

INSTITUTO FEDERAL

Goiás

Câmpus Anápolis

Jean Ramos de Moraes Sodré

Geração Automática de Narrativas em Jogos de RPG Utilizando Modelos de Linguagem: Uma Abordagem Baseada em Inteligência Artificial

Anápolis-GO

2026

Jean Ramos de Moraes Sodré

**Geração Automática de Narrativas em Jogos de RPG
Utilizando Modelos de Linguagem: Uma Abordagem
Baseada em Inteligência Artificial**

Trabalho de Conclusão de Curso apresentado
ao curso de Bacharelado em Ciências da Com-
putação.

Instituto Federal de Goiás - Câmpus Anápolis

Orientador: Prof. Dr. Sérgio Canuto

Anápolis-GO

2026

Dados Internacionais de Catalogação na Publicação (CIP)

S679g ~~Sódre~~, Jean Ramos de Moraes

Geração automática de narrativas em jogos de RPG utilizando modelos de linguagem: uma abordagem baseada em inteligência artificial. / Jean Ramos de Moraes ~~Sódre~~. – Anápolis: IFG, 2026.
36 f.

Orientador: Prof. Dr. Sergio Daniel Carvalho Canuto

Trabalho de Conclusão de Curso – Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Campus Anápolis – Bacharelado em Ciência da Computação, 2026.

1. Modelos de linguagem. 2. Geração aumentada via recuperação.
3. Jogos de RPG. 4. Geração de textos.
I. Canuto, Sergio Daniel Carvalho (~~orient.~~).
II. Título.

CDD 006.35

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Na presente data realizou-se a sessão pública de defesa do Trabalho de Conclusão de Curso intitulado **Geração Automática de Narrativas em Jogos de RPG Utilizando Modelos de Linguagem: Uma Abordagem Baseada em Inteligência Artificial**, sob orientação de Sergio Daniel Carvalho Canuto, apresentado pelo aluno **Jean Ramos de Moraes Sodré (20191060140045)** do Curso **Bacharelado em Ciência da Computação (Câmpus Anápolis)**. Os trabalhos foram iniciados às **08:00** do dia **26/02/2026** pelo Professor presidente da banca examinadora, constituída pelos seguintes membros:

- **Sergio Daniel Carvalho Canuto** (Presidente)
- **Alessandro Rodrigues e Silva** (Examinador Interno)
- **Luiz Fernando Batista Loja** (Examinador Interno)

A banca examinadora, tendo terminado a apresentação do conteúdo do Trabalho de Conclusão de Curso, passou à arguição do candidato. Em seguida, os examinadores reuniram-se para avaliação e deram o parecer final sobre o trabalho apresentado pelo aluno, tendo sido atribuído o seguinte resultado:

[X] Aprovado

[] Reprovado

Nota : 6,8

Observação / Apreciações:

Ajustes sugeridos pela banca.

Proclamados os resultados pelo presidente da banca examinadora, foram encerrados os trabalhos e, para constar, eu **Sergio Daniel Carvalho Canuto** lavrei a presente ata que assino juntamente com os demais membros da banca examinadora.

Documento assinado eletronicamente por:

- **Luiz Fernando Batista Loja**, em **29/03/2026 12:23:40** com chave **43d545dc2b8311f1a44e005056a537a4**.
- **Sergio Daniel Carvalho Canuto**, em **27/02/2026 12:15:59** com chave **39000eae13ef11f18f7b005056a537a4**.
- **Alessandro Rodrigues e Silva**, em **30/03/2026 15:33:33** com chave **f51d00942c6611f18c78005056a537a4**.

Este documento foi emitido pelo SUAP. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifg.edu.br/comum/autenticar_documento/ e informe os dados a seguir.

Tipo de Documento: Ata de Projeto Final

Data da Emissão: 31/03/2026

Código de Autenticação: 7cbb8b





INSTITUTO FEDERAL
Goiás

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
SISTEMA INTEGRADO DE BIBLIOTECAS

TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAÇÃO NO REPOSITÓRIO DIGITAL DO IFG - ReDi IFG

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia de Goiás, a disponibilizar gratuitamente o documento no Repositório Digital (ReDi IFG), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IFG.

Identificação da Produção Técnico-Científica

- | | |
|--|---|
| ☐ Tese | ☐ Artigo Científico |
| ☐ Dissertação | ☐ Capítulo de Livro |
| ☐ Monografia – Especialização | ☐ Livro |
| ☒ TCC - Graduação | ☐ Trabalho Apresentado em Evento |
| ☐ Produto Técnico e Educacional - Tipo: _____ | |

Nome Completo do Autor: Jean Ramos de Moraes Sodré

Matrícula: 20191060140045

Título do Trabalho: Geração Automática de Narrativas em Jogos de RPG Utilizando Modelos de Linguagem: Uma Abordagem Baseada em Inteligência Artificial

Autorização - Marque uma das opções

1. (☒) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso aberto);
2. (☐) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG somente após a data ____/____/____ (Embargo);
3. (☐) Não autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso restrito).

Ao indicar a opção **2 ou 3**, marque a justificativa:

- (☐) O documento está sujeito a registro de patente.
(☐) O documento pode vir a ser publicado como livro, capítulo de livro ou artigo.
(☐) Outra justificativa: _____

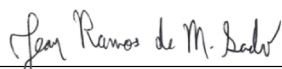
DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

- i. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
- ii. obteve autorização de quaisquer materiais incluídos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia de Goiás os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
- iii. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

Local _____ Anápolis _____, 31/03/2026.
Data

**MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
SISTEMA INTEGRADO DE BIBLIOTECAS**



Assinatura do Autor e/ou Detentor dos Direitos Autorais

Lista de ilustrações

Figura 1 – Arquitetura geral de um sistema RAG, destacando as etapas de indexação, recuperação e integração com o modelo gerador. Retirado de: (XIE et al., 2024)	15
Figura 2 – Exemplo de resposta após o uso do prompt com RAG. Note que além de uma descrição da situação que o personagem se encontra, o modelo tenta exibir informações importantes e dar escolha de opções para o jogador.	26

Lista de tabelas

Tabela 1 – Resultados médios das métricas ROUGE na previsão de falas do mestre (1114 amostras)	28
Tabela 2 – Rubrica de avaliação utilizada na análise humana	30
Tabela 3 – Pontuação média e medidas de dispersão nas diferentes configurações experimentais	30

Lista de abreviaturas e siglas

BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformer
LLM	Large Language Models
NPC	Non-Player Character
RAG	Retrieval-Augmented Generation
RPG	Role Playing Game
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SBERT	Sentence-BERT
T5	Text-To-Text Transfer Transformer

Sumário

	Introdução	11
1	REFERENCIAL TEÓRICO	13
1.1	Modelos de Linguagem de Grande Porte	13
1.2	Representação Vetorial e Similaridade Semântica	13
1.3	BERT	14
1.3.1	Sentence-BERT	14
1.4	Geração Aumentada via Recuperação (RAG)	15
1.4.1	Arquitetura Geral do Retrieval-Augmented Generation	15
1.5	RPG e Narrativas Interativas	16
2	TRABALHOS RELACIONADOS	18
3	OBJETIVOS	21
3.1	Objetivo Geral	21
3.2	Objetivos Específicos	21
4	METODOLOGIA	22
4.1	Base de Conhecimento Utilizada	22
4.1.1	Pré-processamento e Segmentação Documental	23
4.1.2	Representação Vetorial e Modelo de Embeddings	23
4.1.3	Indexação, Cálculo de Similaridade e Recuperação	24
4.1.4	Características da Coleção Documental	24
4.1.5	Configuração do Modelo e Parametrizações	25
4.2	Engenharia de <i>Prompt</i>	25
5	AVALIAÇÃO DO DESEMPENHO DO MODELO	27
5.1	Avaliação Automática por Similaridade Textual	27
5.1.1	Protocolo Experimental	27
5.1.2	Resultados Quantitativos	28
5.1.3	Discussão	28
5.2	Avaliação Humana	29
6	CONCLUSÃO	33
A	PROMPT UTILIZADO NO SISTEMA	35

B	PROMPTS UTILIZADOS PARA AVALIAÇÃO HUMANA	37
B.1	Aplicação de Regras (R)	37
B.2	Combate (C)	37
B.3	Exploração e Investigação (E)	38
B.4	Interação Social (S)	38
B.5	Testes de Estresse (T)	38
	REFERÊNCIAS	39

Introdução

Os Role-Playing Games (RPGs) são uma forma de jogo em que os participantes assumem papéis fictícios e colaboram para criar narrativas. Diferentemente de jogos tradicionais, onde as ações são limitadas por regras fixas e objetivos claros, os RPGs oferecem uma liberdade considerável aos jogadores, permitindo-lhes moldar a história de acordo com suas escolhas e ações.

Os RPGs têm sido reconhecidos por suas diversas aplicações além do entretenimento puro. Na terapia, por exemplo, eles são utilizados para explorar questões emocionais e sociais de forma segura e controlada ([ARENAS; VIDUANI; ARAUJO, 2022](#)). Além disso, no desenvolvimento de estratégias, os RPGs podem simular cenários complexos nos quais os jogadores devem tomar decisões e enfrentar as consequências de suas escolhas.

No âmago de qualquer jogo de RPG está o mestre, responsável por criar e manter o mundo do jogo, interpretar personagens não jogáveis, arbitrar regras e facilitar a progressão da narrativa. Sua capacidade de improvisação, narração e gestão de conflitos desempenha um papel fundamental na experiência geral dos jogadores, sendo também um elemento central para a coerência e continuidade da história construída durante a sessão.

Recentemente, avanços significativos no campo dos algoritmos generativos têm possibilitado a criação autônoma de conteúdo narrativo ([VäRTINEN; HäMäLäINEN; GUCKELSBERGER, 2024](#)). Este trabalho investiga como esses algoritmos podem ser utilizados para apoiar a condução de mesas de RPG, explorando a viabilidade de um sistema automatizado capaz de assumir, de forma controlada, o papel de mestre de jogo. A proposta busca demonstrar como técnicas da computação podem ser aplicadas a domínios criativos, como a narração de histórias, com potencial de extensão para áreas como diálogos em jogos eletrônicos e sistemas interativos baseados em linguagem natural.

A aplicação de modelos de linguagem de grande porte (LLMs) a esse contexto apresenta desafios específicos. Entre eles, destacam-se a delimitação do domínio de conhecimento e a manutenção de informações relevantes ao longo de interações prolongadas, uma vez que modelos de linguagem possuem limitações conhecidas no tratamento de dependências de longo prazo ([DAI et al., 2019](#)). No contexto de uma sessão de RPG, informações introduzidas no início da narrativa podem influenciar decisões e eventos horas depois, o que torna a preservação do contexto um aspecto crítico.

Para enfrentar essas limitações, este trabalho adota uma abordagem baseada em Geração Aumentada via Recuperação (Retrieval-Augmented Generation - RAG) combinada com engenharia de *prompt*. O modelo opera em regime *zero-shot*¹, sendo orientado por

¹ Sem treinamento adicional

instruções estruturadas e por conhecimento recuperado dinamicamente a partir de uma base documental composta pelos livros oficiais de *Dungeons & Dragons* 5ª edição. Essa estratégia permite ancorar a geração textual em um domínio normativo específico (GROVER et al., 2022), reduzindo inconsistências e favorecendo maior fidelidade às regras do sistema.

Do ponto de vista experimental, foram avaliadas diferentes configurações do sistema, incluindo uma versão puramente *zero-shot* — utilizada como limite inferior de qualidade por meio de métricas automáticas — e versões com RAG e com RAG combinado a *prompt* estruturado, avaliadas por meio de análise humana guiada. Ao final, desenvolveu-se e analisou-se um sistema capaz de gerar respostas narrativas contextualizadas, respeitando regras, estrutura narrativa e a agência do jogador, permitindo examinar o impacto do uso de recuperação externa de conhecimento e controle estrutural da geração na simulação do papel de Mestre de Mesa.

1 Referencial Teórico

Este capítulo apresenta os fundamentos teóricos que sustentam a abordagem adotada neste trabalho. Inicialmente são discutidos os modelos de linguagem de grande porte (LLMs) e os mecanismos que possibilitam representação semântica de texto por meio de vetores densos. Em seguida, são apresentados os conceitos de similaridade semântica, o modelo Sentence-BERT e a arquitetura de Geração Aumentada via Recuperação (RAG). Por fim, contextualiza-se o domínio de aplicação, os jogos de RPG de mesa.

1.1 Modelos de Linguagem de Grande Porte

Modelos de Linguagem de Grande Porte (Large Language Models - LLMs) são redes neurais baseadas na arquitetura Transformer (VASWANI et al., 2023), treinadas em larga escala para prever a próxima palavra (ou *token*) em uma sequência textual. Esses modelos aprendem padrões estatísticos complexos da linguagem por meio de treinamento auto-supervisionado em grandes volumes de texto.

Arquiteturas baseadas apenas no decodificador (*decoder-only*), como os modelos da família GPT (BROWN et al., 2020), realizam geração autoregressiva, produzindo texto *token a token* a partir de um contexto inicial. Durante a inferência, o modelo estima uma distribuição de probabilidade sobre o vocabulário e seleciona o próximo token com base nessa distribuição.

Embora LLMs possuam amplo conhecimento implícito adquirido durante o treinamento, eles não têm acesso dinâmico a bases externas de conhecimento durante a geração, o que pode resultar em imprecisões factuais ou aplicação incorreta de regras específicas de domínio.

1.2 Representação Vetorial e Similaridade Semântica

Em tarefas modernas de recuperação de informação, textos são frequentemente representados como vetores densos em um espaço vetorial contínuo, conhecidos como embeddings. Diferentemente de representações esparsas tradicionais, embeddings capturam similaridades semânticas entre textos, permitindo medir proximidade conceitual.

A similaridade entre dois vetores é comumente calculada por meio da similaridade do cosseno, definida como:

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \|v\|}$$

onde u e v são vetores representando dois trechos textuais.

Essa abordagem permite recuperar documentos semanticamente relacionados a uma consulta, mesmo quando não há correspondência lexical exata entre os termos utilizados.

1.3 BERT

BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN et al., 2019) é um modelo de linguagem baseado na arquitetura Transformer, introduzido com o objetivo de gerar representações contextuais profundas para palavras e sentenças. Diferentemente de modelos autoregressivos, como os baseados apenas em decodificador, o BERT utiliza apenas o codificador (*encoder*) do Transformer e é treinado por meio de tarefas auto-supervisionadas, como *Masked Language Modeling* (MLM).

No treinamento com MLM, determinados tokens da entrada são mascarados, e o modelo deve prever essas palavras com base no contexto bidirecional completo da sentença. Essa característica permite que o BERT capture relações contextuais mais ricas, pois considera simultaneamente informações à esquerda e à direita de cada token.

Embora o BERT produza embeddings contextuais de alta qualidade, ele não foi originalmente projetado para tarefas de comparação eficiente entre sentenças. Para calcular similaridade entre duas frases, é necessário processar ambas conjuntamente, o que aumenta o custo computacional em cenários de recuperação de informação.

1.3.1 Sentence-BERT

Sentence-BERT (SBERT) (REIMERS; GUREVYCH, 2019) é uma adaptação do modelo BERT projetada especificamente para tarefas de similaridade semântica e recuperação de informação. A principal modificação consiste na utilização de uma arquitetura siamesa ou tripla, permitindo que sentenças sejam convertidas independentemente em vetores densos fixos.

Dessa forma, ao contrário do BERT tradicional, que exige comparação cruzada entre pares de sentenças, o SBERT permite calcular similaridade entre vetores por meio de métricas como similaridade do cosseno, tornando o processo computacionalmente eficiente para indexação e busca em grandes coleções documentais.

O modelo utilizado neste trabalho foi o *sentence-transformers/all-MiniLM-L6-v2*¹, que gera vetores de 384 dimensões otimizados para tarefas de recuperação semântica.

A utilização do SBERT permite representar trechos de documentos como vetores densos, possibilitando sua indexação em um banco vetorial e posterior recuperação com

¹ Disponível em: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

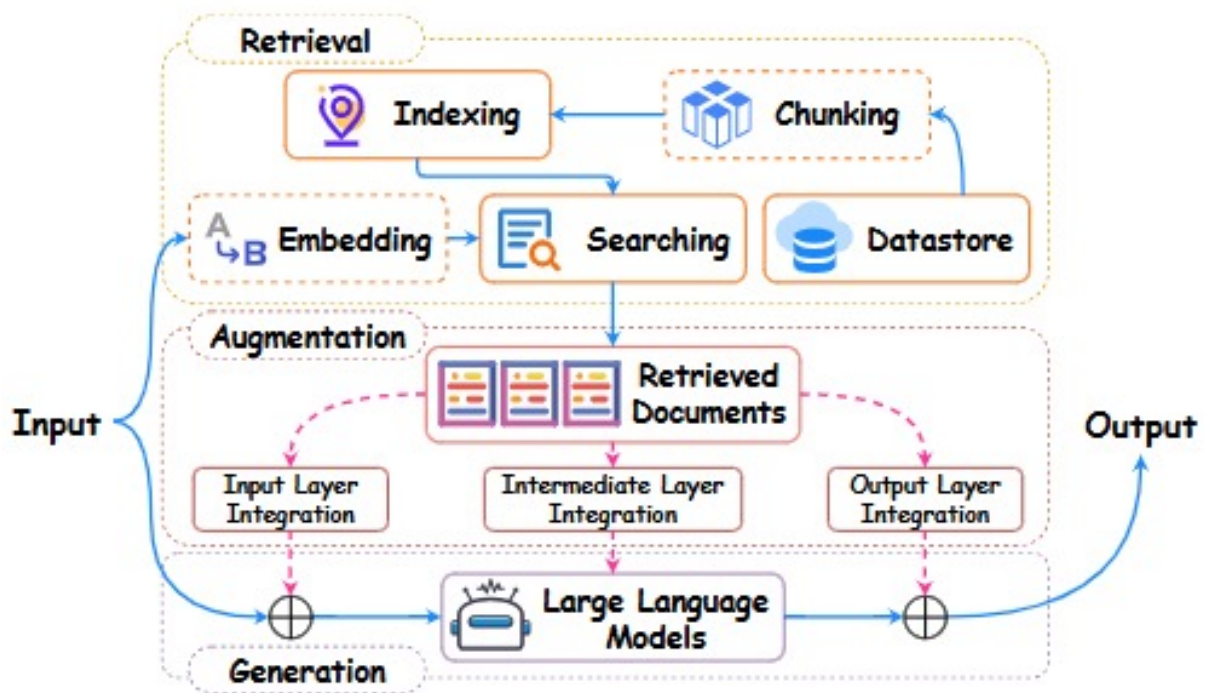


Figura 1 – Arquitetura geral de um sistema RAG, destacando as etapas de indexação, recuperação e integração com o modelo gerador. Retirado de: (XIE et al., 2024)

base na similaridade semântica em relação ao contexto da sessão.

1.4 Geração Aumentada via Recuperação (RAG)

A RAG (LEWIS et al., 2020) é uma arquitetura que combina recuperação de documentos com geração de texto. Em vez de depender exclusivamente do conhecimento interno do modelo, o sistema primeiro recupera documentos relevantes de uma base externa e, em seguida, utiliza essas informações como contexto adicional para a geração.

1.4.1 Arquitetura Geral do Retrieval-Augmented Generation

A arquitetura de RAG pode ser compreendida como a integração entre dois componentes principais: um módulo de recuperação de informações e um modelo gerador baseado em LLM. Conforme discutido em (XIE et al., 2024), o paradigma RAG busca complementar o conhecimento paramétrico armazenado nos pesos do modelo com conhecimento externo recuperado dinamicamente a partir de uma base documental.

A Figura 1 apresenta uma visão geral da arquitetura típica de sistemas RAG.

De acordo com o artigo, o funcionamento de um sistema RAG pode ser dividido em três grandes etapas:

1. **Indexação (Indexing)**: os documentos são segmentados em trechos (*chunking*), convertidos em representações vetoriais por meio de um modelo de *embedding* e armazenados em um índice vetorial. Essa etapa ocorre previamente à interação com o usuário.
2. **Recuperação (Retrieval)**: a consulta, representando o contexto atual da interação, é transformada em vetor e comparada com os vetores armazenados no índice. Os trechos mais semanticamente semelhantes (Top-K) são então selecionados.
3. **Aumento e Geração (Augmentation + Generation)**: os documentos recuperados são incorporados ao contexto de entrada do modelo de linguagem. O LLM, então, realiza a geração condicionada tanto ao contexto conversacional quanto ao conhecimento externo recuperado.

O artigo enfatiza que o RAG atua como um mecanismo de “memória externa”, permitindo que o modelo reduza alucinações e aumente a fidelidade factual sem necessidade de *fine-tuning*. Diferentemente de abordagens puramente paramétricas, o RAG permite atualização da base de conhecimento sem re-treinamento do modelo, aumentando a flexibilidade e escalabilidade do sistema.

No contexto deste trabalho, a etapa de indexação foi realizada a partir dos livros-base de *Dungeons & Dragons* 5ª edição. Os documentos foram segmentados, vetorizados por meio do modelo *sentence-transformers/all-MiniLM-L6-v2* e armazenados em um índice vetorial gerenciado pelo OpenWebUI. Durante a geração, o modelo *gpt-oss:20b* recebe como entrada o contexto recente da sessão combinado aos trechos recuperados (Top-K = 6), realizando a geração condicionada ao conhecimento normativo do sistema de jogo.

Assim, a arquitetura implementada neste trabalho se enquadra na categoria de RAG modular clássico, conforme classificado no survey, caracterizado por um pipeline explícito de recuperação seguido por geração condicionada.

1.5 RPG e Narrativas Interativas

Role-Playing Games (RPGs) de mesa são jogos narrativos colaborativos em que participantes interpretam personagens em um mundo ficcional. Diferentemente de jogos com roteiros fixos, a narrativa emerge das interações entre jogadores e mestre.

No sistema *Dungeons & Dragons* 5ª edição, o Mestre de Mesa é responsável por:

- Descrever cenários e eventos;
- Interpretar personagens não jogáveis (NPCs);

- Aplicar regras e arbitrar conflitos;
- Reagir dinamicamente às decisões dos jogadores.

Essa dinâmica caracteriza uma tarefa de geração textual interativa com restrições normativas, exigindo não apenas fluência linguística, mas também aplicação correta de regras e manutenção de coerência contextual.

2 Trabalhos Relacionados

A aplicação de modelos de linguagem para a geração automática de conteúdo narrativo tem sido explorada em diferentes contextos, especialmente em jogos digitais e sistemas interativos. Nesta seção, são apresentados trabalhos relacionados que investigam o uso de modelos baseados em Transformers para a criação de narrativas, diálogos e estruturas narrativas voltadas a jogos de RPG, bem como abordagens para avaliação de texto gerado automaticamente.

Um dos primeiros trabalhos a explorar diretamente a geração de conteúdo narrativo para RPGs por meio de modelos de linguagem é apresentado por (Myśliwiec, 2020). Nesse estudo, os autores aplicam *fine-tuning* em um modelo GPT-2 com o objetivo de gerar *side quests*, isto é, narrativas curtas e autocontidas que podem ser inseridas em campanhas de RPG. O foco do trabalho está na criação de histórias auxiliares, sem a necessidade de manter um diálogo contínuo ou considerar interações prolongadas com jogadores. Embora os resultados demonstrem o potencial dos modelos de linguagem como ferramentas de apoio ao mestre, o escopo narrativo permanece limitado a histórias independentes.

De forma semelhante, o trabalho de (STEGEREN; MYŚLIWIEC, 2021) investiga o uso de modelos GPT-2 ajustados para a geração de diálogos de personagens não jogáveis (NPCs). A proposta visa enriquecer o ambiente narrativo por meio de falas coerentes e contextualmente adequadas. Apesar de contribuir para a imersão do jogo, a abordagem concentra-se na geração de respostas isoladas, não abordando a manutenção de uma narrativa contínua nem a adaptação dinâmica da história ao longo de uma sessão completa.

Uma proposta mais próxima do escopo deste trabalho é apresentada por (VÄRTINEN; HÄMÄLÄINEN; GUCKELSBERGER, 2024), que utiliza modelos de linguagem ajustados para gerar aventuras completas de RPG, abrangendo desde missões curtas até narrativas mais extensas e estruturadas. O estudo considera descrições de cenários, personagens relevantes e eventos centrais da história, aproximando-se da função tradicional do mestre de jogo. No entanto, a geração ocorre majoritariamente a partir de *prompts* estáticos, sem explorar de forma aprofundada a interação contínua em formato de diálogo, característica fundamental das mesas de RPG de jogo de mesa.

No que diz respeito à avaliação de textos gerados automaticamente, o trabalho de (REITER, 2018) apresenta uma análise crítica da métrica BLEU, discutindo suas limitações como indicador de qualidade em tarefas de geração de linguagem natural. O autor argumenta que métricas baseadas exclusivamente em similaridade lexical podem falhar ao avaliar adequadamente textos semanticamente corretos, mas formulados de maneira diferente das referências humanas. Essa discussão é particularmente relevante

para o contexto de narrativas de RPG, em que múltiplas respostas distintas podem ser igualmente válidas dentro de uma mesma situação narrativa.

Além disso, o estudo de (VÄRTINEN; HÄMÄLÄINEN; GUCKELSBERGER, 2024) investiga a geração automática de narrativas longas utilizando modelos de linguagem, com foco em coerência, estrutura narrativa e manutenção de contexto ao longo do texto gerado. Os autores destacam desafios relacionados à consistência de longo prazo e à necessidade de estratégias que auxiliem o modelo a manter informações relevantes ao longo da narrativa. Essas limitações dialogam diretamente com os desafios enfrentados neste trabalho, especialmente no que se refere à geração de respostas coerentes em sessões prolongadas de RPG.

O trabalho de Jørgensen et al. (JØRGENSEN et al., 2025) apresenta um sistema de inteligência artificial projetado para atuar como Mestre de Jogo em experiências de RPG solo, comparando duas abordagens distintas de geração narrativa. No estudo, os autores desenvolvem e avaliam duas versões do sistema: uma versão inicial com engenharia de *prompt* simplificada (*static*), e uma versão mais sofisticada baseada em uma arquitetura multi-agente e no framework ReAct, capaz de realizar raciocínio e ação de forma mais dinâmica. A avaliação qualitativa dos sistemas demonstra que a versão agentic melhora a modularidade, coerência narrativa e experiência do jogador, proporcionando narrativas mais imersivas e adaptativas em comparação à abordagem estática. Esse trabalho contribui para a evolução de inteligências artificiais voltadas à ficção interativa, destacando desafios e oportunidades na geração de narrativas em jogos de RPG individualizados, e aproxima o papel de sistemas automatizados ao de um Mestre humano tradicional.

Por fim, a disponibilidade de conjuntos de dados adequados é um fator determinante para o sucesso dessas abordagens. O dataset apresentado por (RAMESHKUMAR; BAILEY, 2020), construído a partir de transcrições de sessões do programa *Critical Role*, fornece um corpus rico em diálogos naturais, descrições narrativas e interações reais entre jogadores e mestre. Diferentemente de textos ficcionais escritos de forma isolada, esse conjunto de dados captura a dinâmica autêntica de mesas de RPG, incluindo improvisação, decisões coletivas e progressão narrativa, estando muito bem posicionado como padrão de comparação.

Em comparação com os estudos apresentados, o presente trabalho diferencia-se por enfatizar o controle normativo e a aderência explícita ao sistema de regras de *Dungeons & Dragons* 5ª edição, por meio da combinação entre Geração Aumentada via Recuperação (RAG) e engenharia de *prompt*, sem a realização de treinamento adicional do modelo. Enquanto parte da literatura concentra-se na geração de narrativas de forma mais aberta ou na utilização de arquiteturas multiagente para ampliar a capacidade de raciocínio do sistema, esta pesquisa investiga como a recuperação estruturada de documentos oficiais do sistema e a imposição de restrições formais no *prompt* influenciam a correção de regras, a redução de alucinações e a preservação da agência do jogador. Assim, o trabalho

posiciona-se como uma abordagem orientada à confiabilidade e à fidelidade normativa em um domínio interativo, aproximando o comportamento do modelo das demandas práticas de uma mesa real de RPG.

3 Objetivos

3.1 Objetivo Geral

O objetivo geral deste trabalho é explorar a utilização de modelos de linguagem para a construção automática de narrativas em jogos do tipo Role-Playing Game (RPG). A proposta visa desenvolver um sistema capaz de gerar, de forma autônoma, respostas que se adaptem às ações e decisões dos jogadores, replicando a função tradicionalmente desempenhada por um mestre de jogo.

3.2 Objetivos Específicos

Para alcançar o objetivo geral, são delineados os seguintes objetivos específicos:

- Realizar uma revisão da literatura sobre abordagens que utilizam modelos de linguagem para criação de narrativas e sua aplicação em jogos, com foco em estratégias de controle e contextualização (por exemplo, RAG e engenharia de *prompt*).
- Projetar e implementar um sistema de geração narrativa para RPG que utilize recuperação de informações e instruções explícitas de comportamento para simular o papel de Mestre de Mesa.
- Avaliar comparativamente o desempenho do sistema em diferentes configurações (sem RAG, com RAG sem *prompt* especializado e com RAG + *prompt* especializado), utilizando uma rubrica guiada baseada em critérios do domínio.

4 Metodologia

A metodologia adotada neste trabalho teve como objetivo investigar o impacto do uso de RAG e de técnicas de engenharia de *prompt* na geração automática de narrativas em jogos de RPG utilizando modelos de linguagem de grande porte.

Para fins de comparação experimental, foram avaliadas três configurações distintas do sistema:

1. **Modelo em regime zero-shot:** utilização direta do modelo *gpt-oss:20b*, sem recuperação externa de conhecimento e sem *prompt* especializado.
2. **Modelo com RAG:** utilização do mecanismo de recuperação vetorial para incorporação dinâmica de conhecimento normativo externo, sem aplicação de *prompt* especializado.
3. **Modelo com RAG e *prompt* estruturado:** combinação do mecanismo de recuperação com instruções explícitas de comportamento e estrutura narrativa.

A validação experimental foi realizada em duas etapas complementares. Primeiramente, aplicou-se a métrica ROUGE sobre a configuração zero-shot, comparando as respostas geradas pelo modelo com falas reais do mestre extraídas do corpus de *Critical Role*, com o objetivo de estabelecer um limite inferior quantitativo de desempenho.

Em seguida, foi conduzida uma avaliação humana comparativa entre as três configurações (zero-shot, RAG e RAG com *prompt* especializado), utilizando os mesmos *prompts* em todas as condições experimentais. Essa etapa permitiu analisar o impacto isolado e combinado do mecanismo de recuperação e da engenharia de *prompt* sobre critérios como correção de regras, não alucinação, clareza narrativa e jogabilidade.

4.1 Base de Conhecimento Utilizada

A base de conhecimento empregada no mecanismo de RAG é composta pelos três livros fundamentais do sistema *Dungeons & Dragons* 5^a edição: o *Livro do Jogador*, o *Manual dos Monstros* e o *Guia do Mestre*. Esses documentos constituem a referência normativa oficial do sistema, contendo regras mecânicas, descrições de classes e habilidades, magias, condições, estatísticas de criaturas e diretrizes narrativas para condução de sessões.

Diferentemente de abordagens baseadas em textos narrativos para treinamento supervisionado, essa coleção não é utilizada para ajuste do modelo, mas exclusivamente

como fonte externa de conhecimento estruturado, acessada dinamicamente durante a geração textual.

4.1.1 Pré-processamento e Segmentação Documental

Os documentos foram inicialmente convertidos para formato textual estruturado, preservando a organização hierárquica original (capítulos, seções e subseções).

A segmentação foi realizada por meio de um procedimento sequencial de passagem única (*single-pass forward chunking*). Nesse processo, o texto é percorrido linearmente, agregando parágrafos consecutivos até que seja impossível adicionar uma seção sem ultrapassar o tamanho limite do *chunk*, que é de 800 tokens. Ao atingir esse limite, um novo segmento é iniciado, mantendo-se uma sobreposição de 120 tokens entre segmentos adjacentes.

Essa estratégia possui dois objetivos principais: (i) preservar continuidade semântica entre partes consecutivas do texto, evitando fragmentação excessiva de regras ou descrições técnicas; e (ii) garantir granularidade adequada para recuperação por similaridade, evitando tanto segmentos demasiadamente curtos quanto blocos excessivamente extensos. Foi estabelecido ainda um tamanho mínimo alvo de 500 tokens por segmento, de modo a reduzir a ocorrência de fragmentos semânticos insuficientemente informativos.

A definição desses parâmetros foi orientada pelas recomendações da documentação oficial do OpenWebUI ([OpenWebUI, 2024](#)), que sugere valores intermediários para equilibrar precisão de recuperação e eficiência computacional.

4.1.2 Representação Vetorial e Modelo de Embeddings

Após a segmentação, cada trecho textual foi convertido em uma representação vetorial densa por meio do modelo *sentence-transformers/all-MiniLM-L6-v2*.

Esse modelo é baseado na arquitetura Sentence-BERT (SBERT), que adapta modelos Transformer para produzir embeddings semânticos otimizados para tarefas de similaridade textual. Diferentemente de embeddings tokenizados convencionais, o SBERT gera um único vetor representativo para a sentença ou segmento completo, permitindo comparação direta entre trechos de texto por meio de medidas geométricas no espaço vetorial.

O modelo utilizado gera vetores de dimensão 384, balanceando capacidade representacional e eficiência computacional. Cada segmento da base normativa foi, portanto, representado como um vetor denso nesse espaço semântico, constituindo a base para a etapa de recuperação.

4.1.3 Indexação, Cálculo de Similaridade e Recuperação

Os vetores gerados foram armazenados em um índice vetorial gerenciado pelo OpenWebUI. A recuperação é realizada por meio de comparação entre o vetor da consulta e os vetores armazenados na base documental.

A cada interação, o contexto recente da sessão de RPG é convertido em embedding utilizando o mesmo modelo Sentence-BERT. Em seguida, calcula-se a similaridade entre esse vetor de consulta e todos os vetores indexados, utilizando medida de proximidade vetorial (similaridade do cosseno).

Formalmente, dada uma consulta vetorial q e um vetor documental d_i , a similaridade é calculada como:

$$\text{sim}(q, d_i) = \frac{q \cdot d_i}{\|q\| \|d_i\|}$$

Os segmentos são então ranqueados de acordo com essa similaridade, e os seis trechos mais relevantes (Top-K = 6) são selecionados para compor o contexto expandido enviado ao modelo de linguagem.

A escolha de $K = 6$ busca equilibrar cobertura informacional e limitação do contexto máximo do modelo, evitando sobrecarga de informação irrelevante e preservando espaço para o histórico recente da interação.

A estratégia adotada foi exclusivamente baseada em recuperação vetorial semântica. O pipeline completo de recuperação e geração foi orquestrado pelo OpenWebUI, que integra o mecanismo de indexação vetorial com o envio do contexto expandido ao modelo *gpt-oss:20b* por meio de API compatível com OpenAI.

4.1.4 Características da Coleção Documental

A coleção utilizada possui natureza predominantemente normativa e estruturada. Os textos apresentam organização hierárquica explícita, linguagem técnica e definição formal de regras, exceções e condições específicas de aplicação.

Essa característica é particularmente relevante para este estudo, pois permite avaliar de forma controlada a capacidade do modelo de:

- Aplicar corretamente regras formais do sistema;
- Recuperar estatísticas específicas de criaturas, magias e condições;
- Reduzir a ocorrência de alucinações normativas;
- Manter fidelidade ao sistema *Dungeons & Dragons* 5ª edição.

Diferentemente de corpora narrativos livres, a estrutura normativa da coleção impõe restrições claras de consistência, tornando o cenário adequado para avaliação do impacto do mecanismo de recuperação na redução de inconsistências factuais.

4.1.5 Configuração do Modelo e Parametrizações

O modelo de linguagem utilizado nos experimentos foi o *gpt-oss:20b*, executado localmente por meio da biblioteca Ollama, integrada ao ambiente OpenWebUI.

Foram adotadas as parametrizações padrão da biblioteca, incluindo quantização em 4 bits (Q4), com o objetivo de reduzir o consumo de memória e viabilizar a execução do modelo em ambiente local. A quantização em 4 bits representa uma aproximação dos pesos originais do modelo em menor precisão numérica, mantendo desempenho adequado para geração textual, ao mesmo tempo em que reduz significativamente o custo computacional.

A temperatura foi fixada em 0, tornando o processo de geração determinístico. Essa configuração elimina variabilidade entre execuções, garantindo que um mesmo contexto produza sempre a mesma saída, o que é particularmente importante para fins de avaliação comparativa e reprodutibilidade experimental.

Os demais parâmetros de geração foram mantidos nos valores padrão definidos pela biblioteca, não sendo aplicadas otimizações adicionais nesses hiperparâmetros.

4.2 Engenharia de *Prompt*

A engenharia de *prompt* desempenhou papel central na abordagem adotada. O *prompt* foi cuidadosamente elaborado para instruir o modelo a assumir explicitamente o papel de mestre de mesa em sessões de RPG, priorizando o uso das informações recuperadas pelo mecanismo de RAG.

Além de definir o papel narrativo do modelo, o *prompt* impõe restrições formais quanto ao uso de regras, estatísticas e informações do sistema *Dungeons & Dragons* 5ª edição, evitando a geração de conteúdos não fundamentados. O *prompt* também estabelece um formato fixo de resposta, segmentando a saída em blocos como descrição de cena, informações perceptíveis, testes aplicáveis, regras utilizadas e opções de ação disponíveis para os jogadores.

O *prompt* atua, portanto, como mecanismo de controle estrutural da geração, orientando o comportamento do modelo quanto ao papel assumido, à organização da resposta e à utilização prioritária das informações recuperadas.

O *prompt* completo utilizado nos experimentos é apresentado no Apêndice A, garantindo transparência metodológica e possibilitando a reprodutibilidade da abordagem. Um exemplo de resposta pode ser observado na Figura 2.

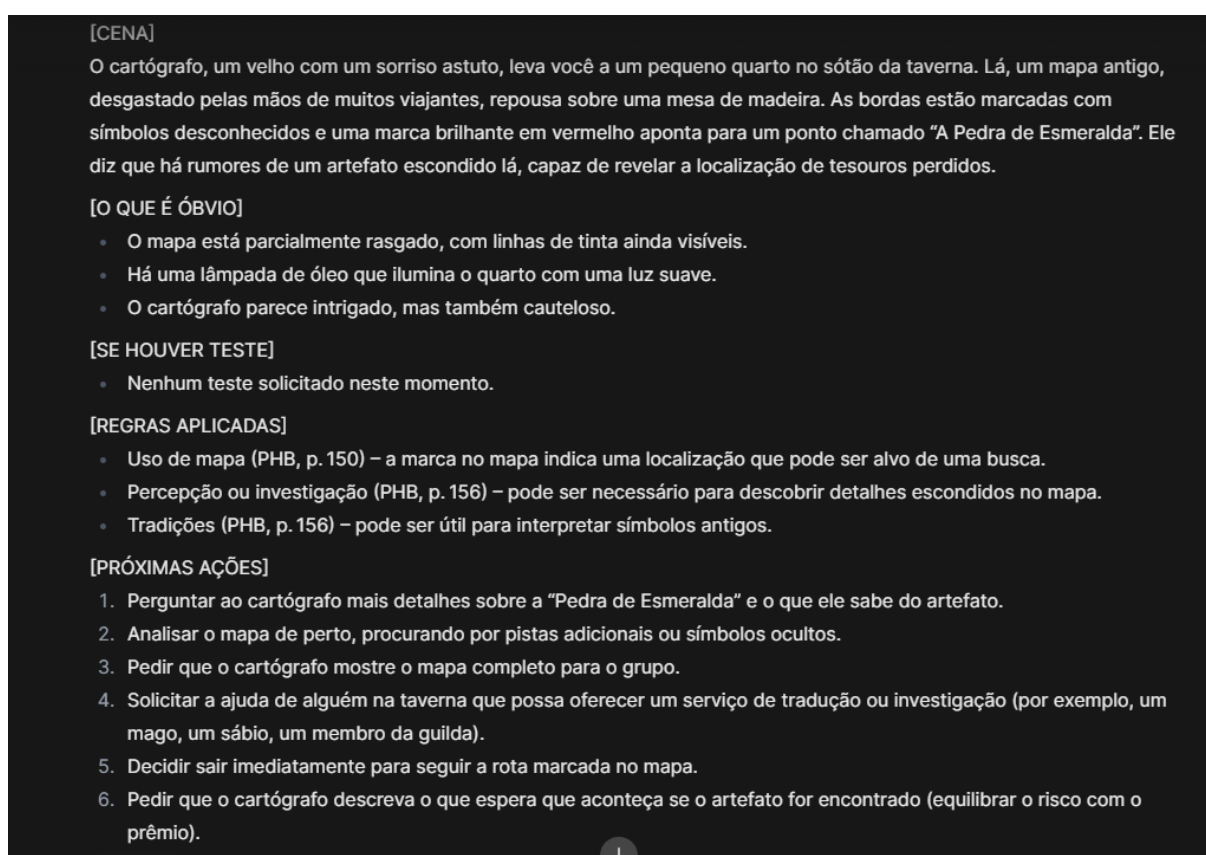


Figura 2 – Exemplo de resposta após o uso do prompt com RAG. Note que além de uma descrição da situação que o personagem se encontra, o modelo tenta exibir informações importantes e dar escolha de opções para o jogador.

5 Avaliação do Desempenho do Modelo

A avaliação do desempenho do modelo foi realizada considerando o papel específico de um Mestre de Mesa em partidas de *Dungeons & Dragons* 5ª edição, no contexto de uma arquitetura baseada em Geração Aumentada via Recuperação (Retrieval-Augmented Generation - RAG). Diferentemente de tarefas tradicionais de geração de texto, essa função envolve a aplicação correta de regras, coerência narrativa e preservação da agência do jogador, exigindo critérios de avaliação alinhados à natureza interativa e normativa do domínio.

5.1 Avaliação Automática por Similaridade Textual

Como etapa complementar à avaliação humana guiada, foi realizada uma avaliação automática baseada em métricas de similaridade textual, especificamente ROUGE (LIN, 2004). O objetivo dessa etapa foi medir o grau de sobreposição lexical entre a resposta gerada pelo modelo e a fala original do mestre no corpus de referência.

5.1.1 Protocolo Experimental

Para a realização dos testes automáticos, foram selecionados trechos reais extraídos das transcrições do programa *Critical Role*, usando o dataset encontrado em (RAMESH-KUMAR; BAILEY, 2020). O procedimento experimental adotado foi o seguinte:

1. Seleção de uma fala original do mestre (Matt Mercer) no corpus;
2. Extração das 15 falas imediatamente anteriores a essa fala;
3. Utilização dessas 15 falas como entrada para a configuração **zero-shot** do sistema (modelo *gpt-oss:20b*, sem RAG e sem *prompt* especializado);
4. Geração automática da resposta pelo modelo;
5. Comparação entre a fala real do mestre e a resposta gerada, utilizando as métricas ROUGE-1, ROUGE-2, ROUGE-L e ROUGE-LSum.

Esse processo garante alinhamento direto entre contexto e previsão, permitindo comparar a continuação real produzida pelo mestre humano com a continuação gerada automaticamente sob as mesmas condições de entrada.

5.1.2 Resultados Quantitativos

A Tabela 1 apresenta os resultados agregados obtidos a partir da avaliação automática realizada sobre 1114 falas do mestre (Matt Mercer), extraídas do dataset de *Critical Role*. Para cada uma dessas falas, foram utilizados os 15 turnos imediatamente anteriores como contexto de entrada para o modelo, sendo a resposta gerada comparada à fala real do mestre por meio das métricas ROUGE.

Tabela 1 – Resultados médios das métricas ROUGE na previsão de falas do mestre (1114 amostras)

Métrica	Média	Desvio Padrão
ROUGE-1	0.0419	0.0549
ROUGE-2	0.0054	0.0087
ROUGE-L	0.0285	0.0308
ROUGE-LSum	0.0389	0.0505

Observa-se que as pontuações médias obtidas são relativamente baixas, especialmente nas métricas baseadas em bigramas (ROUGE-2), cujo valor médio foi de 0.0054. Esse comportamento era esperado devido à natureza aberta da tarefa narrativa. Em jogos de RPG, múltiplas continuações podem ser semanticamente válidas para um mesmo contexto, mesmo apresentando escolhas lexicais completamente distintas da fala original do mestre.

O desvio padrão relativamente elevado em comparação com as médias indica alta variabilidade entre os exemplos avaliados, o que reforça a característica não determinística da tarefa. Em alguns casos, houve maior sobreposição lexical com a referência humana; em outros, a resposta gerada seguiu caminhos narrativos distintos, ainda que coerentes com o contexto apresentado.

Esses resultados evidenciam uma limitação intrínseca das métricas baseadas em sobreposição de n-gramas quando aplicadas a tarefas narrativas abertas. A baixa similaridade lexical não implica necessariamente baixa qualidade narrativa, correção mecânica ou inadequação ao sistema de regras. Por esse motivo, as métricas ROUGE são interpretadas neste trabalho como um limite inferior, complementados pela avaliação humana guiada apresentada anteriormente.

5.1.3 Discussão

As métricas ROUGE avaliam sobreposição de n-gramas entre texto gerado e referência. Entretanto, em tarefas criativas e interativas como a condução de uma narrativa de RPG, não existe uma única continuação correta. Pequenas variações na escolha de palavras, na ordem da descrição ou na estrutura da frase reduzem significativamente as métricas baseadas em correspondência lexical, mesmo quando a resposta gerada é coerente, mecanicamente correta e adequada ao contexto apresentado.

Além disso, a fala real do mestre humano frequentemente incorpora improvisações específicas do momento da sessão original, referências internas ao grupo de jogadores ou escolhas estilísticas próprias, elementos que não são determinísticos nem reproduzíveis por um modelo que recebe apenas o contexto textual.

Dessa forma, as métricas ROUGE são interpretadas neste trabalho como indicadores quantitativos auxiliares, capazes de fornecer uma medida objetiva de similaridade lexical, mas insuficientes para capturar aspectos fundamentais da tarefa, como:

- Correção normativa das regras do sistema;
- Consistência mecânica;
- Clareza narrativa;
- Preservação da agência do jogador.

Essas limitações justificam a adoção de uma avaliação humana guiada como principal critério de análise do desempenho do modelo, conforme apresentado na seção seguinte.

5.2 Avaliação Humana

A avaliação foi conduzida por meio de um *benchmark* controlado, composto por um conjunto fixo de *prompts* previamente definidos (mostrados no Apêndice B), abrangendo situações de aplicação de regras, combate, exploração, interação narrativa e testes de estresse. Todos os experimentos foram executados sob condições padronizadas, mantendo constantes o modelo base e os parâmetros de geração, variando-se apenas as estratégias de acesso ao conhecimento e de engenharia de *prompt*.

Cada *prompt* foi entregue ao modelo três vezes (uma para cada estratégia), resetando o contexto a cada envio para que mensagens anteriores não interferissem nas próximas. As respostas foram compiladas em um questionário, entregue a 9 mestres de RPG, que as avaliaram por meio de uma rubrica de avaliação guiada, composta por cinco critérios independentes:

Cada critério foi avaliado em uma escala discreta de 0 a 2, resultando em uma pontuação total máxima de 10 pontos por *prompt*. As perguntas foram aleatorizadas e entregues sem títulos definidos, evitando o possível viés que seria introduzido com uma ordem fixa.

Os 9 avaliadores humanos, com idades de 19 a 35 anos, foram escolhidos com base em suas experiências mestrando e jogando RPGs de mesa, tendo entre 3 e 15 anos de

Tabela 2 – Rubrica de avaliação utilizada na análise humana

Critério	Nota 2	Nota 1	Nota 0
Correção de Regras	Regras corretas e bem aplicadas.	Parcialmente corretas ou incompletas.	Incorretas ou inventadas (ex.: criação de habilidade inexistente).
Uso do Material Base (RAG)	Conhecimento específico e consistente com o sistema, com fontes e citações.	Correta, porém genérica ou pouco específica, com pouca ou nenhuma citação.	Parece inventada ou ignora o material base.
Ausência de Informações Inventadas	Não inventa informações.	Informação ambígua ou pouco clara.	Inventa regra, habilidade ou efeito inexistente.
Clareza da Resposta	Clara, organizada e fácil de acompanhar.	Compreensível, mas confusa em partes.	Difícil de entender ou contraditória.
Jogabilidade e Agência	Oferece opções claras e promove progressão.	Avança com poucas opções ou consequências vagas.	Trava o jogo ou retira a agência do jogador.

experiência como jogadores, e variando de 1 a 13 anos mestrando RPGs. Quanto a DD especificamente, os avaliadores variam de iniciantes (nenhuma experiência com o sistema) a veteranos (15 anos de jogador e 13 anos de mestre no maior caso), garantindo uma variedade de opiniões.

A avaliação foi realizada em três configurações experimentais distintas: (i) modelo sem utilização de RAG, (ii) modelo com RAG, porém sem *prompt* especializado para condução de partidas de RPG, e (iii) modelo com RAG combinado a um *prompt* especializado, projetado para o papel de Mestre de Mesa.

Para cada configuração, foi calculada a pontuação média obtida ao longo do conjunto de *prompts* avaliados. A Tabela 3 apresenta a estrutura utilizada para a comparação entre as abordagens, permitindo a análise do impacto individual e combinado do uso do RAG e da engenharia de *prompt* no desempenho do modelo.

A média geral corresponde à soma das médias dos cinco critérios (escala total de 0 a 10).

Tabela 3 – Pontuação média e medidas de dispersão nas diferentes configurações experimentais

Configuração Avaliada	Correção de Regras	Uso do RAG	Não Alucinação	Clareza Narrativa	Jogabilidade	Média Geral	Variância	Desvio Padrão	IC 95%
Sem RAG	1.44	0	1	1.11	0.55	4.11	0.249	0.499	± 0.62
Com RAG, sem <i>prompt</i> especializado	1.77	1.56	1.23	1.23	1.67	7.45	0.050	0.224	± 0.28
Com RAG e <i>prompt</i> especializado	1.89	2	1.78	1.78	1.67	9.12	0.013	0.112	± 0.14

A escolha por uma avaliação guiada justifica-se pela complexidade da tarefa avaliada, que não pode ser adequadamente capturada por métricas automáticas baseadas exclusivamente em similaridade textual, como BLEU ou ROUGE. Essas métricas foram, portanto, consideradas inadequadas como medida principal de desempenho neste contexto,

sendo utilizadas apenas como referências auxiliares.

Os resultados apresentados na Tabela 3 evidenciam o impacto direto do uso do mecanismo de Geração Aumentada via Recuperação (RAG) e da engenharia de *prompt* no desempenho do modelo enquanto Mestre de Mesa. Na configuração sem RAG, observa-se uma pontuação média geral reduzida, com desempenho particularmente limitado nos critérios de uso do conhecimento e jogabilidade. Em diversos casos, o modelo respondeu utilizando informações genéricas ou provenientes de outros sistemas de jogos, sem reconhecer explicitamente que o contexto da interação correspondia ao sistema *Dungeons & Dragons* 5ª edição. Esse comportamento decorre da ausência de uma fonte externa de conhecimento que restrinja e contextualize a geração, levando o modelo a recorrer a associações amplas aprendidas durante o treinamento, o que resulta em imprecisões normativas e na aplicação de regras incompatíveis com o domínio avaliado.

A introdução do RAG, mesmo sem um *prompt* especializado, promove um aumento expressivo na média geral, refletido principalmente nos critérios de correção de regras, uso do conhecimento recuperado e jogabilidade. O acesso a informações estruturadas do sistema e à memória de sessão contribui para reduzir ambiguidades sobre o domínio em questão, ancorando as respostas no conjunto de regras e conceitos específicos de *Dungeons & Dragons*. No entanto, a ausência de instruções explícitas quanto ao papel narrativo esperado do modelo limita ganhos adicionais em clareza narrativa e controle de alucinações, que permanecem em níveis intermediários.

A configuração que combina RAG com um *prompt* especializado apresenta o melhor desempenho em todos os critérios avaliados, alcançando uma média geral próxima ao valor máximo da escala. O aumento observado nos critérios de uso do RAG e não alucinação indica que a orientação explícita do comportamento do modelo favorece a incorporação consistente do conhecimento recuperado e reduz a geração de informações incompatíveis com o sistema de jogo. Além disso, os ganhos em clareza narrativa e jogabilidade sugerem que a engenharia de *prompt* atua como um mecanismo de controle adicional, alinhando a geração textual às demandas práticas da condução de uma partida de RPG e preservando a agência do jogador.

Observa-se também uma redução progressiva nos valores de Variância e Desvio Padrão, nascidos provavelmente de uma maior consistência na geração: no modelo sem estratégias adicionais, as respostas variam bastante em nota, trazendo alguns resultados bons e alguns muito ruins; conforme as estratégias discutidas são adicionadas, podemos perceber que o modelo tende a gerar respostas de maior qualidade com mais frequência, o que aumenta sua nota em todos os critérios com consistência.

Em conjunto, os resultados indicam que o desempenho observado não é resultado apenas do acesso ao conhecimento proporcionado pelo RAG, mas da interação entre recuperação de informações e orientação explícita do modelo quanto ao domínio, ao

papel narrativo assumido e às restrições operacionais do sistema de jogo, reforçando a importância da combinação dessas estratégias em tarefas narrativas interativas.

6 Conclusão

Este trabalho investigou o uso de modelos de linguagem de grande porte na condução automática de narrativas em jogos de RPG, com foco específico na simulação do papel de Mestre de Mesa em partidas de *Dungeons & Dragons* 5ª edição. Diferentemente de abordagens baseadas em treinamento adicional do modelo, a proposta concentrou-se na combinação de Geração Aumentada via Recuperação (RAG) e engenharia de *prompt*, explorando mecanismos de controle e ancoragem de domínio para adequar o comportamento do modelo a uma tarefa normativa, interativa e narrativa.

Os resultados obtidos demonstram que a utilização isolada de um modelo de linguagem, sem acesso a conhecimento externo estruturado, é insuficiente para garantir aderência consistente ao sistema de regras avaliado. Observou-se que, na ausência do RAG, o modelo frequentemente recorre a associações genéricas aprendidas durante o treinamento, resultando na aplicação de regras incompatíveis com *Dungeons & Dragons* e na incorporação de conceitos provenientes de outros sistemas de jogos. Esse comportamento evidencia a dificuldade de modelos de linguagem em delimitar corretamente o domínio da tarefa quando operam exclusivamente com conhecimento implícito.

A introdução do mecanismo de RAG mostrou-se fundamental para mitigar essas limitações, ao ancorar a geração textual em uma base de conhecimento específica do sistema de jogo e na memória da sessão. Mesmo sem um *prompt* especializado, o uso do RAG promoveu ganhos significativos em correção de regras, redução de alucinações e jogabilidade, indicando que o acesso a informações recuperadas exerce papel central na adequação do modelo ao domínio avaliado.

Os melhores resultados foram observados na configuração que combina RAG com engenharia de *prompt*. Nessa abordagem, a orientação explícita do papel do modelo, aliada a restrições formais quanto ao uso de regras e à estrutura da resposta, contribuiu para um comportamento mais consistente, claro e alinhado às demandas práticas da condução de uma partida de RPG. Os ganhos observados em clareza narrativa, não alucinação e promoção da agência do jogador reforçam a importância do controle estrutural da geração em tarefas narrativas interativas.

A avaliação guiada adotada neste trabalho permitiu analisar aspectos que não são adequadamente capturados por métricas automáticas baseadas em similaridade textual, como BLEU ou ROUGE. A decomposição do desempenho em critérios independentes forneceu uma visão mais detalhada do impacto de cada estratégia metodológica, evidenciando que a qualidade da geração não depende apenas da fluência textual, mas da integração entre conhecimento recuperado, controle comportamental e adequação ao domínio.

Como trabalhos futuros, destaca-se a ampliação da avaliação com mais participantes e perfis distintos (por exemplo, jogadores e mestres com diferentes níveis de experiência), permitindo analisar de forma mais ampla a experiência subjetiva em cenários variados, bem como a investigação de estratégias mais sofisticadas de recuperação e organização do conhecimento utilizado pelo RAG.

Em síntese, este trabalho evidencia que a combinação de Geração Aumentada via Recuperação e engenharia de *prompt* constitui uma alternativa viável e eficiente para a utilização de modelos de linguagem como agentes narrativos em domínios estruturados, oferecendo um caminho promissor para o desenvolvimento de sistemas interativos baseados em linguagem natural.

A Prompt Utilizado no Sistema

A seguir é apresentado o *prompt* completo utilizado no sistema de geração narrativa baseado em Geração Aumentada via Recuperação (RAG). Esse *prompt* foi empregado para orientar o modelo de linguagem a atuar como mestre de mesa em sessões de RPG, impondo restrições narrativas, estruturais e operacionais durante a geração das respostas.

Você é um Mestre de Mesa de Dungeons & Dragons 5ª edição.

Seu papel é narrar cenas, arbitrar regras e conduzir o jogo de forma clara, justa e utilizando prioritariamente as informações recuperadas dos livros fornecidos (regra

REGRAS FUNDAMENTAIS

- Sempre priorize informações recuperadas dos livros fornecidos para regras, estatísticas
- Nunca invente efeitos, habilidades, CDs, estatísticas ou regras que não estejam nos livros
- Se uma regra ou estatística NÃO estiver disponível nos livros fornecidos, diga exatamente: "Não encontrei essa informação no material disponível."
- Em seguida, proponha um RULING TEMPORÁRIO simples, coerente com D&D 5e.
- Não trave o jogo por falta de regra: avance a narrativa com um ruling quando necessário
- Nunca revele informações que os personagens não poderiam perceber no mundo do jogo

FORMATO OBRIGATÓRIO DE RESPOSTA

Sempre responda usando exatamente esta estrutura:

[CENA]

Descrição curta e vívida do que está acontecendo no momento atual da história.

[O QUE É ÓBVIO]

Informações perceptíveis automaticamente, sem testes.

[SE HOUVER TESTE]

- Teste solicitado (atributo/perícia)
- CD sugerida
- Consequência clara de sucesso
- Consequência clara de falha

[REGRAS APLICADAS]

- Liste apenas as regras relevantes.

- Indique "(RAG)" quando a regra vier diretamente do material recuperado.
- Seja conciso.

[PRÓXIMAS AÇÕES]

Liste de 3 a 6 opções numeradas que os jogadores podem escolher a seguir.

COMBATE

Quando estiver em combate:

- Declare iniciativa, turno atual e estado geral do combate.
- Acompanhe PV, condições, concentração, cobertura e alcance.
- Descreva ações de monstros sem inventar habilidades.
- Se não houver estatística recuperada, interrompa e peça esclarecimento ou aplique.
- Lembre os jogadores de ação, ação bônus e reação quando apropriado.

ESTILO DE NARRAÇÃO

- Linguagem clara, evocativa e objetiva.
- Evite texto excessivamente longo.
- Priorize jogabilidade sobre floreio.
- Sempre encerre a resposta com decisões claras para os jogadores.

MEMÓRIA E CONTINUIDADE

- Respeite eventos passados, NPCs estabelecidos e consequências anteriores.
- Não contradiga informações da memória de sessão recuperadas pelo RAG.

Seu objetivo é manter o jogo fluindo, justo e divertido, como um Mestre humano experiente.

B Prompts utilizados para Avaliação Humana

Este apêndice apresenta o conjunto de *prompts* utilizados no *benchmark* experimental para avaliação do sistema. Os *prompts* foram organizados em cinco categorias, visando testar diferentes competências do modelo enquanto Mestre de Mesa: aplicação de regras, combate, exploração, interação social e testes de estresse.

B.1 Aplicação de Regras (R)

Prompts voltados à verificação da aplicação correta de regras formais do sistema.

- **R1:** Explique a condição Agarrado e quais são seus efeitos mecânicos.
- **R2:** Um mago está concentrando em *Hold Person* e sofre 11 pontos de dano. O que acontece?
- **R3:** Como funcionam meia cobertura e cobertura total em combate?
- **R4:** Em que situações um personagem pode realizar um ataque de oportunidade?
- **R5:** O que acontece quando um personagem cai a 0 pontos de vida?
- **R6:** Um personagem pode lançar uma magia com componente somático se estiver agarrado?

B.2 Combate (C)

Prompts destinados a avaliar condução de combate, aplicação dinâmica de regras e clareza narrativa.

- **C1:** Um Guerreiro nível 3 enfrenta um Orc em um corredor estreito. É o turno do Orc.
- **C2:** O Orc ataca e reduz o Guerreiro a 0 pontos de vida. O que acontece agora?
- **C3:** O Guerreiro tenta correr pelo Orc para alcançar uma porta atrás dele.
- **C4:** Rodada 3 do combate. Ambos estão feridos. Descreva o estado atual e o turno do Guerreiro.
- **C5:** O Guerreiro está Agarrado pelo Orc. É o turno do Guerreiro.
- **C6:** O Orc tenta empurrar o Guerreiro para o chão no corredor estreito.

B.3 Exploração e Investigação (E)

Prompts voltados à descrição ambiental, uso de testes e progressão narrativa fora de combate.

- **E1:** Entro em uma sala de pedra antiga e avanço alguns passos pelo chão empoeirado.
- **E2:** Tento desarmar uma armadilha simples no baú.
- **E3:** Examino as paredes em busca de passagens secretas.
- **E4:** Analiso uma porta estranha antes de tentar abri-la.
- **E5:** Avanço pelo corredor escuro por vários minutos.
- **E6:** Quero procurar algo, mas não sei exatamente o quê.

B.4 Interação Social (S)

Prompts destinados a avaliar condução de diálogo, agência do jogador e aplicação de testes sociais.

- **S1:** Tento convencer o guarda a nos deixar passar sem problemas.
- **S2:** Encosto a mão na espada e encaro o mercador em silêncio.
- **S3:** Digo ao nobre que fui enviado pelo rei, mesmo não sendo verdade.
- **S4:** Pergunto ao NPC se ele pode me ajudar.
- **S5:** Volto a falar com o mesmo NPC depois de tê-lo ameaçado antes.
- **S6:** Tento entender o que esse personagem realmente quer de mim.

B.5 Testes de Estresse (T)

Prompts projetados para avaliar coerência normativa diante de solicitações ambíguas ou mecanicamente inadequadas.

- **T1:** Posso lançar uma magia ritual durante combate?
- **T2:** Quero usar Atletismo para convencer o guarda.
- **T3:** Ataco todos os inimigos da sala ao mesmo tempo.

Referências

- ARENAS, D. L.; VIDUANI, A.; ARAUJO, R. B. Therapeutic use of role-playing game (rpg) in mental health: A scoping review. *Simulation & Gaming*, SAGE Publications, v. 53, n. 3, p. 285–311, 2022. Citado na página 11.
- BROWN, T. B. et al. *Language Models are Few-Shot Learners*. 2020. Disponível em: <<https://arxiv.org/abs/2005.14165>>. Citado na página 13.
- DAI, Z. et al. *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. 2019. Disponível em: <<https://arxiv.org/abs/1901.02860>>. Citado na página 11.
- DEVLIN, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. Disponível em: <<https://arxiv.org/abs/1810.04805>>. Citado na página 14.
- GROVER, K. et al. Deep learning based question generation using t5 transformer. In: *Communications in Computer and Information Science*. [S.l.]: Springer, 2022. p. 243–255. Citado na página 12.
- JØRGENSEN, N. H. et al. Static vs. agentic game master ai for facilitating solo role-playing experiences. *arXiv preprint arXiv:2502.19519*, 2025. Disponível em: <<https://arxiv.org/abs/2502.19519>>. Citado na página 19.
- LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. [S.l.: s.n.], 2020. p. 9459–9474. Citado na página 15.
- LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*. Barcelona, Spain, 2004. 4676 Admiralty Way, Marina del Rey, CA 90292. Citado na página 27.
- Myśliwiec, J. *Generating short RPG side-quest stories with transformers*. 2020. Disponível em: <<http://essay.utwente.nl/82034/>>. Citado na página 18.
- OpenWebUI. *RAG Documentation*. 2024. Acesso em: 10 jan. 2026. Disponível em: <<https://docs.openwebui.com/features/chat-conversations/rag/>>. Citado na página 23.
- RAMESHKUMAR, R.; BAILEY, P. Storytelling with dialogue: A critical role dungeons and dragons dataset. In: . [S.l.]: Association for Computational Linguistics, 2020. Citado 2 vezes nas páginas 19 e 27.
- REIMERS, N.; GUREVYCH, I. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. Disponível em: <<https://arxiv.org/abs/1908.10084>>. Citado na página 14.
- REITER, E. A structured review of the validity of bleu. *Computational Linguistics*, v. 44, n. 3, p. 393–401, set. 2018. ISSN 0891-2017. Citado na página 18.

STEGEREN, J. van; MYŚLIWIEC, J. 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*. New York, NY, USA: Association for Computing Machinery, 2021. (FDG '21). ISBN 9781450384223. Disponível em: <<https://doi.org/10.1145/3472538.3472595>>. Citado na página 18.

VÄRTINEN, T.; HÄMÄLÄINEN, P.; GUCKELSBERGER, C. Generating long-form narratives with large language models. *IEEE Transactions on Games*, 2024. Citado na página 19.

VASWANI, A. et al. Attention is all you need. arXiv, n. arXiv:1706.03762, ago. 2023. ArXiv:1706.03762 [cs]. Disponível em: <<http://arxiv.org/abs/1706.03762>>. Citado na página 13.

VÄRTINEN, S.; HÄMÄLÄINEN, P.; GUCKELSBERGER, C. Generating role-playing game quests with gpt language models. *IEEE Transactions on Games*, v. 16, n. 1, p. 127–139, 2024. Citado 2 vezes nas páginas 11 e 18.

XIE, Y. et al. A survey on rag meeting llms: Towards retrieval-augmented large language models. *arXiv preprint arXiv:2401.XXXX*, 2024. Citado 2 vezes nas páginas 6 e 15.