Creating Tabletop RPG Dialogues via Retrieval-Augmented Generation

Gabriel Rudan Sales Matos[1], José Wellington Franco da Silva[2], Artur de Oliveira da Rocha Franco[1], José Gilvan Rodrigues Maia[1], José Antônio Fernandes de Macêdo[1]

[1] Programa de Pós-Graduação em Ciência da Computação (MDCC)
Departamento de Computação – Universidade Federal do Ceará (UFC)
Fortaleza – CE – Brasil

[2]Departamento de Computação – Universidade Federal do Ceará (UFC)
Campus Crateús – Crateús – CE – Brasil

gabrielrudan@alu.ufc.br, wellington@crateus.ufc.br, arturfranco@ufc.br

gilvanmaia@virtual.ufc.br, jose.macedo@dc.ufc.br

***Abstract. Introduction**: Tabletop RPGs (TRPGs) rely heavily on narrative, but Game Masters face challenges creating coherent and engaging dialogues while managing extensive rulebooks. **Objective**: This work investigates the adoption of Retrieval-Augmented Generation (RAG) to improve dialogue generation in TRPGs using Large Language Models. **Methodology or Steps**: Four dialogues were analyzed: two handwritten and two generated by LLMs—with and without RAG. Eleven participants rated them across five criteria—Engagement, Coherence, Cohesion, Creativity, and Surprise. **Results**: RAG-based generation outperformed standard LLMs across all categories, improving coherence (+0.18), cohesion (+0.46), creativity (+0.55), engagement (+0.91), and surprise (+0.64). Compared to handwritten dialogues, generated ones were rated higher in cohesion (3.91 vs. 3.68) and matched in coherence (3.82), although handwritten dialogues remained superior in engagement (3.82 vs. 3.09) and creativity (3.68 vs. 3.14).*
***Keywords** TRPG, Dialogue Generation, RAG, LLM.*

1. Introduction

Tabletop role-playing games (TRPGs) are renowned for their capacity to provide immersive experiences and collaborative narratives, wherein players and Game Masters (GMs) collectively construct plentiful and dynamic worlds [Barton e Stacks 2019]. The GM plays a pivotal role in this process since she is responsible for portraying non-player characters (NPCs), describing settings, and arbitrating the game rules [Bowman 2007]. Consequently, dialogues assume a critical function, as it is through these interactions that players unravel enigmas, obtain cues for their quests, and engage with the game world [Cover 2014]. The GM's ability to generate coherent and engaging dialogues not only facilitates the comprehension of missions and supports players' decision-making processes, but it is also fundamental in sustaining the narrative pace. It is therefore concluded that the dialogue dynamics establish interactions crucial for immersing players in the game universe.

However, the complex scalability of TRPGs – with multiple editions, expansions, and rule supplements – imposes considerable challenges on the GM, who may need to consult numerous sources to ensure the narrative's cohesion and adherence to the specific TRPG's official rules. The difficulty in accessing and integrating such information directly impacts the narrative fluidity and the overall quality of the gameplay experience [Mäyrä 2017]: the GM is frequently compelled to interrupt the session for consulting rulebooks or resort to improvisation, which can lead to inconsistencies or decisions lacking adequate substantiation. Furthermore, traditional rapid reference systems still require a high degree of familiarity with the material, thereby impeding the adoption of new players and game masters.

Recent advances in Artificial Intelligence, particularly with Large Language Models (LLMs), have opened novel possibilities to assist Game Masters. Nonetheless, these models face significant challenges in incorporating and updating new information without necessitating a complete retraining [Kaplan et al. 2020].

The conventional solution for updating LLMs entails the complete retraining of the model to assimilate new information. This process, however, is highly costly in computational terms, demanding specialized hardware such as high-performance GPUs [Bai et al. 2024] [1]. In addition to the cost, the time required for a new training session can range from weeks to months, rendering the frequent incorporation of new RPG rules and expansions unfeasible [Patil e Gudivada 2024]. Another critical problem is the limitation of tokens [Xue et al. 2023] [2] in inference time: even if a model is updated, it still has a context limit that restricts the amount of information it can process simultaneously, leading to losses in coherence and difficulties in retrieving specific rules.

Within this context, the Retrieval Augmented Generation (RAG) [Lewis et al. 2020] approach emerges as a viable alternative, as it allows an LLM to dynamically access updated knowledge bases, thereby ensuring coherent responses aligned with the most recent RPG rules [Gao et al. 2023]. The present study investigates the use of the RAG technique to aid Game Masters in generating cohesive narratives and dialogues, thereby reducing the need for manual consultations of rulebooks. To evaluate the efficiency of the proposed approach, an assessment based on crowdsourcing [Chiu et al. 2014] will be employed, considering five essential categories for the gameplay experience: engagement, coherence, cohesion, creativity, and surprise [Moser e Fang 2014].

This work is organized as follows: Section 2 presents the key concepts and related works; Section 3 discusses the application of the RAG technique in assisting Game Masters; Section 4 details the validation process for the generated dialogues; and finally, Section 5 discusses the results and outlines directions for future research.

[1] Studies indicate that the training of a model like GPT-3 can cost between 4 and 12 million dollars [Kaplan et al. 2020].

[2] In the context of LLMs, a "token" is a basic unit of text processing. It may represent a word, a part of a word, or even a character, depending on the tokenization employed by the underlying model. The number of tokens is an important factor for the efficiency and computational cost of LLMs, as the model processes and generates text in token units.

2. Related Work / State of the Art

The generation of dialogues for NPCs in RPGs has been the focus of several studies aiming to improve the coherence and relevance of interactions with AI-based systems [da Rocha Franco et al. 2024]. Among the main techniques explored are context sensitivity [Luu et al. 2024], fine-tuning [Lin et al. 2024], knowledge graphs [Meyer et al. 2023], and, more recently, Retrieval-Augmented Generation (RAG).

[Csepregi 2021] investigated the impact of incorporating context sensitivity in LLMs for dialogue generation. The study showed that adding contextual information improves coherence and fluidity but requires continuous input updates, making it less scalable for dynamic games. [van Stegeren e Myśliwiec 2021] applied fine-tuning to personalize NPC dialogues, achieving good alignment with specific narratives. However, this approach demands frequent retraining to adapt to changes in game content, which limits its scalability in games with regular expansions.

Knowledge graphs were explored by [Ashby et al. 2023] to dynamically adapt dialogues to narrative shifts. Although effective in enhancing immersion, this method depends on manual construction and maintenance of the graphs, reducing scalability when new variables emerge. [Nananukul e Wongkamjan 2024] proposed combining GPT-4 with knowledge graphs built from structured sources like Wikis. Despite generating detailed dialogues, the approach tends to generalize NPC personalities, resulting in overly positive or simplified character behaviors, even when nuanced traits are expected.

Unlike context sensitivity, fine-tuning, or knowledge graphs—which rely on static data, handcrafted structures, or retraining—RAG allows dynamic retrieval of information without modifying the underlying model. This enhances flexibility, scalability, and maintainability. Furthermore, while prior works focus mainly on English-language RPGs, this paper advances the field by applying RAG to generate dialogues in Brazilian Portuguese (PT-BR), addressing a language and cultural gap rarely explored in this research area [Rolim 2023].

To enhance the understanding of the state-of-the-art works in comparison with the study presented herein, Table 1 provides a schematic comparative summary. In this overview, the related works are highlighted according to their applicability based on foundations in LLMs, competence in generating dialogues in PT-BR, capacity for generalization of results (GR)[3], quality in generating continuous conversations (CC)[4], and the employment of enhancement techniques, such as Context Sensitivity (SC), Fine-Tuning (FT), and Knowledge Graphs (GC), as well as, evidently, the RAG method.

Compared to previous approaches, our work stands out for (i) enabling dialogue generation without retraining, (ii) supporting multiple RPG contexts by dynamically retrieving rule-based or narrative information, (iii) delivering competitive results in coherence and cohesion—even surpassing human-written dialogues in cohesion—and (iv) pioneering the development of dialogue systems for TRPGs in Brazilian Portuguese.

[3]The generalization of results refers to the model's ability to operate across different RPG systems, without being confined to a single set of rules or setting. Some works are developed for a specific RPG, limiting their applicability to other styles and game systems.

[4]In this context, "continuous conversations" refer to dialogues composed of multiple interactions between the user and the generating model, in contrast with approaches that limit the generation to a single response for an initial interaction.

Table 1. State of the Art: schematic summary.

Works	LLM	PT-BR	GR	CC	SC	FT	GC	RAG
[Csepregi 2021]	✓	-	-	✓	✓	-	-	-
[van Stegeren e Myśliwiec 2021]	✓	-	-	-	-	✓	-	-
[Ashby et al. 2023]	✓	-	✓	-	-	✓	✓	-
[Nananukul e Wongkamjan 2024]	✓	-	✓	✓	✓	-	✓	-
This Work	✓	✓	✓	✓	✓	-	-	✓

Unlike works that optimize single-turn responses or require rigid graph structures, our RAG-based approach supports multi-turn, fluid conversations with scalable adaptability.

3. Methodology

This paper proposes the application of the RAG technique for generating immersive dialogues in the missions of the RPG *Ordem Paranormal*[5], chosen due to its narrative appeal and popularity. The RAG technique integrates a document retrieval model, responsible for searching for pertinent text excerpts using *embeddings*[6], with a dialogue generation model, which employs this information to generate well-founded responses [Wang et al. 2024].

In addition to implementing the RAG technique, this work included validation experiments using the crowdsourcing method, involving online communities of players active on Reddit and Discord to evaluate the quality of the generated dialogues. The use of crowdsourcing is essential to ensure a diversified and collaborative evaluation [Blohm et al. 2013], as it enables the collection of insights from actual players, who represent the target audience of *Ordem Paranormal*. This approach is particularly valuable since the direct involvement of an active community ensures that the dialogues meet the players' expectations [Daniel et al. 2018].

Unlike previous studies on dialogue generation for digital RPGs or NPCs [Nananukul e Wongkamjan 2024], [Ashby et al. 2023], which often evaluate models using static datasets or pre-defined metrics (such as BLEU or ROUGE), our approach adopts a user-centered evaluation based on subjective criteria—Engagement, Coherence, Cohesion, Creativity, and Surprise—collected from real tabletop RPG players. While we acknowledge the value of those prior works, their baselines are not directly comparable to our context, as they target different genres (e.g., digital RPGs or video game NPCs) and rely on objective measures that may not capture the nuanced narrative quality required in TRPGs. Therefore, we opted for a qualitative comparison between human-written and LLM-generated dialogues (with and without RAG), which better reflects the narrative expectations in tabletop settings.

[5]Ordem Paranormal is a Brazilian RPG created by Rafael Lange and his team, set in a universe of mystery, supernatural elements, and investigation. The game combines immersive narrative elements with classic TRPG mechanics. More information can be found on the official website: `https://ordemparanormal.com.br/`.

[6]*Embeddings* are vectors that represent data numerically, allowing for the retrieval of relevant information from the database based on semantic similarities [Singh et al. 2023].

3.1. Process Overview

Figure 1 illustrates the six main steps for applying the RAG technique in this study. This structure is composed of four central components: the User (interacting with the system), the Game Master (the algorithm responsible for orchestrating the retrieval and the dialogue generation), the Vector DB (a vector database containing the RPG rules converted into *embeddings*), and the LLM (the language model used for generating responses). The following describes the steps detailing the interaction flow among these elements:

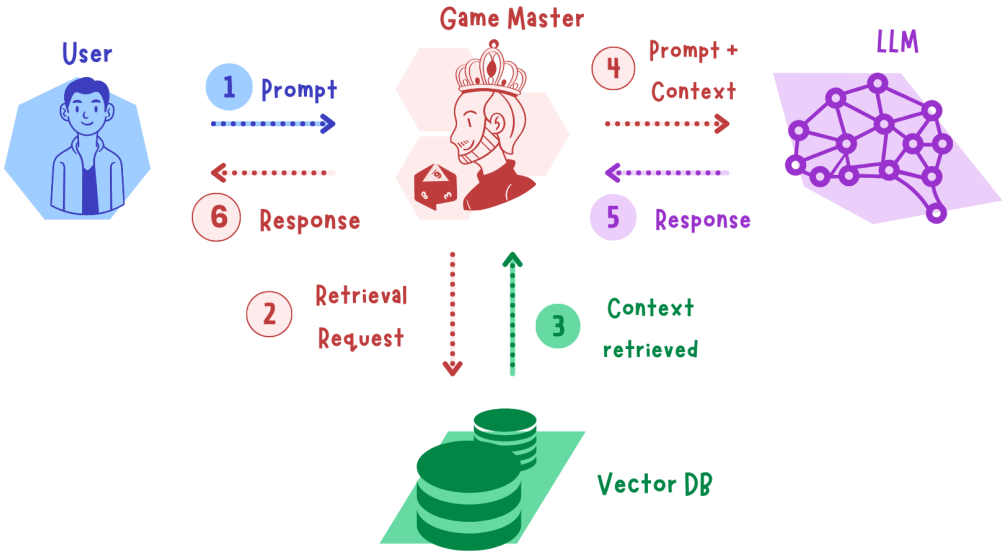

Figure 1. The primary structure of the RAG technique as applied to dialogue generation.

- **Step 1**: The User initiates the interaction with the Game Master by submitting a *prompt*, i.e., a textual input provided to the language model to generate a response or to perform a specific task.
- **Step 2**: The Game Master processes the *prompt* and makes a request to the Vector DB to retrieve excerpts from the rulebook which are pertinent to the User's query.
- **Step 3**: The Vector DB converts the *prompt* into an *embedding* and uses this vector to locate the two most similar *embeddings* stored. The retrieval is performed through a comparison of vector proximities, selecting the *embeddings* with the highest similarity to the embedding generated from the *prompt*, thereby ensuring a relevant context for the interaction.
- **Step 4**: The Game Master combines the original *prompt* with the retrieved context, forming a new *contextualized prompt* that is sent to the LLM.
- **Step 5**: The LLM processes the *contextualized prompt* and generates a response based on the information from the rulebook.
- **Step 6**: The Game Master receives the response generated by the LLM and presents it to the User, thus concluding the interaction.

3.2. Initial Proposition

At the start of the interaction, players submit a proposition to the game master, who acts as an NPC [Tychsen et al. 2005]. This free text defines the mission and the characters'

roles while following the RPG rules. The input provided by the players is used to create dynamic and contextualized dialogues.

Starting from the initial proposition, the LangChain system[7] is configured to generate dialogues grounded in the game's rules. This framework was chosen due to its extensive use in developing systems with LLMs, as well as its comprehensive and diverse online documentation, which facilitates its implementation and customization. The model processes the input until a termination command, such as "exit", is provided, at which point the model adjusts its behavior to act as the game master.

3.3. Large Language Model

The model employed for generating the dialogues is *GPT4o*, chosen for its capacity to handle complex contexts and extended dialogues—an essential requirement for RPGs [Nye et al. 2023]. Compared to *GPT-3*, *GPT4o* offers enhanced processing capabilities and flexibility in continuous dialogues [Lin et al. 2023], which is crucial for ensuring immersion in the RPG universe. The development was carried out using *Jupyter Notebook* [Kluyver et al. 2016], with *Google Colab* [Bisong e Bisong 2019] employed to facilitate the execution of the experiments.

The model receives the players' input along with the context retrieved from the vector database, generating coherent responses that are in line with the RPG universe.

3.4. Implementation Aspects

To facilitate the retrieval of information, integration with *OpenAIEmbeddings* was used, a functionality that converts text into numerical vectors, enabling semantic representation of textual content [Xian et al. 2024]. This process transforms the "Ordem Paranormal" Rulebook into vectors that semantically represent its content. The rulebook was divided into sections, each indexed and stored as individual documents using LangChain's `Document` class.

The embeddings were generated using the `text-embedding-3-small` model from OpenAI, chosen for its balance between performance and efficiency. These vectorized documents were then stored and indexed in *Qdrant* [8], another library chosen for its popularity and extensive online documentation, which utilizes cosine similarity as the primary metric [Singh et al. 2023]. Qdrant serves as the vector database, storing all embedded excerpts from the rulebook [9].

When a new query is submitted, the system encodes it into a vector and performs a similarity search to retrieve the two most relevant excerpts from the database ($k = 2$), balancing context richness and token limits [Ray 2023]. These excerpts are concatenated and inserted into a fixed prompt template: "Use the context below to answer the question." This augmented prompt is then sent to the LLM, allowing the model to generate contextually grounded and coherent dialogue responses based on knowledge from the "Ordem Paranormal" Rulebook.

[7]LangChain is a framework designed for building applications based on language models. For more details, see: `https://docs.langchain.com`.

[8]*Qdrant* is a vector database engine that facilitates the search and storage of vectors. More details can be found at: `https://qdrant.tech`

[9]For details regarding the complete implementation, please consult the documentation available at this link

4. Experimental Evaluation & Results

We conducted an online crowdsourcing study to evaluate the quality of dialogues generated by the GPT-4.0 model using the RAG technique, comparing them with dialogues written by humans. To do this, we created four distinct dialogues: dialogue A was generated automatically without the use of RAG; dialogue C was generated with RAG; dialogue B was written entirely by humans; and dialogue D was written by humans but with the help of LLM, aiming to test the collaboration between the model and human creativity.

Dialogues B and D were written by two university students who have extensive knowledge of the game and its setting, both of whom have experience as both Game Masters of Ordem Paranormal and as players participating in the table. In this way, they were able to develop dialogues that authentically and plausibly represent the narrative and dynamics of a real RPG game, providing an adequate level of fidelity to the original material. Table 2 shows the arrangement of characteristics of each of the four dialogues.

Table 2. Schematic summary of dialogues for validation

Dialogues	Generated	Written	With RAG	Without RAG
Dialogue A	✓	-	-	✓
Dialogue B	-	✓	-	-
Dialogue C	✓	-	✓	-
Dialogue D	-	✓	-	-

The study consisted of 26 questions[10], developed by the authors of this work in collaboration with a small group of experienced players of the RPG "Ordem Paranormal". These questions assessed five main categories – the definitions for the evaluation categories were based on the Aurélio Dictionary and the studies of [Csepregi 2021] and [Tapscott et al. 2018], to measure key aspects of a good dialogue, as indicated in the literature:

- **Engagement**: How much the dialogue immersed the participant in the story.
- **Coherence**: How consistent the dialogue is with the RPG content.
- **Cohesion**: How internally cohesive the dialogue is within its narrative information.
- **Creativity**: How creative the dialogue is.
- **Surprise**: How surprising the dialogue was.

The study included 11 participants, all of whom were encouraged to read and evaluate the four dialogues without any indication of their origin. Respondents were not given any information regarding which dialogues were AI-generated and which were human-written. Additionally, the dialogues were presented anonymously and in random order in the questionnaire, ensuring that participants had no clues about the authorship of each text. This approach guaranteed an unbiased evaluation, as evaluators only knew that some dialogues had been generated automatically and others written manually, but

[10]Refer to the documentation for details on the questions and the four dialogues at this link. It is worth noting that both the dialogue texts and the questions are in Brazilian Portuguese, as the evaluators are Brazilian.

had no means of identifying which belonged to each category. It is worth noting that the sample of participants in the experiment was small, which may limit the generalization of the results.

It is crucial that participants remain unaware of the origin of the dialogues (whether AI-generated or human-written) to avoid biases that may influence their evaluations of the AI model [Cozman e Kaufman 2022]. According to Cozman and Kaufman, when evaluators have prior knowledge of text authorship, their perceptions can be affected by conscious or unconscious biases, thus compromising the impartiality of the results.

Most participants (approximately 64%) reported having advanced knowledge of the "Ordem Paranormal" RPG, and an additional 18% had previously played or acted as Game Masters. This high level of familiarity with the system may have positively influenced the evaluation of the dialogues, as participants were more attuned to the nuances of the game's narrative style and mechanics. However, it may also have introduced bias, as prior expectations or preferences could affect how generated content was perceived.

4.1. Phase 1

In the first phase of the study, participants evaluated the dialogues in each of the five categories using a 5-point Likert scale (ranging from 1, meaning "Very poor", to 5, meaning "Very good"). Based on these responses, the weighted average of each dialogue in each category was calculated:

$$Weighted\ Average = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

where:

- x_i are the values to be weighted;
- w_i are the weights corresponding to each value;
- n is the total number of values.

Figure 2 presents a comparison between the average ratings of the manually written and generated dialogues in each category.

As shown in the chart, the generated dialogues matched the manually written ones in coherence (both scoring 3.82) and surpassed them in cohesion (3.91 vs. 3.68). These two categories are essential for maintaining a logical flow and structural consistency in TRPG narratives, indicating that LLMs can effectively support Game Masters by generating well-organized and consistent storylines.

However, human-written dialogues remained superior in engagement (3.82 vs. 3.09), creativity (3.68 vs. 3.14), and surprise (3.41 vs. 3.14), highlighting the current limitations of LLMs in replicating human emotional nuance and narrative originality. These findings suggest that while LLMs are promising tools for generating coherent and cohesive content, they may still require human refinement to achieve emotionally resonant and captivating storytelling.

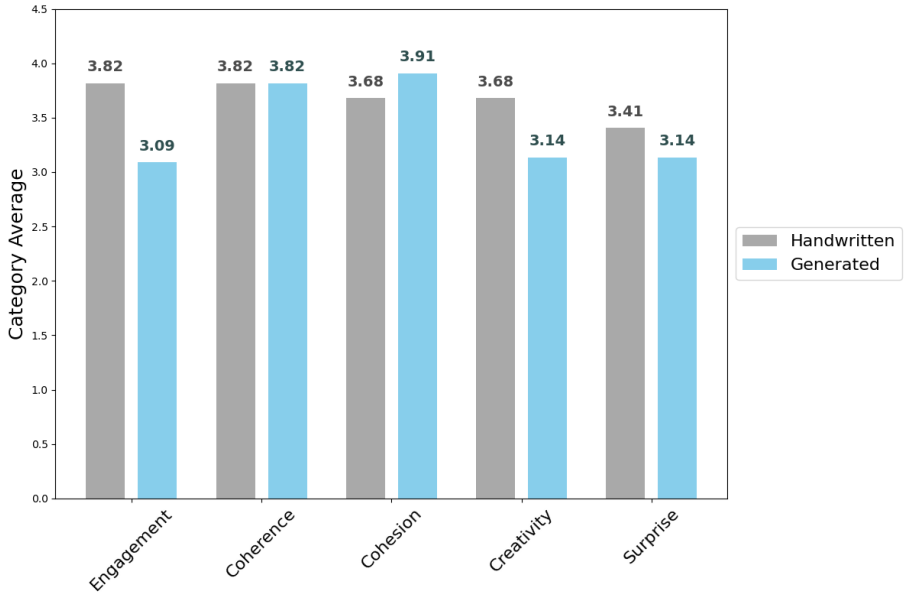

Figure 2. Comparison of dialogue evaluations.

4.2. Phase 2

In the second phase, we aimed to identify which automatically generated dialogue was best rated in each evaluation category. Figure 3 shows the distribution of participant preferences for each dialogue.

The dialogue generated with RAG (Dialogue C) clearly outperformed the one generated without RAG (Dialogue A) in four out of five categories: engagement (27.3% versus 9.1%), coherence (27.3% versus 18.2%), creativity (27.3% versus 18.2%), and surprise (36.4% versus 18.2%). In the Cohesion category, both dialogues performed similarly.

These results indicate that the use of RAG provides a noticeable benefit to dialogue generation in TRPGs, especially in aspects related to narrative richness and contextual alignment. The improvement can be attributed to the fact that RAG grants the language model access to precise information about the RPG system—such as character archetypes, thematic tone, and lore—enabling the generation of more grounded and contextually appropriate responses. In contrast, the dialogue without RAG lacked access to these semantic cues, leading to more generic and less immersive content. This reinforces the importance of retrieval mechanisms when generating text in specific fictional universes.

When compared to the manually written dialogues, participants tended to favor human-created content in categories like Engagement and Creativity. These dialogues were perceived as more emotionally compelling and original, largely because the authors were already familiar with the "Ordem Paranormal" universe. This prior knowledge enabled them to craft narratives that were both consistent with the setting and creatively surprising, adding personal flair that current LLMs still struggle to emulate.

Overall, while LLMs equipped with RAG showed strong potential for supporting Game Masters—especially in automating coherent and context-sensitive interactions—they still fall short in reproducing the emotional nuance and ingenuity

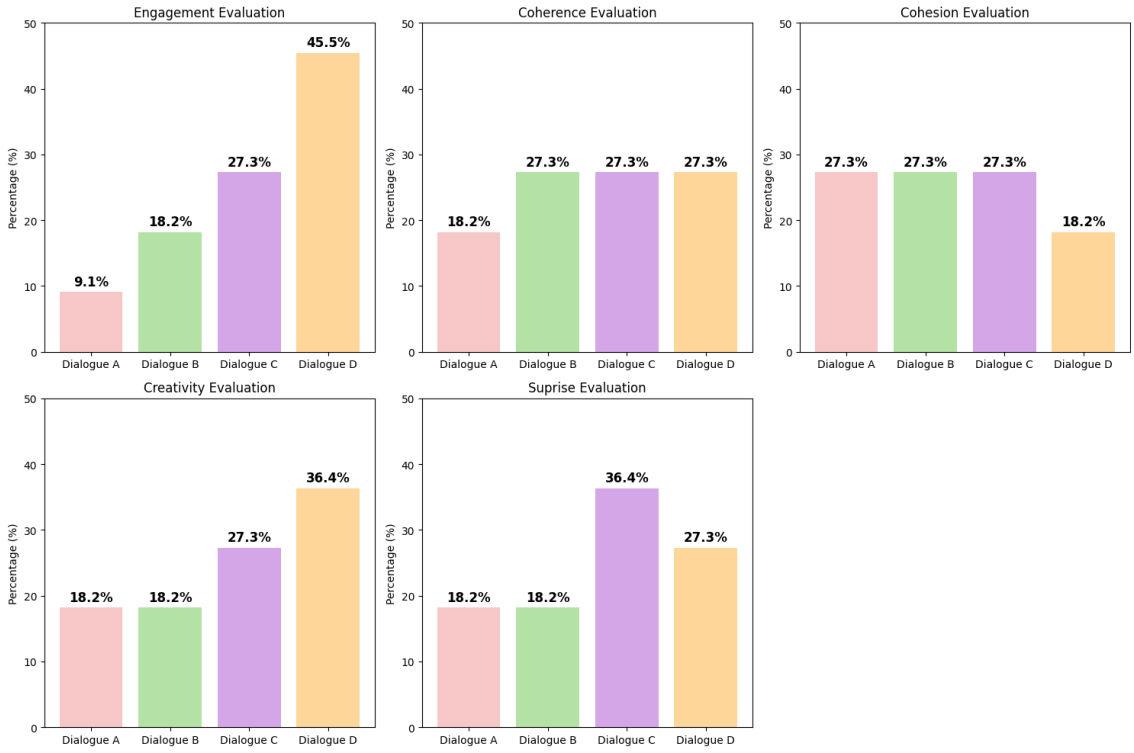

Figure 3. Comparison of best performance for each dialogue by category.

typical of experienced human storytellers.

5. Concluding Remarks

This work contributes to the advancement of dialogue generation for tabletop RPGs in Brazilian Portuguese, a context that remains underexplored in both academic and technical domains. By employing LLMs in combination with RAG, we explored how these technologies can support Game Masters in producing more immersive and coherent narratives. The findings suggest that, while current models can approximate human-level performance in structural aspects such as coherence and cohesion, they still face notable limitations in capturing emotional depth, originality, and the personal nuances that characterize engaging human storytelling. This gap highlights the importance of continued research in augmenting creative and affective capacities in LLMs.

Despite the promising outcomes, the results are exploratory and should be interpreted with caution due to the limited number of participants and the potential bias introduced by their familiarity with the RPG used. Future work will aim to expand the number and diversity of participants, including those with varying levels of RPG experience, and to test the method across different tabletop RPG systems to assess generalization. We also recommend embedding generated dialogues within actual RPG sessions to more directly assess player engagement and narrative impact. Exploring real-time adaptation and hybrid approaches combining model generation with human input may further enhance usability in dynamic gameplay scenarios.

References

Ashby, T., Webb, B. K., Knapp, G., Searle, J., e Fulda, N. (2023). Personalized quest and dialogue generation in role-playing games: A knowledge graph-and language model-based approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., et al. (2024). Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*.

Barton, M. e Stacks, S. (2019). *Dungeons and desktops: The history of computer role-playing games 2e*. AK Peters/CRC Press.

Bisong, E. e Bisong, E. (2019). Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64.

Blohm, I., Leimeister, J. M., e Krcmar, H. (2013). Crowdsourcing: How to benefit from (too) many great ideas. *MIS quarterly executive*, 12(4).

Bowman, S. L. (2007). The psychological power of the role-playing experience. *Journal of Interactive Drama*, 2(1):1–15.

Chiu, C.-M., Liang, T.-P., e Turban, E. (2014). What can crowdsourcing do for decision support? *Decision Support Systems*, 65:40–49.

Cover, J. G. (2014). *The creation of narrative in tabletop role-playing games*. McFarland.

Cozman, F. G. e Kaufman, D. (2022). Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. *Revista USP*, (135):195–210.

Csepregi, L. M. (2021). The effect of context-aware llm-based npc conversations on player engagement in role-playing video games. *Unpublished manuscript*.

da Rocha Franco, A. d. O., de Carvalho, W. V., da Silva, J. W. F., Maia, J. G. R., e de Castro, M. F. (2024). Managing and controlling digital role-playing game elements: A current state of affairs. *Entertainment Computing*, 51:100708.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., e Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., e Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., e Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al. (2016). Jupyter notebooks–a publishing format for reproducible computational workflows. In *Positioning and power in academic publishing: Players, agents and agendas*, pages 87–90. IOS press.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lin, J. C., Younessi, D. N., Kurapati, S. S., Tang, O. Y., e Scott, I. U. (2023). Comparison of gpt-3.5, gpt-4, and human user performance on a practice ophthalmology written examination. *Eye*, 37(17):3694–3695.

Lin, X., Wang, W., Li, Y., Yang, S., Feng, F., Wei, Y., e Chua, T.-S. (2024). Data-efficient fine-tuning for llm-based recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 365–374.

Luu, Q. K., Deng, X., Van Ho, A., e Nakahira, Y. (2024). Context-aware llm-based safe control against latent risks. *arXiv preprint arXiv:2403.11863*.

Mäyrä, F. (2017). Dialogue and interaction in role-playing games. *Dialogue across Media*, 28:271.

Meyer, L.-P., Stadler, C., Frey, J., Radtke, N., Junghanns, K., Meissner, R., Dziwis, G., Bulert, K., e Martin, M. (2023). Llm-assisted knowledge graph engineering: Experiments with chatgpt. In *Working conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow*, pages 103–115. Springer Fachmedien Wiesbaden Wiesbaden.

Moser, C. e Fang, X. (2014). Narrative structure and player experience in role-playing games. *International Journal of Human-Computer Interaction*, 31:146–156.

Nananukul, N. e Wongkamjan, W. (2024). What if red can talk? dynamic dialogue generation using large language models. *arXiv preprint arXiv:2407.20382*.

Nye, B. D., Mee, D., e Core, M. G. (2023). Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In *LLM@ AIED*, pages 78–88.

Patil, R. e Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074.

Ray, P. P. (2023). Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Rolim, F. M. (2023). Llms e ia generativa em jogos.

Singh, P. N., Talasila, S., e Banakar, S. V. (2023). Analyzing embedding models for embedding vectors in vector databases. In *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, pages 1–7. IEEE.

Tapscott, A., León, C., e Gervás, P. (2018). Generating stories using role-playing games and simulated human-like conversations. In *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018)*, pages 34–42.

Tychsen, A., Hitchens, M., Brolund, T., e Kavakli, M. (2005). The game master. In *ACM International Conference Proceeding Series*, volume 123, pages 215–222.

van Stegeren, J. e Myśliwiec, J. (2021). Fine-tuning gpt-2 on annotated rpg quests for npc dialogue generation. In *Proceedings of the 16th International Conference on the Foundations of Digital Games*, pages 1–8.

Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., et al. (2024). Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.

Xian, J., Teofili, T., Pradeep, R., e Lin, J. (2024). Vector search with openai embeddings: Lucene is all you need. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1090–1093.

Xue, F., Fu, Y., Zhou, W., Zheng, Z., e You, Y. (2023). To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36:59304–59322.