# When Avatars Debate: Design and Evaluation of an LLM-Based Debate Platform for Educational Applications

Antonio Cvitković[1], Luka Batistić[1,2], Antonio Maffei[3], and Sandi Ljubić[1,2(✉)]

[1] University of Rijeka, Faculty of Engineering, Vukovarska 58, 51000 Rijeka, Croatia
acvitkovic@uniri.hr, luka.batistic@riteh.uniri.hr
[2] Center for Artificial Intelligence and Cybersecurity, University of Rijeka, R. Matejčić 2, 51000 Rijeka, Croatia
sandi.ljubic@riteh.uniri.hr
[3] Department of Production Engineering, Royal Institute of Technology (KTH), 114 28 Stockholm, Sweden
maffei@kth.se

**Abstract.** Large language models (LLMs) create new opportunities for interactive learning environments that foster critical thinking and exposure to diverse perspectives. This paper introduces a prototype e-learning system that uses these models to conduct argumentative debates between digital avatars representing opposing viewpoints on a given topic. The debates are generated using locally deployed LLMs, enabling controlled, scalable discussion scenarios while avoiding reliance on external cloud services. By dynamically presenting contrasting arguments, the system is designed to encourage learners to consider alternative perspectives. To support active engagement with the debate content, the platform also features automatic quiz generation, allowing users to assess their comprehension immediately after observing the discussion. The system was evaluated in a study with 18 participants who observed AI-driven debates on two topics: a technical trade-off in software engineering and a controversial socio-technical issue. The evaluation had two objectives: assessing perceived educational usefulness and willingness to use the system for learning, and comparing the debate output quality of three instruction-tuned models from different providers (Alibaba Qwen 3, Meta Llama 3.1, and IBM Granite 3.0), specifically their 8-billion-parameter variants with 4-bit quantization. The study is designed as a proof-of-concept exploratory evaluation, focusing on perceived usefulness and user preferences rather than actual learning outcomes. The results indicate generally positive user perceptions of the debate-based learning format and reveal differences in model performance across specific evaluation criteria.

**Keywords:** e-learning · digital avatars · large language models · AI-supported debate · AI in education

© The Author(s) 2027
Published by Springer Nature Switzerland AG 2026
G. Pezzotta et al. (Eds.): APMS 2026, IFIP AICT 813, pp. 210–225, 2027.
https://doi.org/10.1007/978-3-032-38618-2_14

## 1 Introduction

Digital learning platforms often rely on relatively static materials such as documents, slides, and quizzes. Recent work on AI-supported education, however, explores more interactive learning experiences, including chatbot-assisted learning and classroom debate with LLM support [10,12]. Among interactive formats, previous work suggests that debate-based learning can expose learners to multiple perspectives and support argumentative development [4]. However, organizing debates in educational settings can be logistically challenging, requiring coordination among participants, moderation, and preparation of structured discussion topics.

Recent advances in large language models (LLMs) offer new opportunities to address these challenges. Recent studies indicate that LLM-based systems can generate arguments, respond to opposing viewpoints, and sustain structured debate-like interactions in experimental debate platforms [1]. Related educational studies suggest that similar debate-like interactions can also be integrated into classroom and digital-human settings [9,12]. This creates an opportunity to simulate debates in a controlled and scalable manner, allowing students to observe argumentative exchanges without the logistical overhead of organizing human participants. This motivates interest in locally deployable debate systems built on openly available model families. At the same time, the use of LLMs in educational contexts raises several concerns. These include bias in debate interactions [3], the risk of factual inaccuracies presented in an authoritative tone [5], and the possibility that students may become over-reliant on generated content in ways that reduce learner autonomy [12]. Addressing these concerns requires carefully designed interaction mechanisms that encourage active learning and critical reflection.

We present a prototype e-learning platform in which digital avatars powered by locally deployed LLMs conduct structured debates on user-defined topics. The system allows configurable debate parameters and integrates text-to-speech synthesis to enhance accessibility and immersion. To reduce passive consumption, the platform also generates quizzes to assess understanding of the debate's key points. The proposed platform was evaluated in an experiment involving 18 participants who observed AI-driven debates on two topics: a technical trade-off in software engineering and a controversial socio-technical issue. The evaluation pursued two objectives: assessing the perceived educational usefulness of the system and participants' willingness to use it for learning, and comparing the debate output quality of three LLMs: Alibaba Qwen 3, Meta Llama 3.1, and IBM Granite 3.0.

This paper makes three contributions. First, we design and implement a prototype platform that enables configurable AI-driven debates between digital avatars powered by locally deployed LLMs. Second, we integrate automatic quiz generation to reinforce learning and encourage active engagement with debate content. Third, we present an empirical evaluation that assesses both the perceived educational usefulness of the debate-based learning format and differences in debate performance across three instruction-tuned LLMs.

The paper is organized as follows. Section 2 reviews related work on AI-assisted education and multi-agent debate systems. Section 3 describes the architecture of the proposed system and its debate configuration mechanisms. Section 4 presents the experimental setup and evaluation results. Section 5 discusses the findings and their implications for AI-supported learning environments, and Sect. 6 concludes the paper.

## 2 Related Work

Related work relevant to this paper can be broadly categorized into three research streams: multi-agent debate (MAD) systems based on large language models, AI-supported argumentation in educational contexts, and interactive debate platforms involving digital avatars.

The first stream focuses on MAD approaches using LLMs and their potential to improve reasoning quality. Earlier work has shown that debates between multiple LLM agents can enhance factuality and reasoning compared to responses generated by a single model instance [5]. Other studies have investigated debate setups in which human or LLM judges evaluate competing arguments and select better-supported responses [8]. Recent research has also explored ways to optimize the efficiency of multi-agent interactions. For example, $S^2$-MAD reduces token overhead through selective participation and redundancy filtering [11], while OPTAGENT uses verbal reinforcement learning to refine multi-agent collaboration [2]. Additional studies examine how anonymization strategies can reduce self-bias and sycophancy in multi-agent debate dynamics [3].

The second stream investigates the role of argumentation and debate in educational settings. Previous work suggests that debate-based learning activities can expose learners to multiple perspectives and support argumentative development [4]. At the system level, persona-driven multi-agent planning has been shown to increase the diversity of perspectives and the persuasiveness of generated argumentative content [7]. More broadly, research on chatbot-assisted learning suggests that chatbots can foster interactive engagement and, for learners with stronger learning perception, be associated with higher motivation [10]. Recent studies have also explored the use of AI-supported debate preparation activities in educational settings. For example, chatbot-assisted classroom debates have been shown to help students generate arguments, anticipate opposing viewpoints, and improve the organization and elaboration of argumentative content [6]. Such systems can reduce the effort required for information gathering and initial idea formulation, allowing learners to focus more on reflective reasoning and discussion. However, the educational value of these activities depends strongly on careful pedagogical design, as students must still critically evaluate the generated content, assess the credibility of arguments, and make their own strategic decisions during debate activities. Thus, generative AI functions primarily as a cognitive support tool rather than a substitute for critical thinking.

The third stream examines AI-driven debate systems as interactive applications. DebateBrawl combines LLM-based argument generation with adversarial search and evolutionary mechanisms to produce adaptive debates [1]. Other

studies have explored digital-human avatar debates in which participants design personas and then reflect while observing the resulting discussions [9].

In contrast to previous work that focuses primarily on either the algorithmic aspects of multi-agent debate or educational studies of argumentation, this paper seeks to bridge these perspectives by advancing multi-agent debate systems toward user-ready deployment in educational settings. We present a Unity-based prototype platform that enables configurable avatar debates powered by locally deployed LLMs, supporting structured discussions on user-defined topics. Beyond debate functionality, the platform includes automated quiz generation to promote active engagement and reinforce knowledge. The system is evaluated through a controlled user study that examines both the perceived educational usefulness of the debate-based learning format and differences in debate quality across three instruction-tuned models.

## 3 A Framework for AI-Driven Debates

This section presents the framework underlying the proposed AI-driven debate platform. The prototype is implemented in Unity and coordinates components responsible for debate orchestration, LLM inference, user interaction, and post-debate assessment. The system operates in three main phases: session setup, live debate loop, and post-debate quiz generation.

### 3.1 Phase 1: Session Setup

The first phase begins at the application's start screen, where the user can access the debate interface or configure debate parameters. The control flow for this initialization stage is shown in Fig. 1.

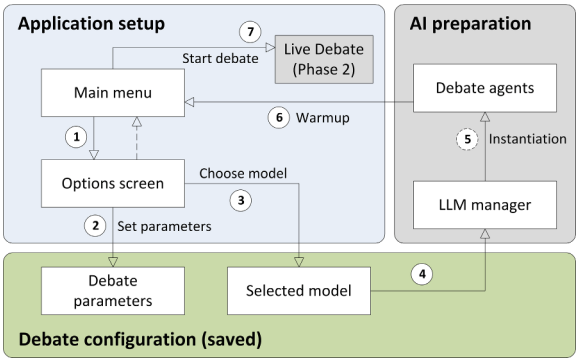


**Fig. 1.** Phase 1 workflow: system initialization and debate session configuration.

The illustrated workflow depicts a scenario in which a debate administrator prepares a new debate session by configuring system parameters and selecting the

LLM to be used. In practice, users can bypass this step by loading saved settings, but this path is omitted from the figure for clarity. The startup process therefore includes preparing the user-selected debate configuration and initializing the LLM components.

Within the configuration panel, users can adjust parameters that define the debate structure, including the number of participating avatars, the number of debate rounds, and the approximate maximum length of responses generated by each avatar (Fig. 2). The interface also allows users to select the LLM for the debate from a set of locally stored models in the GGUF format. Additional models can be added to the designated model directory. An optional setting enables text-to-speech (TTS) synthesis for avatar responses.

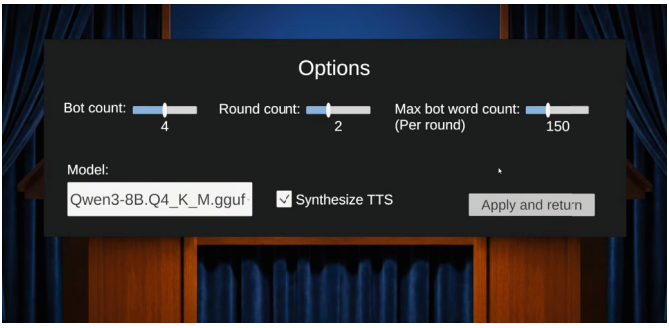


**Fig. 2.** Debate configuration panel allowing the user to adjust debate parameters and select the LLM model.

During initialization, the system performs a warm-up step to load and prepare the LLM instances for inference. This reduces first-response latency. At this stage, system prompts are also assigned to the debate agents. These prompts dynamically adapt to the debate configuration, informing agents about their teammates, opposing agents, and the approximate response length expected for each message. Once initialization is complete, the user can proceed to the debate interface.

### 3.2 Phase 2: Live Debate Loop

The second phase involves conducting the debate itself. As shown in Fig. 3, the system coordinates several modules that generate debate messages, synthesize speech output, and maintain the session transcript. The term *loop* refers to the iterative structure of the debate, in which each avatar generates a response in turn, and this sequence repeats for a predefined number of debate rounds.

At the beginning of the debate, the user enters the discussion topic in a text input field. After the topic is submitted, the debate agents generate responses according to the predefined debate configuration. All models were prompted with

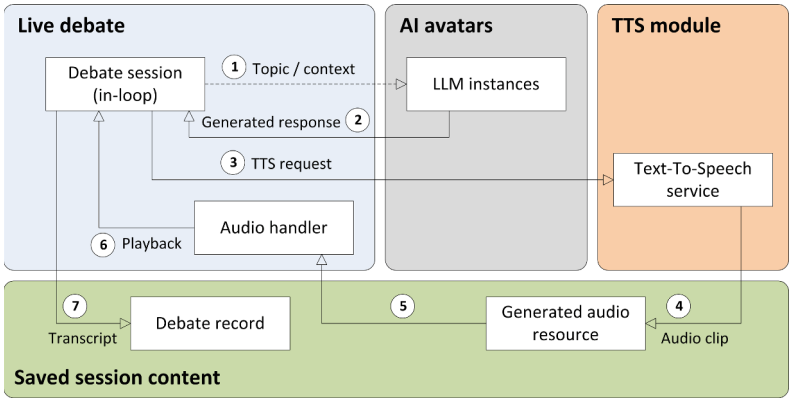


**Fig. 3.** Phase 2 workflow: live debate loop with LLM response generation and text-to-speech synthesis.

the same template, which defined debate roles, assigned opposing stances, and instructed each agent to maintain its position throughout the discussion. System-level instructions were included at each turn to reinforce the agent's assigned role and stance. The prompts also specified response structure and length to ensure consistency across all models. An example of the debate interface during this phase is shown in Fig. 4.

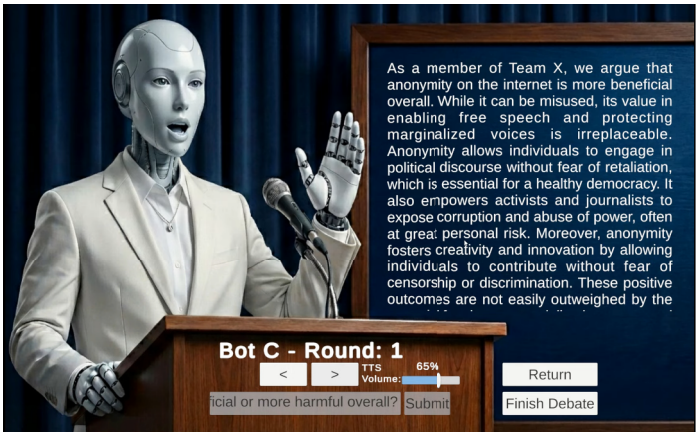


**Fig. 4.** Example of the debate interface during the live debate phase.

To improve responsiveness, the system overlaps text generation and speech synthesis. While the text-to-speech module processes the previously generated message, the LLM simultaneously generates the next response. This reduces perceived waiting time between messages. Messages are presented sequentially,

and navigation controls let the user move forward or backward through the discussion as needed.

Because the system imposes an approximate token limit per response, some generated messages may occasionally be truncated. To minimize disruptions to coherence, this limit is also communicated to the debate agents through the system prompt, encouraging responses within the expected size range. Once all debate rounds are completed and each avatar has spoken, the user may proceed to the final phase of the system.

### 3.3 Phase 3: Quiz Generation and Session Wrap-Up

The third phase begins after the debate concludes. The operational workflow for this phase is shown in Fig. 5. During this step, the system processes the debate transcript and prepares it as input for quiz generation. The transcript is first anonymized to remove references to specific agents, ensuring that the quiz generation model focuses on informational content rather than avatar identity.

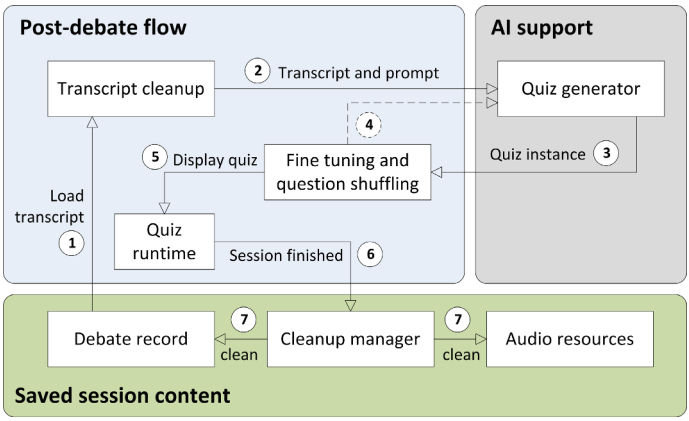


**Fig. 5.** Phase 3 workflow: quiz generation and session wrap-up after the debate.

The processed transcript is then provided to the LLM responsible for generating the quiz. The system enforces a predefined quiz structure (via system prompt) and applies post-processing steps to correct common formatting inconsistencies that may occur in LLM-generated output. The quiz answers are then shuffled to vary the answer order.

After the quiz is prepared, the system loads the quiz interface in Unity. The generated quiz consists of ten multiple-choice questions with four possible answers each, reflecting the simplified format adopted in this proof-of-concept implementation. The quiz is not time-restricted, allowing users to review questions at their own pace before submitting answers. Upon submission, immediate feedback is provided, indicating the number of correct responses. The user may then return to the main menu and start a new session.

## 4 Evaluation of the System and Utilized Models

To evaluate the proposed prototype platform, we conducted a user study with 18 participants. The evaluation was guided by the following research questions:

- How do users perceive the quality of debate outputs generated by different instruction-tuned LLMs in a multi-agent setting?
- How do users perceive the usefulness of AI-driven debate, presented alongside the post-quiz session, as a support tool in an educational context?

Because the platform is a proof-of-concept implementation, the study was designed as an exploratory evaluation to collect initial user feedback and identify potential differences in model performance within the debate setting.

### 4.1 Experiment Procedure

The experiment was conducted online with 18 volunteer participants, primarily from the education sector, including mostly university students and a smaller number of teaching staff. Participants were recruited from several countries in Europe and the United States, and all provided informed consent.

Instead of interacting directly with the prototype system, participants viewed recorded demonstrations of the system's operation (i.e., the debates). This ensured that all participants evaluated identical debate outputs under the same system configuration, providing a controlled basis for comparing models. However, this design reduces ecological validity compared to direct interaction with the system, since participants did not actively engage with the platform during the evaluation. The recordings showed debates between digital avatars generated using three LLMs: Alibaba Qwen 3, Meta Llama 3.1, and IBM Granite 3.0. The demonstrations also briefly showed the system's configuration interface and the quiz generated after each debate. Each recorded debate session, including the quiz-generation demonstration, lasted no more than 8 min.

Two debate topics were prepared for the study. The first topic addressed a technical trade-off in software engineering, specifically whether Java or C++ is a more suitable programming language for game development. The second topic concerned a controversial socio-technical issue, namely whether anonymity on the Internet is beneficial or harmful for online communities.

After viewing the debates, participants completed a questionnaire consisting of two parts. The first part evaluated the perceived quality of debate outputs generated by each model. Participants rated their agreement with the following statements on a scale from 1 (strong disagreement) to 10 (strong agreement):

- **Argument clarity:** "The arguments presented by this model were clear and understandable."
- **Counterargument handling:** "This model responded well to opposing viewpoints."
- **Insight generation:** "This model introduced new and interesting points during the discussion."

- **Trustworthiness:** "I trusted the accuracy of the information presented by this model."
- **Future use intention:** "I would choose this model again for future discussions."

The second part of the questionnaire assessed participants' perceptions of the proposed platform as a learning support tool. Participants responded using the same scale:

- **Conceptual understanding:** "The debate format helped me understand the topic better."
- **Learning from debate:** "Seeing opposing viewpoints improved my learning."
- **Knowledge reinforcement (quiz):** "The generated quiz helped reinforce what I learned."
- **System adoption intention:** "I would use this system to learn other topics."

All debate instances shown to participants were generated using the same system configuration: four digital avatars participated in each debate, the discussion consisted of two rounds, the maximum message length was approximately 150 words, and text-to-speech synthesis was enabled. To minimize ordering effects, the presentation order was counterbalanced across participants using six questionnaire variants, one for each model-order permutation.

### 4.2 Results

Before presenting the user evaluation results, we briefly summarize the technical characteristics of the evaluated models. All three models are instruction-tuned LLMs of comparable scale, each with approximately 8 billion parameters and deployed using the same quantization scheme (Q4_K_M). This configuration enables efficient local inference while maintaining similar computational requirements across models.

**Table 1.** Technical characteristics and measured inference throughput of the evaluated models.

| Model | Parameters | Quantization | Size (GB) | Throughput (tok/s) |
|---|---|---|---|---|
| IBM Granite 3.0 | 8B | Q4_K_M | 4.60 | $61.73 \pm 0.68$ |
| Meta Llama 3.1 | 8B | Q4_K_M | 4.58 | $62.36 \pm 2.89$ |
| Alibaba Qwen 3 | 8B | Q4_K_M | 4.68 | $62.14 \pm 1.51$ |

To verify that the evaluated models operate under comparable runtime conditions, we measured their inference throughput. All models were executed locally on a system equipped with an NVIDIA RTX 3080 laptop GPU using the same

quantized GGUF format. Table 1 summarizes the measured model sizes and token generation throughput. Throughput was measured in tokens per second and averaged over 24 consecutive inference runs for each model. The three evaluated models exhibit very similar memory footprints and inference speeds, indicating that observed differences in user ratings cannot be attributed to computational disparities.

To obtain an overall indicator of perceived model quality, participant ratings across the five evaluated features were aggregated. For each participant and model, the ratings for the five evaluation criteria were averaged to produce a single composite score representing the model's overall perceived performance. The resulting descriptive statistics are visualized in Fig. 6.

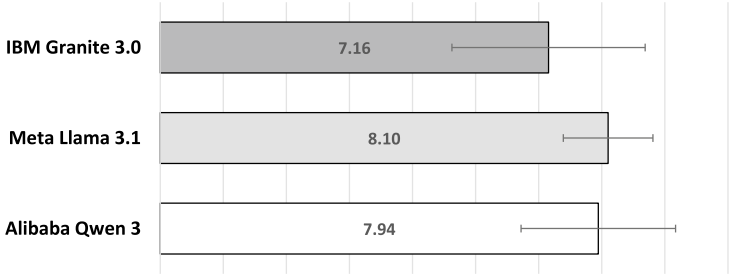


**Fig. 6.** Mean participant ratings of the evaluated LLMs aggregated across all assessed features (error bars show standard deviation).

As shown, Meta Llama received the highest mean ratings, followed closely by Alibaba Qwen, while IBM Granite received the lowest overall scores. However, a Friedman test conducted on the aggregated ratings did not reveal statistically significant differences among the three models ($\chi^2(2) = 4.254$, $p = 0.119$). This result suggests that, when considering the overall evaluation across all assessed aspects, the models were perceived similarly by participants. Nevertheless, aggregating ratings across multiple evaluation criteria may obscure potential differences in specific model capabilities. To gain a more detailed understanding of model performance, we therefore analyzed the ratings on a feature-by-feature basis. Figure 7 shows the mean participant ratings for each evaluated feature and model.

Visual inspection indicates that differences between models vary across individual criteria. In general, Meta Llama tends to receive the highest ratings across the evaluated features, while IBM Granite typically receives the lowest scores. Ratings for Alibaba Qwen generally fall between those of the other two models. To statistically examine these differences, separate Friedman tests were conducted for each evaluated feature.

The analysis revealed statistically significant differences for *Insight generation* ($\chi^2(2) = 7.651$, $p = 0.022$) and *Future use intention* ($\chi^2(2) = 7.387$,

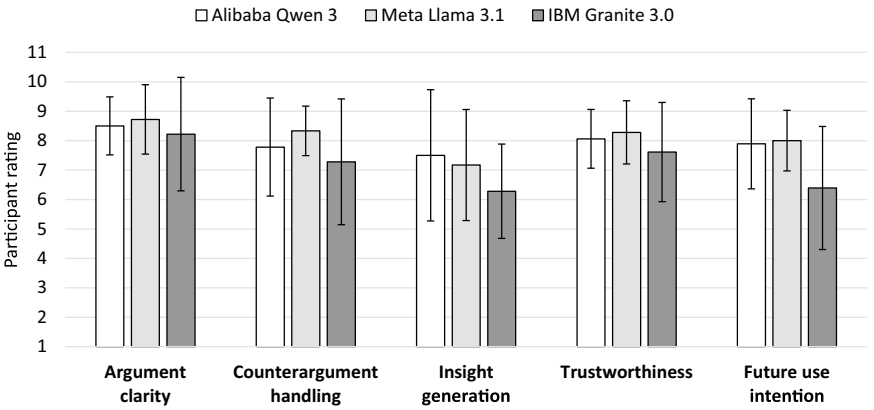


**Fig. 7.** Mean participant ratings for each evaluated feature across the three models (error bars show standard deviation).

$p = 0.025$), while no significant differences were observed for *Argument clarity*, *Counterargument handling*, or *Trustworthiness* (Table 2).

**Table 2.** The Friedman test results on evaluated model features ($n = 18$).

| Feature | Friedman $\chi^2(2)$ | $p$ | Kendall's $W$ |
|---|---|---|---|
| Argument clarity | 1.042 | 0.594 | 0.029 |
| Counterargument handling | 2.847 | 0.241 | 0.079 |
| Insight generation | 7.651 | 0.022* | 0.213 |
| Trustworthiness | 1.733 | 0.420 | 0.048 |
| Future use intention | 7.387 | 0.025* | 0.205 |

To further investigate these effects, post-hoc pairwise comparisons were performed using Wilcoxon signed-rank tests with Holm correction for multiple comparisons (Table 3).

For the *Insight generation* feature, none of the pairwise comparisons remained statistically significant after adjustment, indicating that although the overall Friedman test suggested differences among models, no specific model pair could be reliably distinguished. This outcome may reflect reduced statistical power associated with multiple-comparison corrections in small-sample exploratory studies. In contrast, post-hoc analysis for *Future use intention* revealed two statistically significant differences. Specifically, IBM Granite received significantly lower ratings than both Meta Llama and Alibaba Qwen, whereas the difference between Meta Llama and Alibaba Qwen was not statistically significant. These results suggest that participants were significantly less inclined to use IBM Granite for future discussions compared to the other two models.

**Table 3.** Post-hoc pairwise comparisons using Wilcoxon signed-rank tests with Holm correction for multiple comparisons.

| Feature | Comparison | Z | $p_{adj}$ (Holm) |
|---|---|---|---|
| Insight generation | Llama vs Qwen | -0.688 | 0.491 |
| | Granite vs Qwen | -1.959 | 0.150 |
| | Granite vs Llama | -1.395 | 0.326 |
| Future use intention | Llama vs Qwen | -0.214 | 0.831 |
| | Granite vs Qwen | -2.297 | 0.044* |
| | Granite vs Llama | -2.579 | 0.030* |

Regarding the evaluation of the educational potential of the proposed solution, participants responded positively to the prototype platform as a whole (Fig. 8), which combines a debate-based learning format with automatically generated quizzes. Mean participant ratings across the four evaluated aspects ranged from 7.89 to 8.50 on the 10-point scale, indicating generally favorable perceptions of the system's usefulness as a learning support tool.

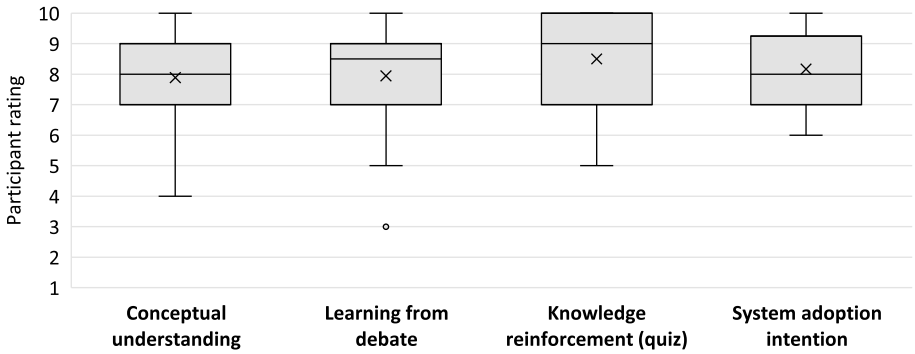


**Fig. 8.** Distribution of participant ratings for the prototype's perceived impact on learning across the four evaluated aspects.

Somewhat greater variability was observed in responses related to conceptual understanding of the debate content. This variation may reflect differences in participants' prior familiarity with the selected debate topics. Despite this variability, all four evaluated aspects received consistently high ratings. In particular, participants responded very positively to the role of opposing viewpoints in supporting learning and to the quiz component designed to reinforce knowledge after the debate, with median ratings above 8.5. These results suggest that the combination of structured AI-generated debates and post-discussion assessment may be a promising approach for supporting reflective learning and engagement.

Furthermore, participants expressed a strong willingness to adopt such a system for learning additional topics. This finding indicates that AI-driven debates may have potential as complementary learning tools in educational settings, particularly in contexts where exposure to diverse perspectives and active reflection are desirable.

## 5 Discussion

The obtained results provide several insights into both the behavior of the evaluated LLMs in a debate setting and the educational potential of the proposed debate-based learning platform. It is important to note that the present study evaluates perceived usefulness rather than objective learning gains or educational effectiveness, and should therefore be interpreted as exploratory.

Although the aggregated evaluation across all features did not reveal statistically significant differences between the models, the feature-level analysis identified meaningful distinctions in how the models behaved during debates. In particular, IBM Granite received significantly lower ratings in the *Future use intention* criterion, suggesting that participants perceived its debate contributions as less suitable for sustained argumentative discussions. During experimentation with the prototype, we observed that debate agents powered by IBM Granite occasionally switched their argumentative stance despite prompts explicitly instructing them to maintain a predefined position. When such divergence occurred, other agents sometimes followed the shift in stance, leading to a gradual collapse of the adversarial debate structure. In longer debates, this effect may result in multiple agents slowly converging toward agreement rather than maintaining opposing viewpoints. This tendency may help explain why participants expressed a lower willingness to use IBM Granite in future discussions and why the model also received the lowest ratings in the *Insight generation* criterion.

Beyond comparing individual models, the results also highlight the educational potential of AI-driven debate environments. Participants responded positively to the debate-plus-quiz format, suggesting that combining contrasting viewpoints with post-discussion assessment may support understanding and reflective engagement with complex topics. This interpretation is broadly consistent with prior work linking AI-mediated debate settings to improved argumentative structure and metacognitive awareness [4]. It also aligns with studies showing that AI-supported classroom debate can support learner engagement and collaborative learning, while introducing potential risks such as dependency and reduced autonomy [12]. However, the quiz component represents an additional feature specific to the proposed platform. The quality, factual accuracy, and difficulty of the generated quizzes were not systematically evaluated in this study and should be assessed in future research. Importantly, our results suggest that even relatively small, locally deployable models (approximately 8B parameters) can support debate-based learning experiences that participants perceive as useful. This is relevant for institutions seeking AI-assisted learning tools that do not rely on external cloud services.

While the results affirm educational potential, deploying AI-driven debate systems also raises several important considerations. LLMs may exhibit bias in debate interactions [3] or present information confidently yet inaccurately [5]. In educational contexts, these risks, together with cognitive dependency and reduced learner autonomy [12], necessitate careful system design that encourages critical reflection rather than passive acceptance of generated content. Including post-debate quizzes in the proposed platform is one mechanism for promoting active engagement with the material. Additional safeguards may include transparency mechanisms that clearly indicate content is AI-generated, as well as integrating external knowledge sources to verify factual claims. These considerations align with emerging regulatory frameworks, such as the EU AI Act, which emphasize transparency, accountability, and human oversight.

Several directions for future work arise from this study. Debate topics could be more closely aligned with specific course curricula using retrieval-augmented generation (RAG), enabling the system to incorporate authoritative educational materials when constructing arguments. This integration could reduce the risk of factual inaccuracies, support verification of generated content, and enhance the pedagogical relevance of generated debates. Additionally, model behavior in debate environments may benefit from targeted adaptation strategies. Techniques such as low-rank adaptation (LoRA) could specialize models for argumentative interaction patterns, helping agents maintain consistent debate roles and improving the reliability of auxiliary tasks such as quiz generation. Improved context management strategies, such as context summarization or compaction between debate turns, may also help models remain coherent during longer discussions.

The evaluated models were chosen for comparable configurations, not to represent the strongest locally deployable models currently available. Given the rapid progress in compact language models, future AI-driven debate platforms may achieve significantly higher debate quality while remaining deployable on consumer hardware.

## 6 Conclusion

This paper presented a prototype platform that enables customizable AI-driven debates between digital avatars powered by locally deployed LLMs. An exploratory user study involving 18 participants evaluated both the perceived educational usefulness of the debate-based learning format and differences in debate quality across three instruction-tuned models, without measuring objective learning outcomes. Although no statistically significant differences were observed when models were compared using aggregated ratings, feature-level analysis revealed differences in specific evaluation criteria, particularly future use intention. Overall, the findings suggest that locally deployable LLMs can support debate-based learning experiences while maintaining accessibility and independence from external cloud services. While the results should be interpreted cautiously due to the exploratory nature of the study and the limited

sample size, the prototype provides a foundation for future research on easily deployable AI-supported debate environments and their role in higher education.

**Acknowledgments.** This work was supported by EU under the Erasmus+ project PATH-AI (grant 2025-1-PL01-KA220-HED-000359600), and by the NextGenerationEU - BIOHCI - uniri-iz-25-198 grant.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Aryan, P.: LLMs as debate partners: utilizing genetic algorithms and adversarial search for adaptive arguments. In: 2024 IEEE Conference on Engineering Informatics (ICEI), pp. 1–11 (2024). https://doi.org/10.1109/ICEI64305.2024.10912343
2. Bi, Z., Lu, M., Li, Y., Roy, S., Guan, W., Ziyadi, M., Wang, X.: OPTAGENT: optimizing multi-agent LLM interactions through verbal reinforcement learning for enhanced reasoning. In: Proceedings of the 14th International Joint Conference on NLP and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pp. 1713–1728. The Asian Federation of NLP and The Association for Computational Linguistics (2025). https://doi.org/10.18653/v1/2025.findings-ijcnlp.107
3. Choi, H.K., Zhu, X., Li, S.: When identity skews debate: anonymization for bias-reduced multi-agent reasoning (2026). https://doi.org/10.48550/arXiv.2510.07517
4. Chu, T.: Multi-agent debate in L2 argumentative writing: an exploratory study of Chinese EFL learners. In: Proceedings of the International CALL Research Conference, pp. 21–28 (2025). https://doi.org/10.29140/97817637116240-03
5. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24 (2024). https://doi.org/10.5555/3692070.3692537
6. Guo, K., Zhong, Y., Li, D., Chu, S.K.W.: Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation. Comput. Educ. **203**, 104862 (2023). https://doi.org/10.1016/j.compedu.2023.104862
7. Hu, Z., Chan, H.P., Li, J., Yin, Y.: Debate-to-write: a persona-driven multi-agent framework for diverse argument generation. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 4689–4703. Association for Computational Linguistics, Abu Dhabi, UAE (2025). https://aclanthology.org/2025.coling-main.314/
8. Khan, A., et al.: Debating with more persuasive LLMs leads to more truthful answers. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24 (2024). https://doi.org/10.5555/3692070.3693020
9. Matsuda, I., et al.: Knowing ourselves through others: reflecting with AI in digital human debates (2025). https://doi.org/10.48550/arXiv.2511.13046
10. Septiyanti, N., Luthfi, M., Darmawansah, D.: Effect of chatbot-assisted learning on students' learning motivation and its pedagogical approaches. Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika 10(1), 69–77 (2024). https://doi.org/10.23917/khif.v10i1.4246

11. Zeng, Y., et al.: S2-MAD: Breaking the token barrier to enhance multi-agent debate efficiency. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics, pp. 9393–9408. Association for Computational Linguistics (2025). https://doi.org/10.18653/v1/2025.naacl-long.475
12. Zhang, Z., Sun, B., An, P.: Breaking barriers or building dependency? Exploring team-LLM collaboration in AI-infused classroom debate. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25, ACM, New York, NY, USA (2025). https://doi.org/10.1145/3706598.3713853

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.