Prager, R. H. P. (2019). Exploring The Use of Role-playing Games In Education. *The MT Review*.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect?

Snow, C. (2008). Dragons in the stacks: An introduction to role-playing games and their value to libraries. *Collection Building*, *27*(2), 63–70. https://doi.org/10.1108/01604950810870218

Swanson, D. A., & Morrison, P. A. (2010). Teaching Business Demography Using Case Studies. *Population Research and Policy Review*, *29*(1), 93–104. https://doi.org/10.1007/s11113-009-9155-4

Talboy, A. N., & Fuller, E. (2023). Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption. https://doi.org/10.48550/ARXIV.2304.01358

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2023). Auditing and Mitigating Cultural Bias in LLMs. https://doi.org/10.48550/ARXIV.2311.14096

Tarng, W., & Tsai, W. (2010). The design and analysis of learning effects for a game-based learning system. *61*, 336–345.

van Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., & van der Putten, P. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests. In J. Jiang, D. Reitter, & S. Deng (Eds.), *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 389–402). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.25

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. https://doi.org/10.48550/arXiv.2210.03629

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2023). Explainability for Large Language Models: A Survey.

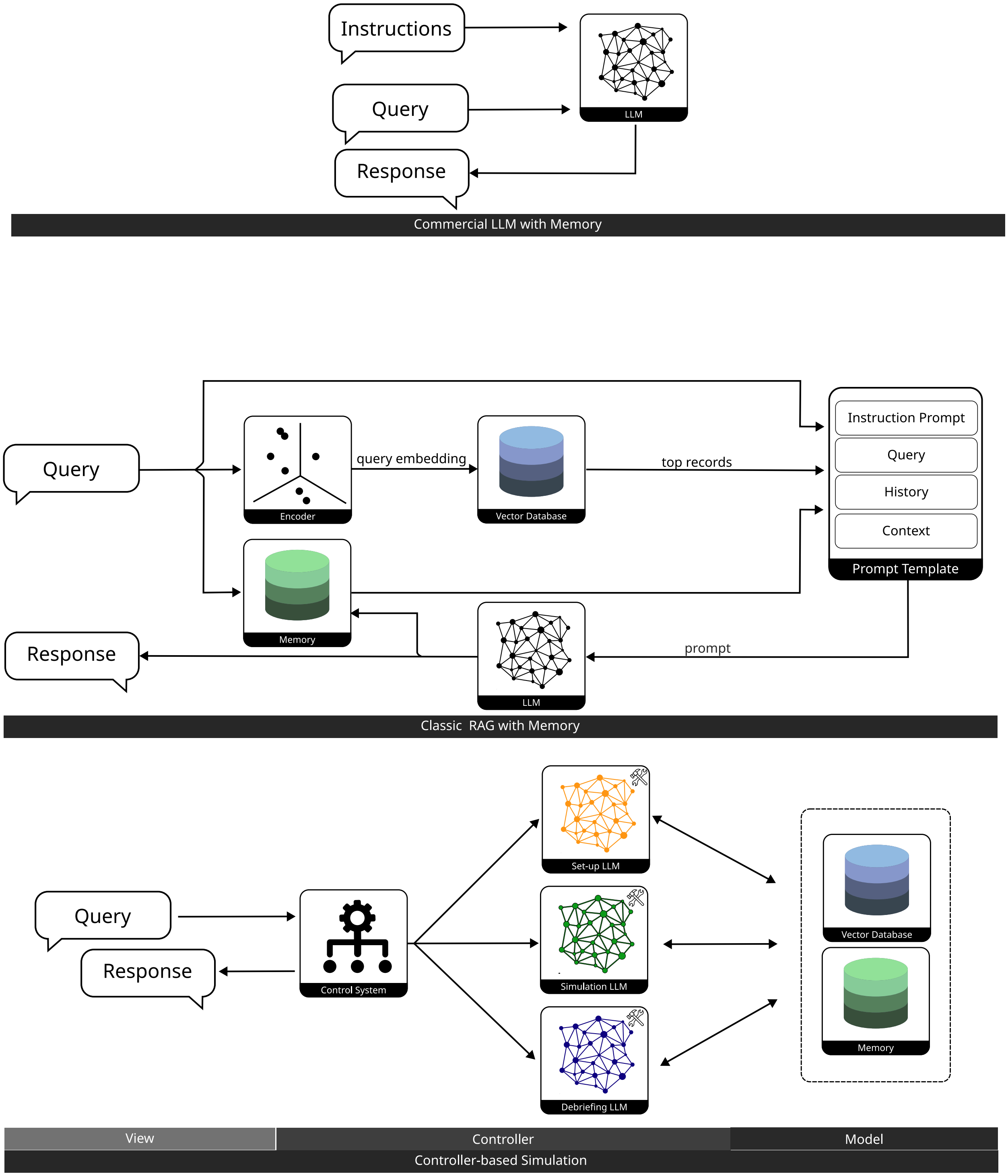

Figure 3. The three architectures explored in our approach.