Avaliação da Geração Automatizada de Narrações Esportivas com o Modelo de Linguagem LLaMA:

Wendson Carlos Souza da Silva

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, PB

2023

Wendson Carlos Souza da Silva

Avaliação da Geração Automatizada de Narrações Esportivas com o Modelo de Linguagem LLaMA

Monografia apresentada ao curso Engenharia da Computação
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Bacharel em Engenharia da Computação

Orientador: Prof. Dr. Yuri de Almeida Malheiros Barbosa

Dezembro de 2023

Catalogação na publicação
Seção de Catalogação e Classificação

S586a Silva, Wendson Carlos Souza da.
Avaliação da geração automatizada de narrações esportivas com o modelo de linguagem LLaMA / Wendson Carlos Souza da Silva. - João Pessoa, 2023.
51 f. : il.

Orientação: Yuri de Almeida Malheiros Barbosa Barbosa.
TCC (Graduação) - UFPB/CI.

1. Inteligência artificial generativa. 2. LLaMA. 3. Narração esportiva. 4. LLMs. I. Barbosa, Yuri de Almeida Malheiros. II. Título.

UFPB/CI CDU 004.8

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Engenharia da Computação intitulado ***Avaliação da Geração Automatizada de Narrações Esportivas com o Modelo de Linguagem LLaMA*** de autoria de Wendson Carlos Souza da Silva, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Thaís Gaudencio do Rêgo
Universidade Federal da Paraíba

Prof. Dr. Yuri de Almeida Malheiros Barbosa
Universidade Federal da Paraíba

Prof. Dr. Tiago Maritan Ugulino de Araújo
Universidade Federal da Paraíba

João Pessoa, 11 de dezembro de 2023

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

"Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer."

Alan Turing

DEDICATÓRIA

Dedico a você, vovó Maria :)

AGRADECIMENTOS

Dedico este trabalho a todos que fizeram parte da jornada desse meu grande sonho: à minha família que sempre me apoiou e acreditou no meu potencial, especialmente a meus irmãos, Rayane, Wesley e Rebeca, que me motivam e inspiram todos os dias.

Aos meus amigos, Laura e Jóison, que compartilharam momentos incríveis comigo, me deram todo o apoio para continuar e me fortaleceram em situações difíceis, sou muito grato a vocês. Aos meus professores, que me guiaram, inspiraram e compartilharam seu conhecimento, moldando o meu caminho acadêmico.

À minha vovó Terezinha por todo amor e suporte. Aos meus maiores motivadores, minhas inspirações, meus pais, Luciene e Carlos, cujo amor e apoio inabaláveis me impulsionaram a alcançar este objetivo.

Este trabalho é dedicado a todos vocês, em reconhecimento da importância que tiveram em minha jornada. Obrigado por fazerem parte desta conquista.

RESUMO

Considerando a vasta quantidade de elementos midiáticos, dados desportivos e estatísticos de futebol, bem como das questões que envolvem a escassez de cobertura esportiva em eventos de menor abrangência, surge a necessidade de empregar estudos experimentais que se concentrem na automação da narração de partidas de futebol. Além disso, a imperativa demanda por aprimorar a acessibilidade cultural para indivíduos com deficiências, por meio da utilização de tecnologias emergentes, reforça a relevância deste domínio de pesquisa. Frente a isso, os Modelos de Linguagem de Larga Escala (LLMs) mostram-se relevantes por apresentarem resultados notáveis na geração de conteúdo. Dessa forma, pesquisa-se sobre o uso do Large Language Model Meta AI - LLaMA a fim compreender o desempenho desse sistema em conceber narrações desportivas a partir de eventos de jogos de futebol. Para tanto, é necessário conduzir experimentos, comparar e avaliar o sistema por parâmetros distintos, analisar a qualidade da narrativa em contexto de legibilidade textual e identificar padrões de comportamentos, falhas e virtudes do modelo. Procedeu-se, então, uma pesquisa com metodologia exploratória, a qual se valeu das informações granulares, obtidas lance-a-lance, para estruturar as ocorrências em agrupamentos de eventos, sendo estes utilizados como comando de entrada para o LLaMA. As amostram também foram submetidas a estruturação em diferentes conjuntos buscando aferir a relevância da alteração dos parâmetros no modelo. Diante disso, verificou-se que o LLaMA apresentou resultados favoráveis relacionados a compreensão de leitura, principalmente no conjunto com parâmetros de temperatura de amostragem, top-k e top-p mais altos, contudo exibiram uma baixa taxa de acerto. O modelo também obteve resultados esperados no que diz respeito ao objetivo de possuir características relacionadas ao gênero narrativo que condiz com a personalidade de um comentarista esportivo. O incremento de informações, característico dos LLMs se mostrou como um fator preocupante e determinístico para avaliação dos erros no texto de saída. Essas informações evidenciam a capacidade do sistema em gerar narrações automatizadas de eventos esportivos, embora revelem a presença de algumas limitações que demandam investigações adicionais e aprimoramentos por parte dos desenvolvedores.

Palavras-chave: Inteligência Artificial Generativa; LLaMA; Narração Esportiva; LLMs.

ABSTRACT

Considering the vast amount of media elements, sports data, and football statistics, as well as the issues surrounding the lack of sports coverage in less prominent events, there arises the need to employ experimental studies that focus on automating football match commentary. Additionally, the imperative demand to improve cultural accessibility for individuals with disabilities through the use of emerging technologies reinforces the relevance of this research domain. In this context, Large Language Models (LLMs) prove relevant for yielding remarkable results in content generation. Thus, research is conducted on the use of the Large Language Model Meta AI - LLaMA to understand the performance of this system in crafting soccer commentary from football game events. To do so, it is necessary to conduct experiments, compare and evaluate the system using different parameters, analyze the quality of the commentary in the context of textual readability, and identify patterns of behaviors, flaws, and virtues of the model. An exploratory research methodology was then employed, utilizing granular information obtained play-by-play to structure occurrences into event clusters, which were used as input commands for LLaMA. The samples were also subjected to structuring in different sets, aiming to assess the relevance of parameter changes in the model. In this regard, it was observed that LLaMA showed favorable results related to reading comprehension, especially in the set with higher sampling temperature, top-k, and top-p parameters. However, they exhibited a low accuracy rate. The model also achieved expected results concerning the goal of possessing narrative genre characteristics that align with the personality of a sports commentator. The increase in information, characteristic of LLMs, proved to be a concerning and deterministic factor for evaluating errors in the output text. These findings highlight the system's ability to generate automated sports event commentary, although they reveal the presence of some limitations that require further investigation and improvements from developers.

Key-words: Generative Artificial Intelligence; LLaMA; Sports Commentary; LLMs.

LISTA DE FIGURAS

1 Processo de treinamento do LLama-2 versão chat. 22

2 Estratificação das ações da partida 32

3 Parte 1: Categorização dos grupos com baseado nos eventos esportivos . . 33

4 Parte 2: Categorização dos grupos com baseado nos eventos esportivos . . 33

5 Divisão dos conjuntos com base na natureza da narração 34

6 Estrutura para o comando de entrada no modelo 36

7 Gráfico de boxplot contendo medidas de posição e dispersão do IFFL . . . 37

8 Gráfico de boxplot contendo a frequência atributos nas amostras 39

9 Gráfico de barras comparando o mesmo evento a partir de duas perspectivas: com e sem o atributo do oponente 39

10 Histograma com a distribuição de frequência entre os conjuntos 40

LISTA DE TABELAS

1 Comparação entre LLMs em tarefas de Common Sense Reasoning. 22

2 Comparação entre quantização dos modelos usando a técnica de diferença de parâmetros de compressão 23

3 Comparação dos trabalhos relacionados evidenciando as metodologias implementadas, métricas de avaliação e contexto 28

4 Proporção da base de treino por idiomas 29

5 Avaliação da taxa de acerto para os conjuntos 41

6 Principais frases característica da narração esportiva extraído da saída do LLaMA 42

7 Categorização dos erros em todas as amostras avaliadas como incorretas no experimento 44

LISTA DE ABREVIATURAS

LLaMA – Large Language Model Meta AI

IFFL – Índice Flesch de Facilidade de Leitura

IA – Inteligencia Artificial

LLM - Modelos de Linguagem de Grande Escala

PLN - Processamento de Linguagem Natural

GPT - Transformador pré-treinado Generativo

RNA - Redes Neurais Artificiais

CPU - Unidade central de processamento

API - Interface de Programação de Aplicações

IAG - Inteligencia Artificial Generativa

Sumário

1 INTRODUÇÃO **18**

1.0.1 Objetivo geral 19

1.0.2 Objetivos específicos 19

1.1 Estrutura da monografia 19

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA **20**

2.1 Processamento de Linguagem Natural (PLN) 20

2.2 Modelos de Linguagem de Larga Escala (LLMs) 21

2.3 Quantização em modelos LLMs 22

2.4 Métodos de avaliação 23

2.4.1 Índice de Flesch de Facilidade de Leitura 24

2.4.2 Menção aos atributos de entrada 24

2.4.3 Avaliação binária dos erros 24

3 TRABALHOS RELACIONADOS **25**

3.1 Gerando narrações ao vivo de partidas de futebol a partir de dados do jogo 25

3.2 Narração ao vivo de um videogame de futebol gerado por uma IA 26

3.3 Treinando um gerador de comentários de jogos com expressões de movimento no campo 26

3.4 Análise comparativa entre os trabalhos 27

4 PROCESSOS METODOLÓGICOS **29**

4.1 Configuração do Modelo de Linguagem de Larga Escala (LLM) 29

4.1.1 Interface do modelo em C/C++ 30

4.2 Conjunto de dados 30

4.3 Preparação do experimento e configuração dos parâmetros 31

4.3.1 Seleção dos eventos esportivos 31

4.3.2 Configuração dos parâmetros 32

4.3.3 Processamento das informações 32

4.3.4 Configuração do *prompt* de entrada 35

4.4 Métricas de avaliação dos resultados 35

5 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS 37

5.1 Análise comparativa entre os grupos 37

5.2 Análise comparativa entre os conjuntos 39

5.3 Caracterização do narrador esportivo 41

5.4 Estilo de narrativa do modelo 42

5.5 Principais falhas do modelo 43

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS 45

6.1 Trabalhos futuros 45

REFERÊNCIAS 47

ANEXO A - ANEXOS E APÊNDICES 1 51

1 INTRODUÇÃO

É notável a crescente evolução dos modelos generativos de linguagem e a popularização em diversos âmbitos sociais, devido a sua aptidão em elaborar materiais similares aos produzidos por humanos (NASCIMENTO, 2023). Os *chatbots*, como a exemplo do ChatGPT e o LLaMA vem ganhando destaque na estruturação e escrita de textos de cunho comuns e científicos, em virtude da utilização de técnicas avanças de Processamento de Linguagem Natural (PLN) e Aprendizagem Profunda de Máquinas, produzindo conteúdos tão excepcionais que torna-se difícil a detecção do autor (SALVAGNO, 2023). No âmbito jornalístico, tal pratica é emprega para transformar dados gerais em textos narrativas ou notícias com ou sem nenhuma intervenção humana (PEÑA-FERNÁNDEZ et al., 2023) sendo prudente no uso dos termos empregados (BRAZ, 2023).

Diversos trabalhos recentes vêm empregando essa ferramenta em contextos jornalísticos (MONTEIRO, 2023) e reforçam o potencial para cenários similares como da narração esportiva, gênero esse que se assemelha as narrativas tradicionais (SANTOS, 2010). Não obstante da atualidade, a automatização de textos no jornalismo desportivo, por meio da transformação de dados estruturados em texto, tornou-se pioneiro dentro do processo de redação de noticiais principalmente com o uso de técnicas relacionada a sistema da informação ou inteligência artificial (CANAVILHAS, 2022). A relevância desse tipo de aplicação surgiu em decorrência da disponibilidade de uma vasta quantidade de dados esportivos para abastecer os sistemas (LEWIS et al., 2019). Tais dados podem ser empregados para conceber resumos completos de partidas de futebol (GAUTAM et al., 2022) a partir de gerais do jogo, legendas e metadados. Estas práticas auxiliam na redução de problemas sociais relacionadas de cobertura em esportivos locais e, principalmente, em gerar acessibilidade para pessoas com deficiência, sendo este um complexo desafio quando falamos em acessibilidade cultural (LEITE, 2016).

A massiva quantidade de conteúdos midiáticos e de dados desportivos enfatizam a necessidade de operar em frentes que vão além da sumarização e criação de notícias no esporte, se expandindo para geração de uma narrativa lance-a-lance. Tal recurso também auxilia na escassez de narradores em eventos esportivos de menor abrangência e no apoio a pessoas com deficiência como já citado. Observa-se alguns trabalhos desenvolvidos nessa linha de pesquisa, que utiliza as informações coletados durante a partida para inseri-las como entrada para geração da narração automatizada (TANIGUCHI et al., 2019; CZAPLICK, 2023). No entanto, é importante destacar que ainda há carência na literatura acadêmica e as estruturas a serem empregada necessitam de mais aprofundamento. Portanto, torna-se importante conduzir experimentos de pesquisa adicionais, a fim de aprofundar a compreensão das alternativas viáveis. Adicionalmente, o contínuo surgimento de novos modelos de linguagem enfatiza a importância de permanecer engajado nesse campo

temático. Nesse contexto, o foco deste estudo é avaliar a possibilidade de geração automatizada de narrações esportivas a partir do modelo de linguagem LLaMA desenvolvido pela Meta AI. Este sistema, que teve um lançamento recente, suscita a necessidade de considerar estudos experimentais dessa nova tecnologia, estimulando a realização de pesquisas nesse contexto. Notavelmente, esse sistema se destaca por contribuir para a democratização dos estudos relacionados aos Modelos de Linguagem de Larga Escala (LLMs), uma vez que é compatível com a filosofia de código aberto e foi treinado com base em dados de acesso público, conforme mencionado por TOUVRON (2023).

1.0.1 Objetivo geral

O objetivo principal desse estudo é avaliar e compreender o desempenho do modelo de linguagem de larga escala - LLaMA 2 em conceber narrações esportivas a partir de eventos de partidas de futebol.

1.0.2 Objetivos específicos

1. Comparar e avaliar o desempenho do modelo com base em diferentes parâmetros de inferência de saída como top-k, top-p e temperatura de amostragem;
2. Verificar o comportamento para diferentes tipos de ações no futebol;
3. Avaliar a qualidade da narração esportiva gerada em termos de legibilidade textual;
4. Identificar padrões de comportamento, deficiências e virtudes do sistema.

1.1 Estrutura da monografia

Este trabalho será dividido em 6 capítulos. No Capítulo 2 serão contemplados os principais conceitos relacionados a inteligência artificial, com ênfase em modelos generativos e processamento de linguagem natural. O Capítulo 3 expõe trabalhos semelhantes vistos na literatura, apontando os métodos empregados, tal como os resultados obtidos por cada um deles. O Capítulo 4 contém todos os processos metodológicos aplicados ao estudo, o que inclui a configuração do modelo, pré-processamento dos dados, processo de quantização entre outros. No Capítulo 5 serão apresentados os resultados dessa pesquisa a partir de uma análise estatística descritiva, que compara os grupos de eventos e os conjuntos entre si, como também as discussões acerca desses resultados e a correlação com os marcadores característicos na narração esportiva. Por fim, o Capítulo 6 irá conter os desfechos do experimento e das avaliações, tal como as limitações encontradas na pesquisa e os trabalhos futuros.

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA

A crescente adoção da Inteligência Artificial (IA) como uma ferramenta tecnológica de apoio em diversos contextos tem se destacado, especialmente durante os últimos anos. As tecnologias emergentes trazem consigo alterações em todas as esferas socais. As mudanças destacam a transformação da percepção humana em relação ao modo como as ferramentas se tornaram ativas e intuitivas, desempenhando um papel de destaque nas atividades cotidianas.

O presente capítulo será dividido em seções que abordam os principais conceitos relacionados à temática que engloba o processamento de linguagem natural, arquitetura e como se apresentam os LLMs, procedimentos para a diminuição dos custos computacionais para execução desse sistema e as principais métricas de avaliação do texto de saída.

2.1 Processamento de Linguagem Natural (PLN)

Os sistemas de processamento de linguagem se caracterizam pelo conhecimento da linguagem com um saber amplo que vai desde o reconhecimento do que é uma palavra até abordagens mais profundas. Os autores Jurafsky e Martin (2008) destacam que o conceito relacionado à capacidade de permitir que computadores processem a linguagem natural remete aos primórdios da concepção dos próprios computadores. Tem-se aplicações de conversação, sistemas de tradução e de pergunta e resposta como exemplos desse tipo de tecnologia. De forma mais ampla, o ramo do Processamento de Linguagem Natural (PLN) é definido como:

> "...uma área de pesquisa e aplicação que explora como os computadores podem ser usados para compreender e manipular texto ou fala em linguagem natural para fazer coisas úteis. Os pesquisadores da PLN pretendem reunir conhecimento sobre como os seres humanos entendem e usam a linguagem para que ferramentas e técnicas apropriadas possam ser desenvolvidas para fazer com que os sistemas de computador entendam e manipulem as línguas naturais para executar as tarefas desejadas." (CHOWDHURY, 2003, p. 51)

Um dos experimentos mais notáveis no contexto da PLN foi a criação de ELIZA, desenvolvida pelo cientista Joseph Weizenbaum. O software foi concebida como um artefato ilustrativo de uma teoria sobre a inteligência artificial e as interações entre humano e o computador. A teoria tomou como base uma abordagem comportamental, na qual a inteligência da máquina era apresentada como o resultado de um efeito enganoso. Este feito descadeou diversas discussões gerando visões alternativas sobre a IA e como a narrativa de máquinas pensantes e a narrativa do engano se apresentavam como importante no meio cientifico (NATALE, 2018).

2.2 Modelos de Linguagem de Larga Escala (LLMs)

As tecnologias relacionados a Inteligencia Artificial Generativa (IAG), ramo da IA que abarca a elaboração de conteúdos novos a partir de dados existentes empregando algoritmos e redes neurais (FRANGANILLO, 2023), incorporam uma série de transformações na forma como acessamos o conhecimento e na forma como interagimos com o mundo, trazendo impactos profundos para os contextos que estão inseridos (LÓPEZ, 2023). Essa evolução desencadeou o desenvolvimento de diversos sistemas e *chatbots* para criação automática de textos como o LLaMA, ChatGPT, Writesonic, Copy.ai, os chamados Modelos de Linguagem de Grande Escala (LLMs). O conteúdo extraído a partir do processamento de linguagem natural pode ser aplicado a esses sistemas, gerando uma conexão entre tecnologias.

Os sistemas são capazes efetuar tarefas a partir de instruções textuais ou até mesmo através de exemplos (Brown et al., 2020). Os sistemas que ganham destaque são os que levam em consideração o dimensionamento do conjunto de dados e o tamanho do modelo, pensando no desempenho para inferência e velocidade de resposta (HOFFMAN et al., 2022). A exemplo podemos citar o LLaMA-13B que supera o GPT-3 em vários *benchamarks*, mesmo possuindo um tamanho 10x menor do que seu concorrente, também sendo competitivo ao PaLM-540B e Chinchilla, na versão 65B (TOUVRAN et al, 2023).

Esses modelos assumem funções importantes no processo de construção de conteúdos, o que inclui a narração automatizada para gerar comentários a partir do lance-a-lance, objetivo desse trabalho. Outra aplicação é nos sistemas de recomendação, oferecendo sugestões personalizadas para os usuários a fim de facilitar o acesso a informação e reduzir o tempo de busca. A ideia principal está em relacionar as interações entre os usuários e os dados secundários associados a eles como avaliações e títulos pesquisados em uma plataforma de *streaming*, para predizer uma sugestão (FAN et.al, 2023)e os dados textuais são os principais itens a serem utilizados durante essa tarefa

Dando atenção ao Large Language Model Meta AI (LLaMA), temos que este apresenta uma arquitetura de transformadores semelhantes a maioria dos modelos de linguagem, acrescido de técnicas para melhorar a estabilização no treinamento, função de ativação SwiGLU para melhoria do desempenho entre outros (TOUVRAN et al., 2023). O LLaMA-2-Chat, por exemplo, aplica aprendizado por reforço com *feedback* humano, dialogando com o ambiente e sendo realimentado em forma de recompensa, ou penalidade, em que faz-se o uso, principalmente, de algoritmos de Otimização de Política Proximal (do inglês, *Proximal Policy Optimization* - PPO) e amostragem de rejeição para torna-se um modelo de qualidade e seguro. O fluxograma desse processo pode ser visto na Figura 1. Obtém-se destaque também por possuir um treinamento que decorre de informações públicas disponíveis na rede, executando 2 trilhões de *tokens* de dados para garantir um

excelente desempenho e evitar alucinações do modelo (TOUVRON et al., 2023).

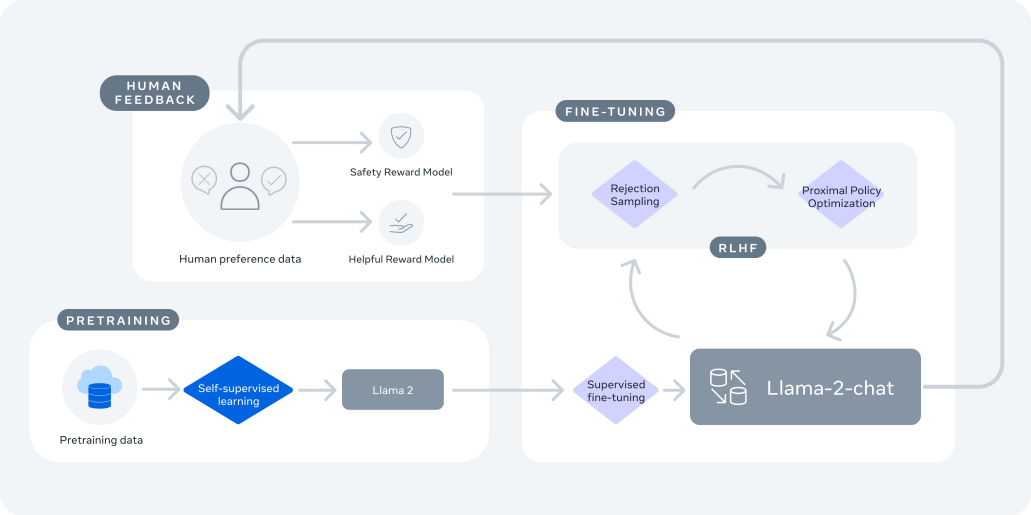

Figura 1: Processo de treinamento do LLama-2 versão chat. O método segue três grandes etapas que se inicia com o pré-treino com dados públicos, seguido do *fine-tuning* do modelo original e aplicação de técnicas de RLHF e PPO. Fonte: Touvron et al. (2023).

Na Tabela 1 é possível visualizar ainda uma comparação entre os modelos de linguagem mais conhecidos versus o sistema da Meta AI em que as colunas representam os conjuntos de desafios que os modelos foram submetidos. Evidencia-se que o LLaMA supera o sistema da OpenAI em quase todos os desafios, mostrando-se competitivo no ramo.

		BoolQ	PIQA	SIQA	HellaSwang	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60,5	81,0	-	78,9	70,2	68,8	51,4	57,6
Gopher	280B	79,3	81,8	50,6	79,2	70,1	-	-	-
Chinchilla	70B	83,7	81,8	51,3	80,8	74,9	-	-	-
PaLM	62B	84,8	80,5	-	79,7	77,0	75,2	52,5	50,4
PaLM-cont	62B	83,9	81,4	-	80,6	77,0	-	-	-
PaLM	540B	**88,0**	82,3	-	83,4	**81,1**	76,6	53,0	53,4
LLaMA	7B	76,5	79,8	48,9	76,1	70,1	72,8	47,6	57,2
	13B	78,1	80,1	50,4	79,2	73,0	74,8	52,7	56,4
	33B	83,1	82,3	50,4	82,8	76,0	**80,0**	**57,8**	58,6
	65B	85,3	**82,8**	**52,3**	**84,2**	77,0	78,9	56,0	**60,2**

Tabela 1: Comparação entre LLMs em tarefas de Common Sense Reasoning. Tem-se o destaque para o PaLM, LLaMA na versão 33B e 65B. **Fonte:** Autor baseado em Touvron et al. (2023).

2.3 Quantização em modelos LLMs

A técnica de quantização, também conhecida como quantização de bits, denomina-se como um instrumento para reduzir o número de bits responsáveis pelo armazenamento de uma variável. Nesse processo, uma variável com intervalo contínuo é discretizado em graus de quantização. Esse procedimento pode ser aplicado aos modelos de linguagem citados na seção anterior buscando compactá-los, e dar-se a partir da quantização de probabilidade de N-gramas e os pesos de *back-off* e a supressão de alguns parâmetros.

Os métodos como a compressão absoluta dos parâmetros e pela diferença deles são comumente aplicados (WHITTAKER e RAJ, 2001). Tal necessidade surgiu visto que os LLMs necessitam de uma massiva capacidade de memória e alto custo computacional. No entanto, o uso da quantização pode desencadear a perda de dados importantes do modelo, ou até mesmo uma baixa inferência, quando aplicado em altos níveis. Assim, novos métodos estão emergindo, como o *Dual Grained Quantization* (DGQ), uma nova técnica de quantização A8W4 para LLMs. Essa abordagem preserva um desempenho superior ao mesmo tempo em que assegura uma velocidade de inferência rápida (ZHANG et al., 2023). A Tabela 2 exibe a diminuição no tamanho dos modelos com base em dois níveis de quantização: 2 e 4 bits.

LM	Bits	3-gram dels.	Tamanho (Mb)	WER (%)
1	4	1119492	40,7	22,1
1	2	3526503	32,8	22,5
2	4	201013	57,5	21,1
2	2	932822	52,0	21,7
3	4	3379	0,126	34,1
3	2	9651	0,106	34,2

Tabela 2: Comparação entre quantização dos modelos usando a técnica de diferença de parâmetros de compressão. A tabela compara diferentes graus de quantização. **Fonte:** Autor baseado em Whittaker e Raj (2001).

2.4 Métodos de avaliação

A avaliação do conteúdo produzido por modelos de linguagem se apresenta como um elemento crucial para quantificar o desempenho e compará-lo com as metas estabelecidas no início do projeto. Segundo Godoy (1995), a necessita de dedicação a quantificação, precisão e margem de segurança nas deduções é importante para mitigar distorções na análise e interpretação dos resultados. Dentre as opções disponíveis, inclui-se a análise da legibilidade do texto, a verificação da incorporação dos atributos mencionados no *prompt*, a avaliação dos erros mais significativos e a investigação relacionada ao tópico do conteúdo gerado, procedimentos esses que foram aplicados no âmbito deste estudo. Tais técnicas são aplicadas aos LLMs após passar por todas as etapas citadas nas seções anteriores.

Há outras abordagens que podem igualmente servir como métricas de avaliação, tais como o coeficiente de coesão textual, o tempo de resposta e o consumo de recursos computacionais, entre outras. Adicionalmente, é viável conduzir pesquisas de campo de natureza qualitativa, com o intuito de entrevistar indivíduos e obter novas perspectivas. Para facilitar o entendimento do leitor, as seções seguintes descreverão o funcionamento das métricas avaliativas e a base matemática e lógica por trás delas.

2.4.1 Índice de Flesch de Facilidade de Leitura

O modelo matemático que rege o Índice de Flesch de Facilidade de Leitura (FLESCH, 1979) desempenha um papel importante na compreensão da legibilidade de um determinado texto, avaliando a dificuldade na leitura. Apresentada na Equação 1, a métrica propõe identificar a relação entre o comprimento médio das palavras e frases, e a compreensibilidade do texto em língua inglesa. Nas pesquisas que abrangem o campo da IA e, principalmente, na avaliação de modelos de linguagem de larga escala é notório o uso dessa medida durante os processos metodológicos para aferir as respostas do sistema (DAVIS et al., 2023; TALONI et al., 2023; SETH et al., 2023; HUYNH et al., 2023).

$$IFFL = 206,835 - (1,015 \cdot M_{PF}) - (84,6 \cdot M_{SP}) \tag{1}$$

onde:

M_{PF} - Média do comprimento da frase
M_{SP} - Média de sílabas por palavra

2.4.2 Menção aos atributos de entrada

No processo de abordar os atributos mencionados, o tratamento para menção dos atributos toma como base a citação dos deles dentro do texto gerado. A análise é pontuada através de uma escala que varia entre 0% - 100%, em que zero representação a ausência total de referências e 100 o aparecimento de todos dos qualificadores. Nesse caso, se existirem 10 atributos de entrada e o modelo citar oito deles, temos 80% das menções dos atributos.

2.4.3 Avaliação binária dos erros

Durante a avaliação das inconsistências, os eventos são examinados manualmente e separadas em dois grupos: respostas corretas e incorretas. Caso ocorra alguma distorção ou informação falsa no conteúdo gerado, a amostra inteira é tratada como erro. Essa investigação baseia-se exclusivamente na comparação com os dados introduzidos no *prompt*, podendo haver divergência com outras fontes, principalmente as feitas por humanos.

3 TRABALHOS RELACIONADOS

Nesse capítulo serão apresentados os principais trabalhos relacionados, suas implicações para o tema e uma análise comparativa. Os trabalhos selecionadas contemplam métodos e resultados semelhantes ao objeto principal desse projeto.

3.1 Gerando narrações ao vivo de partidas de futebol a partir de dados do jogo

Taniguchi *et al.* (2019) propuseram um sistema para geração de narração de jogos esportivos de futebol, em tempo real, tomando como base a sequência de eventos da partida, resultando em textos que descrevem os lances do jogo. A implementação desse trabalho processa-se a partir de um conjunto de ocorrências de partidas de futebol da *English Premier League* referente a temporada 2015/2016, que alternam entre informações lance-a-lance e comentários realizados por narradores reais, fornecidos pela empresa de dados esportivos *OptaSports*. Por existir uma discrepância entre o volume desses dois tipos de registros, uma vez que os eventos individuais superam os comentários codificados por humanos, as narrações resultantes são compostas pela união de mais de um evento evidenciando um alinhamento parcial em relação as dados de entrada.

Toda a narração resulta de um modelo de linguagem que aplica uma RNA multicamadas no processo de codificação das entradas categóricas e contínuas, que sofrem tratamentos e são posteriormente concatenadas para compor a estrutura da rede neural. Em seguida, as representações são inseridas no decodificador, resultando em comentários para a partida. Além disso, implantou-se algoritmos para melhorar o erro de reconstrução e concatenação de palavras, técnicas para contornar problemas com a menção do nome de jogadores e do time durante a decodificação, e métodos para diferenciar os eventos mais relevantes.

Os resultados foram analisados a partir de avaliações humanas verificando aspectos gramaticais e de informatividade dos textos, e pontuações do algoritmo BLEU (PAPINENI *et al.*, 2002), que comparava o texto gerado com o da base inicial. Os resultados experimentais demonstraram que o sistema possui um desempenho interessante sendo avaliada com pontuações intermediários na avaliação humana e no algoritmo BLEU, referente ao modelo contendo a junção de todas as estratégias de melhoria descritas no projeto. Ademais, narrações corretas, porém com expressões distintas das referências, repetição no nome dos jogares e menções equivocadas dos eventos puderam ser vistas no experimento.

3.2 Narração ao vivo de um videogame de futebol gerado por uma IA

O objetivo desse trabalho desenvolvido por Czaplicki (2023) é elaborar narrações para partidas de futebol de videogame valendo-se do GPT-3.5 para contornar restrições das narrativas pré-gravadas que enfrentam limitadores de tempo, idiomas, diversidade e profundidade do instrumento narrado. O artigo dispõe de uma aplicação que obtém dados do estado atual da partida, gera uma estrutura fixa com essas informações e utiliza-os como entrada para o sistema.

Para coletar os registros do jogo, utilizou-se o ambiente de desenvolvimento para jogos de futebol populares de videogame do *Google AI*, para simular uma competição. Posteriormente, os dados foram formatados em um estrutura textual fixa e enviada à API do modelo de inteligência generativa e, por fim, convertidos em áudio com base em programas de produção de fala humana. O emprego da ideia de cliente-servidor na IA para tornar contínua e dinâmica a passagem de informação ao *chat* se mostrou uma estratégia positiva nos resultados.

A avaliação desse estudo consistiu em confrontar a saída textual de 20 jogos com os dados coletados inicialmente, contabilizando as saídas corretas e incorretas com uma precisão de acerto de 78%. Tornou-se evidente que erros oriundos da plataforma da *Google AI* comprometeram 67% das imprecisões gerais, uma vez que gerou eventos diferentes para o mesmo estado do jogo. O excesso de precisão também se mostrou como um problema, já que ocorreram situações em que a IA informa que o gol foi realizado pelo próprio goleiro e não por um atacante do outro time, uma vez que o último a tocar na bola foi quem a defendeu. De forma semelhante, o GPT-3.5 foi responsável pela incoerência da saída em virtude da adição de eventos inexistentes e cometendo graves erros de lógica.

3.3 Treinando um gerador de comentários de jogos com expressões de movimento no campo

A busca por propor soluções em aplicações que demandam a transformação de registros sequenciais em linguagem natural por meio da IA também tornou-se objeto de estudo relevante na pesquisa de Kameko *et al.* (2015). Os autores expõem um sistema para produção de comentários narrativos para o jogo de tabuleiro *Shogi* (conhecido como xadrez japonês), como recurso para driblar programas que fornecem apenas descrições sequenciais de maneira técnica e sem interpretações, requerendo, assim, que o espectador possua um elevado conhecimento para desfrutar de uma experiência satisfatória. Dessa forma, ocorreu a implementação de um sistema que realiza o mapeamento e previsão das palavras que caracterizam as posições das jogadas e concebe automaticamente um comentário usando modelos de entropia máxima.

A definição das expressões de movimento no tabuleiro decorreu de métodos baseados em regras, fazendo uso de árvores, para associar a jogada com a respectiva descrição. Também desenvolveu-se um modelo de predição de palavras com o uso de uma RNA multicamadas com saídas binárias gerando um vetor que indica se o termo em específico deve ou não aparecer no comentário. Na produção da narrativa, a probabilidade estimada dos elementos textuais aparecerem nos comentários finais é baseada em um modelo de linguagem log-linear, com o uso do algoritmo de regressão *softmax*. A composição final é constituída por palavras específicas já previstas anteriormente e palavras geradas para características linguísticas. O conjunto de dados considerou eventos de jogos comentados por especialistas humanos filtrando apenas por comentários que possuam alguma relação com os estados do jogo.

Os comentários gerados exibem resultados satisfatórios, porém com muitas limitações por conter poucas informações e não esclarecer os lances. Apareceram ainda erros textuais ocasionados por uma base de treino com descuidos dos especialistas humanos. As falhas de lógica relativas a eventos que não podem acontecer no mesmo jogo e a falta de distinção entre registros de potenciais posições, dos que realmente aconteceram, foram observados no experimento.

3.4 Análise comparativa entre os trabalhos

Ao confrontarmos os estudos citados anteriormente ao objeto de investigação dessa monografia, torna-se perceptível que não existem limitações de informações sobre a partida de futebol para compor o conjunto de dados de entrada, problema que acomete Kameko *et al.* (2015), uma vez que o estudo atual utiliza uma base que dispõe de referências detalhadas sobre os acontecimentos na partida, da competição/temporada e escalações. Diferente dos outros autores, não houve o uso de limitadores de palavras aplicada para coincidir com os tempos da narração. De maneira análogo ao ensaio de Czaplicki (2023), este trabalho fez uso de modelos de linguagem LLM e a ideia de agente-servidor para maximizar os resultados. Todavia, buscou aperfeiçoar a avaliação dos resultados adicionando novas métricas como legibilidade e menção dos atributos de entrada, dessa forma a análise torna-se mais completa indo além da avaliação binária (acerto e erro). A Tabela 3 apresenta um quadro comparativo entre os trabalhos citados anteriormente. A conferência dos estudos associa a metodologia aplicada na base de dados, com as métricas de avaliação e o contexto do trabalho.

Autor/ano	Metodologia aplicada	Métricas de avaliação	Base de Dados	Contexto
TANIGUCHI *et al.*, 2019	RNA: Codificador - Decodificador	BLEU entre comentários ao vivo gerados e o texto padrão-ouro; avaliação humana; gramaticalidade e informatividade.	Dados estruturados de partida de futebol (OptaSports)	Futebol de campo
CZAPLICKI, 2023	LLM: GPT-3.5	Teste de precisão do sistema, análise de erros e resultado das entrevistas	Estados do jogo da simulação	Futebol de video-game
KAMEKO *et al.*, 2015	RNA e métodos probabilísticos	Análise dos erros e precisão do analisador	Registro dos jogos e comentários dos especialistas	Jogo de tabuleiro Shogi
AUTOR, 2023	LLM: LLama C++	Indice de legibilidade, análise binária do erro, menção dos atributos e análise do conteúdo	Dados estruturados de partida de futebol (StatsBomb)	Futebol de campo

Tabela 3: Comparação dos trabalhos relacionados evidenciando as metodologias implementadas, métricas de avaliação e contexto. Os trabalhos apresentam características semelhantes como a metodologia aplicada por modelos de RNA ou LLM e limtações no conjunto de dados. **Fonte:** Autor (2023).

4 PROCESSOS METODOLÓGICOS

Esta trabalho tem por base uma metodologia exploratória com abordagem quantitativa que visa investigar e avaliar a aplicação do modelo de linguagem de grande escala LLaMA 2 da Meta AI para reproduzir narrações esportivas, a partir dos eventos de partidas de futebol. A definição de uma abordagem exploratório para este estudo apoia-se em conceber uma nova perspectiva para a temática fornecendo mais informações para o assunto investigado e gerando uma delimitação do tema da pesquisa, que se encontra em fase preliminar (PRODANOV e FREITAS, 2013).

A Seção 4.1 demostra a configuração do modelo, processo de quantização e conexão com as interfaces em C/C++ e *Python*. A Seção 4.2 exibe a evolução da coleta e processamento dos dados, a Seção 4.3 expõe as etapas de preparação do experimento e ajuste dos parâmetros e, por fim, a Seção 4.4 apresenta as técnicas de avaliação que serão posteriormente empregadas nos resultados.

4.1 Configuração do Modelo de Linguagem de Larga Escala (LLM)

Nesse estudo, optou-se pelo modelo **LLaMA 2 13-B chat**, uma variação da versão original, aperfeiçoada e treinada com uma gama maior de dados públicos. O modelo aplicado no estudo foi obtida a partir de uma requisição feita à Meta AI através do site oficial do produto. Optou-se por avaliar o modelo apenas em língua inglesa, isto é, entradas e saídas nesse idioma, uma vez que o sistema de linguagem escolhido foi treinado predominantemente em inglês. A Tabela 4 exibe a disparidade da composição dos dados de treinamento em relação aos idiomas, evidenciando a taxa baixa de textos que não são da língua inglesa, como por exemplo, em português.

Idioma	**Base de dados (em %)**
Inglês	89,70%
Desconhecido	8,38%
Alemão	0,17%
Francês	0,16%
Espanhol	0,13%
Russo	0,13%
Italiano	0,11%
Português	0,09%

Tabela 4: Proporção entre base de treino por idiomas. O material desenvolvido pela Meta AI evidência que a modelo foi treinado predominantemente com base de dados em língua inglesa. **Fonte:** Autor baseado em Touvron et al. (2023).

4.1.1 Interface do modelo em C/C++

Desenvolvido pelo programador e físico Georgi Gerganov em 2023 objetivando executar modelos de LLaMA usando quantização em n-bits para fins, principalmente, educacionais, a biblioteca *llama.cpp* fornece os componentes necessários de aprendizagem de máquina para executar modelos de LLM e permite facilitar o carregamento deles (em formato GGUF V2) e executá-los com uso exclusivo da CPU (com GPU integrada). Essa estratégia de operação permite reduzir o custo computacional pela diminuição da alocação de memória, aumentando a velocidade de inferências e preservando ainda uma parte significativa da performance do modelo (LABONNE, 2023).

A configuração adotada teve como ponto de partida um modelo não interativo, com um número de contextos fixado em 2048 *tokens*, tendo um limite superior estabelecido em 4096 *tokens*. O método de quantização empregado foi o de 4-bits, especificamente o Q4_K_M que se caracteriza por utilizar a matriz de chaves Q6_K em metade das operações da técnica de atenção, bem como em tensores e na arquitetura de rede neural *feedforward*. Esta abordagem de quantização é recomendada em situações em que há uma necessidade premente de equilibrar a qualidade do modelo com a redução da precisão dos pesos durante o processo de quantização. É particularmente adequada para casos de uso em que esses dois fatores são críticos e precisam ser cuidadosamente balanceados (XIAO et al., 2023; LIU et al., 2023)

Com o propósito de simplificar a gestão dos dados, decidiu-se pela utilização do pacote *llama-cpp-python*, o qual concede acesso de baixo nível às funcionalidades da API em linguagem C, bem como oferece recursos de alto nível para a inserção de texto. Além disso, esse pacote demonstra compatibilidade com o *framework LangChain*, possibilitando a sua integração em um ambiente baseado na linguagem de programação Python.

4.2 Conjunto de dados

O repositório do *StatsBomb Open Data*, que compartilha gratuitamente dados de partidas de futebol e incentiva a criação de pesquisas nesse segmento, serviu como fonte de extração para compor os eventos de entrada. A base de dados está disponível em formato JSON e acompanha documentos que auxiliam a compreendê-la e utilizá-la. A disposição dos arquivos segue o arranjo descrito abaixo.

competição: contém as informações gerais sobre a competição e a partida, como nome/ID da competição e temporada, modalidade entre outros.

partidas: identificado pelo ID da competição e temporada, contém a data da partida, time da casa, time adversário, nome do treinador, nome do estádio, pontuação final etc.

eventos: Este documento contém os eventos individuais, os quais foram categorizados por identificadores (ID's) e indexados na ordem em que se manifestaram durante a partida. Cada evento é acompanhado por informações genéricas acerca do momento em que ocorreu e de sua posição no campo, bem como por atributos específicos correlacionados ao tipo de ação empreendida.

4.3 Preparação do experimento e configuração dos parâmetros

Esta seção contém as etapas que decorrem para elaboração do experimento, incluindo todas as configurações dos parâmetros referente aos grupos e conjuntos, e seleção dos eventos esportivos.

4.3.1 Seleção dos eventos esportivos

Essa etapa decorre de uma amostragem probabilística estratificada que busca formar 12 grupos a partir de 16 diferentes ações durante a partida de futebol. Esses grupos são compostos por subconjuntos contendo pares de eventos que se relacionam ou não entre si. Para alguns grupos, a associação dos pares traz uma lógica de ação e reação: um evento de chute estar conectado a uma ação do goleiro, ou um passe de bola estar relacionado a recepção da bola por outro jogador. Na base de dados isso torna-se claro, uma vez que um evento de ação traz consigo o identificador do evento de reação. A demonstração desse esquema é evidenciado na Figura 3. Tal escolha se justifica uma vez que gerar narrações de um evento isolado poderia trazer problemas de coerência para a narrativa final.

A definição da amostra a ser avaliada segue o cálculo sugerida por GIL (2008) com o intuito de estabelecer limites amostrais em uma população de natureza finita. O tamanho da população em questão corresponde ao número de eventos ocorridos em uma única partida, totalizando 3.904 eventos. Para fins de análise, considerou-se uma margem de erro de 7,90% e um nível de confiança de 90%. Devido à característica heterogênea da distribuição dos eventos, o efeito deste cálculo resultou em aproximadamente 105 eventos a serem avaliados.

A triagem dos tipos de lances, posteriormente categorizados em grupos, foi realizada com base na periodicidade com que ocorrem durante a partida. Isso se deve ao fato de que eventos mais frequentes são escolhas preferenciais para análise, uma vez que refletem a predominância das ações no jogo. As Figuras 4 e 5 expõem como se deu a formação dos grupos a partir das interações entre os diferentes tipos de ações durante a partida. É possível notar que um mesmo gênero de evento pode estar contido em grupos diferentes, porém em nenhuma hipótese representam acontecimentos idênticos.

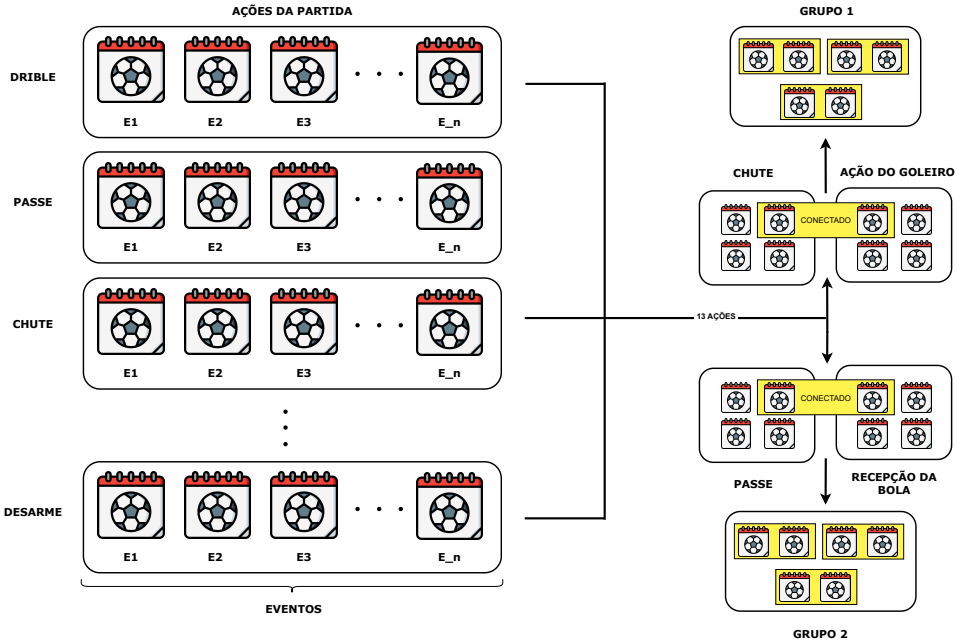

Figura 2: Estratificação das ações da partida. Os grupos foram formados por conjuntos de combinações entre dois eventos que estão relacionados entre si. **Fonte:** Autor (2023).

4.3.2 Configuração dos parâmetros

Os parâmetros que regulam as condições de saída do texto final no sistema foram configurados de forma distinta, dividindo-os em dois conjuntos: 1) narrativa criativa e mais aleatória e 2) narrativa controlada e mais determinística. Essa separação pode ser vista na Figura 6. O Conjunto 1 compreende eventos que requerem uma narrativa mais emotiva e é composto, neste contexto, pelas ocorrências relacionados a chutes ao gol, incidentes com penalidade, inicio/fim da partida e impedimento. Para esse grupo, foram estabelecidos valores de temperatura de amostragem maiores, fixados em 1,2 (com um limite superior de 1,5), juntamente com métricas de precisão top-k = 60 (com um limite superior de 100) e top-p = 0,85 (com um limite superior de 1,0).

No caso do grupo com narrativa controlada, o objetivo foi alcançar um equilíbrio desses parâmetros, utilizando valores de temperatura de amostragem fixados em 0,8, juntamente com métricas de precisão estabelecidas em top-k= 40 e top-p = 0,60. Para ambos os conjuntos, foi aplicada uma penalidade de repetição abaixo do padrão. Não foi imposta qualquer restrição quanto ao limite máximo para a geração de *tokens* de saída, a fim de garantir que nenhuma informação relevante fosse omitida durante o processo de geração de texto.

4.3.3 Processamento das informações

As informações dos eventos f, resultando em três grupos distintos: 1) atributos globais, 2) generalistas e 3) específicos. Os atributos globais, devido à sua natureza abran-

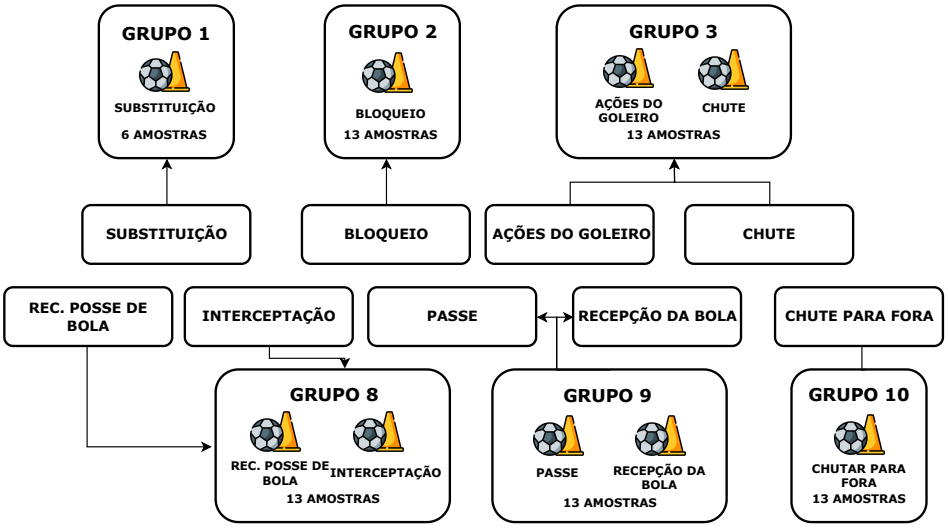

Figura 3: Parte 1: Categorização dos grupos com baseado nos eventos esportivos. Os eventos listados apresentam boa periodicidade na partida e englobando uma parte proveitosa da população total. **Fonte:** Autor (2023).

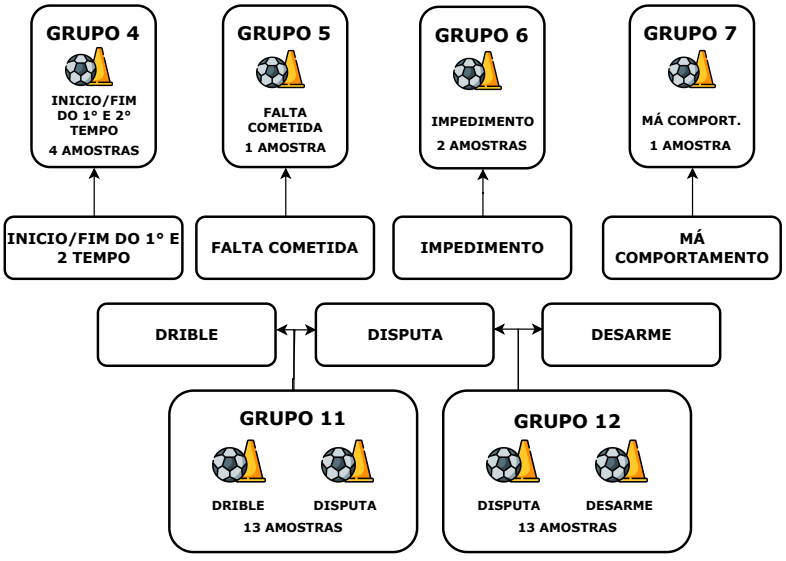

Figura 4: Parte 2: Categorização dos grupos com baseado nos eventos esportivos. A disparidade na amostragem no grupo 4 é justificado pela falta de eventos, uma vez que essa ação ocorre com pouca frequência nos jogos, porém apresenta-se como ato relevante a ser analisado. **Fonte:** Autor (2023).

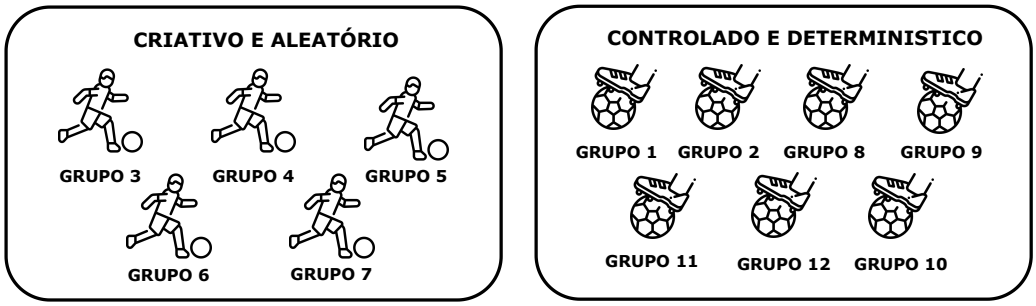

Figura 5: Divisão dos conjuntos com base na natureza da narração. A atribuição do conjunto revela os parâmetros que serão configurados na saída do modelo. O conjunto a esquerda representa o grupo com narrativa criativa. Já o conjunto a direita é composto pelo grupo de narrativa controlada
Fonte: Autor (2023).

gente em relação ao contexto da competição, foram incorporados diretamente ao *prompt*, logo após as instruções referentes à persona. Os atributos generalistas e específicos foram estruturados em um dicionário chave-valor para facilitar a indexação dos eventos e foram concatenados imediatamente após a solicitação de geração da narrativa esportiva. As características generalistas são consideradas elementos comuns a todos os eventos, pois fornecem informações básicas, já as específicas têm particularidades que variam dependendo do evento em questão. Tais peculiaridades desempenham um papel fundamental no enriquecimento da narrativa, conferindo-lhe maior detalhamento e contextualização.

1) Atributos globais: nome da competição, gênero da competição, nome da temporada, nome do país, nome do time da casa, gênero do time da casa, país do time da casa, treinador do time da casa, nome do time visitante, gênero do time visitante, país do time visitante, treinador do time visitante, pontuação final do time da casa, pontuação final do time visitante, estádio e nome do árbitro.

2) Atributos generalistas: tempo, evento, padrão, time, jogador e posição do jogador.

3) Atributos específicos: *início/fim do 1° e 2° tempo:* apenas atributos generalistas (com exceção do nome e posição do jogador); *ação do goleiro e chute:* técnica, parte do corpo, saída, posição do goleiro, tipo do chute, ação do goleiro, time do goleiro; *passe e recepção da bola:* altura do passe, parte do corpo, jogador receptor, time do receptor e posição do receptor; *drible e disputa:* ação do drible, se está sob pressão, enfrentamento, oponente da disputa, time do oponente da disputa e resultado do enfrentamento; *disputa e desarme:* tipo do enfrentamento, se está sob pressão, oponente da disputa, time do oponente da disputa e resultado do enfrentamento; *substituição:* motivo da substituição e nome do jogador que entrou no campo; *recuperação da bola e interceptação:* time que está com a bola, resultado da interceptação, se a bola foi recuperada, time com posse após o evento; *bloqueio:* apenas atributos generalistas; *chutar para fora:* parte do corpo. *falta*

cometida: tipo do cartão de penalidade; *impedimento:* jogador receptor, time do receptor e posição do receptor; *má comportamento:* tipo do cartão de penalidade;

4.3.4 Configuração do *prompt* de entrada

No contexto do presente trabalho, o objeto de entrada para o modelo foi concebido com base nas observações do procedimento de treinamento do modelo LLaMA-2, conforme delineado no estudo conduzido por Touvron et al. (2023). O *prompt* de entrada é apresentado com uma estrutura definida: as etiquetas [INST] e [/INST] delimitam a área de interação entre os eventos de entrada e o assistente, que assume a forma do narrador esportivo. Além disso, as marcações $<< SYS >>$ e $<< /SYS >>$ estabelecem o espaço destinado à inserção do texto de contexto, o qual orientará o sistema na execução da tarefa como é observado na Figura 2. Embora outras abordagens de formatação de *prompts* possam ser adotadas, é importante ressaltar que resultados mais favoráveis tendem a ser alcançados quando se mantém uma conexão direta e coerente com a estrutura empregada durante a fase de treinamento do sistema LLaMA (SCHMID et al., 2023).

Para configurar o comportamento do sistema, optou-se por fornecer determinadas diretrizes que são recomendadas por Touvron et al. (2023) para LLaMA 2-Chat, ChatGPT, PaLM-chat e Falcon, buscando definir uma identidade e persona de comentarista esportivo ao modelo, além de mitigar a geração de conteúdo ofensivo, perigoso, racista, ilegal, entre outros. O texto a seguir exibe um trecho das informações de contexto repassada ao sistema.

> *"You are a helpful, respectful, and honest assistant. Your personality is a sportscaster who has total knowledge about soccer techniques, the ability to improvise, and a great vocabulary avoiding repeat words. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive."* [1]

4.4 Métricas de avaliação dos resultados

Esta seção apresenta as análises quantitativas realizadas sobre o experimento e como decorreu-se a aplicação delas. As avaliações qualitativas e discussões serão apresentadas no capitulo de resultados.

[1] Tradução: "Você é um assistente prestativo, respeitoso e honesto. Sua personalidade é de um comentarista esportivo que possui total conhecimento sobre técnicas de futebol, habilidades de improviso e um ótimo vocabulário evitando repetir palavras. Sempre responda da maneira mais prestativa possível e segura. Suas respostas não podem incluir qualquer conteúdo prejudicial, antiético, racista, sexista, tóxico, perigoso ou ilegal. Por favor, certifique-se de que suas respostas sejam socialmente imparciais e positivas."

```
<s>[INST] <<SYS>>
{{info_persona} {atributos_globais}}
<</SYS>>
{{ (comando_narracao): (descricao_evento)
+ (atributos_generalista)
 + (descricao_evento) + (atributos_especificos) }}
```

Figura 6: Estrutura para o comando de entrada no modelo. A organização segue um modelo próximo ao sugerido pelos desenvolvedores do LLaMA, com a adição dos atributos e respectivas descrições. **Fonte:** Autor (2023).

Índice de Flesch de Facilidade de Leitura: Para esse indicador, fez-se o da biblioteca *textstat* em linguagem *Python* que oferece cálculos estatísticos para determinar a complexidade, grau de leitura e facilidade dos textos e em seguida os valores foram categorizados em níveis de compreensão de leitura.

Menção do atributo de entrada: *Essa métrica deve ser avaliada a partir da quantidade de atributos citados no texto final. A contagem ocorreu de maneira manual e levou em conta a aplicação correta do termo, com o valor final dado em porcentagem.*

Avaliação dos erros: *As narrativas geradas devem ser categorizados em 2 grupos: correto e incorretos e a avaliação final apresentada de maneira binária.*

5 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Esta seção apresenta os resultados desse estudo obtidos através do percurso metodológico descrito no Capítulo 4. Serão exploradas análises comparativas, demonstração de métricas de estatística descritiva e associação de variáveis segredados por grupos e conjuntos a fim de identificar padrões de comportamento, falhas e qualidades do modelo. Buscou-se, também, verificar marcadores de linguagem do padrão narrativo, variações no uso da linguagem nos conjuntos e as principais falhas do modelo.

5.1 Análise comparativa entre os grupos

A análise do Índice de Flesch de Facilidade de Leitura (IFFL) em relação aos grupos revelou resultados consideráveis uma vez que possuem uma distribuição de frequência que varia com médias para os 12 grupos entre 65,32 - 86,87, como pode ser visto na Figura 7. Essas médias foram classificadas nas categorias “Inglês Comum”, “Bastante Fácil” e “Fácil”. Notavelmente, nenhum dos grupos apresentaram classes de complexa compreensão como “Difícil” e “Muito Difícil” (IFFL abaixo de 50). Este fato evidencia que as narrações geradas surtiram efeitos positivos para esse indicador e que podem ser facilmente compreendidas pela maioria dos ouvintes dentro de um estádio. É importante deixar claro ainda que o Grupo 5 e 7 não aparecem no gráfico, uma vez que eram compostos por um único ponto, não sendo possível representá-los com o *boxplot*. Esses grupos representam, respectivamente, ações de falta cometida e má comportamento, situações essas que ocorreram apenas um vez na amostra analisada, porém se mostraram como eventos importantes a serem narrados.

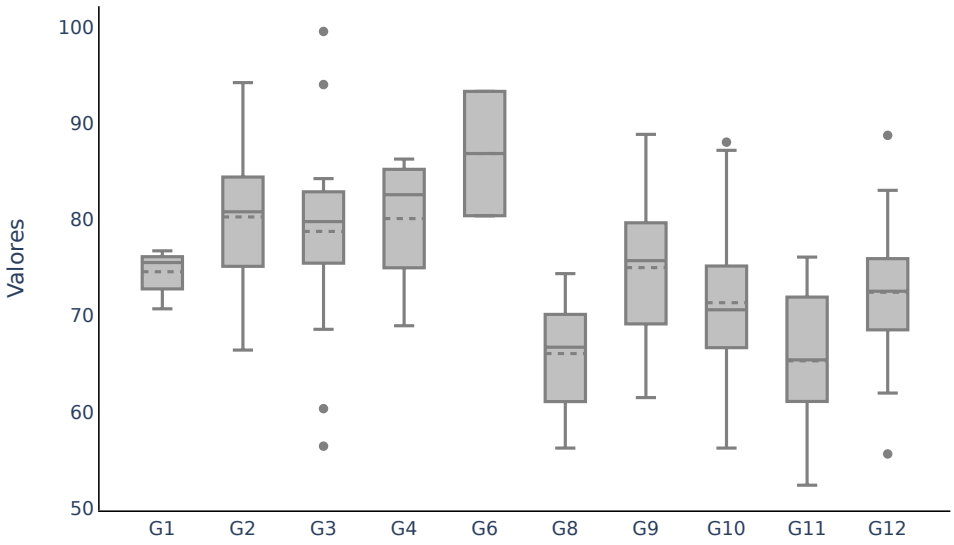

Figura 7: Gráfico de boxplot contendo medidas de posição e dispersão do IFFL É exibido a distribuição dos níveis de compreensão do texto por grupo com escalas de 0 a 100% **Fonte:** Autor (2023).

Acerca da dispersão dos valores de IFFL, é possível observa que os Grupos 2, 9 e 10 apresentam as maiores variações mostrando que existem flutuações no nível de compreensão do texto de saída. Um padrão encontrado nesses três grupos é o fato de possuírem 13 atributos de entradas e em algumas amostras o modelo repetir a informação já apresentada, no final do texto, só que dessa vez de maneira literal, influenciando o algoritmo do IFFL. O Grupo 3 também se destaca por possuir quatro *outliers*: as duas variações positivas são de amostras com textos pequenos, o que pode acarretar uma fácil compreensão, visto que o IFFL leva em conta a quantidade de palavras por frases, por outro lado os motivos para as variações negativas não ficaram claros, podendo ser atribuído a falhas intrínsecos do modelo. De maneira semelhante, o Grupo 12 apresenta pontos anômalos: o valor atípico negativo tem essa condição já que a amostra possui uma frase longa dentro do texto. O *outilier* positivo possui um texto menor em que todas as frases detêm um tamanho regular dentro do texto. O Grupo 6 também ganha visibilidade por possuir a melhor das médias entre os grupos. Esse fato dar-se por igualmente dispor de conteúdo com frases reduzidas, desta vez ainda mais curta.

Em relação a menção dos atributos, isto é, a proporção entre a quantidade de atributos informados e o aparecimentos deles durante a narração, observa-se que para vários grupos se verificou uma abrangência maior que 60% e para alguns outros valores próximos como pode ser visto na Figura 8. A referência ao atributo de entrada no texto é importante para caracterizar o evento e favorecer o entendimento sobre o que aconteceu. O Grupo 4, que contém o menor número de atributos iniciais (como evidenciado na seção 4.2.2) apresentou a maior discrepância, tendo uma amostra com alta cobertura e outra zerada. Para o caso negativo, o modelo deixou de informar dados importantes e para o caso positivo criou eventos fictícias no texto. Isso pode demonstrar que um baixo número de atributos iniciais pode dar espaço para o modelo fugir do foco narrativo. Todavia, não é possível validar essa hipótese, para esse estudo, a partir de um teste de significância estatística em detrimento a baixa quantidade de dados amostrais. No cenário em que esse empasse fosse suprida, tornaria-se possível aprofundar esse entendimento a partir do Teste Exato de Fisher ou Teste Qui-quadrado, com a hipótese de existe correlação entre o número de atributos de entrada e os erros do texto narrativo.

Uma investigação adicional foi conduzida com o propósito de elucidar se a omissão de um atributo de significativa relevância para a narrativa de um evento esportivo acarretaria consequências no resultado final. Dessa forma, optou-se por selecionar, de forma aleatória, o Grupo 11, compreendendo os eventos relacionados a dribles e disputas, e procedeu-se à exclusão do nome do jogador adversário envolvido no embate. A análise centrou-se na avaliação da taxa de erro entre o grupo de controle, contendo todos os atributos, e o grupo experimental, no qual excluiu-se um atributo importante. Diferente do que se esperava, para maioria dos casos o modelo não trouxe um nome fictício para o

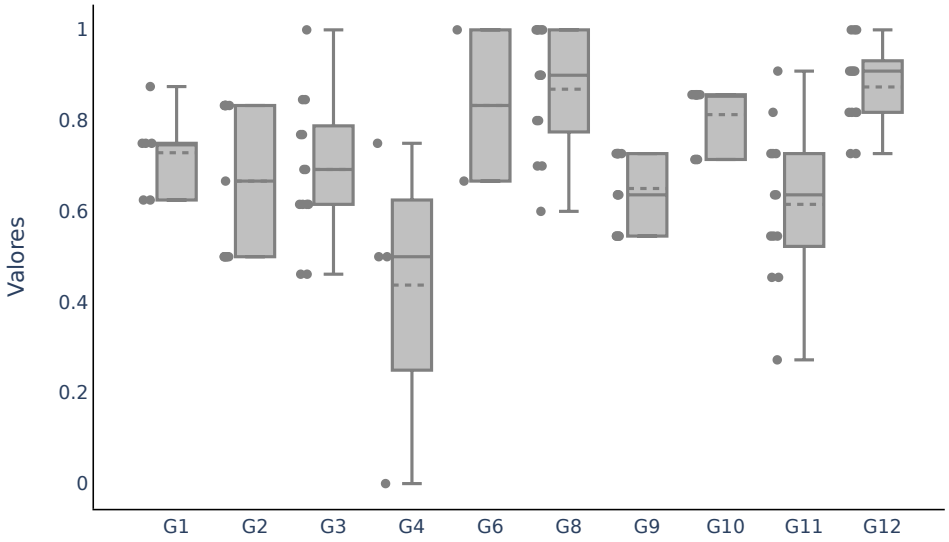

Figura 8: Gráfico de boxplot contendo a frequência atributos nas amostras É evidenciado uma alta frequência de amostras com mais de 60% de citação de atributos no texto final. **Fonte:** Autor (2023).

jogador (com o nome excluído inicialmente), pois é comum que o sistema preenche as lacunas faltantes a partir de padrões apreendidos com base na probabilidade de determinadas palavras aparecerem em sequência e, nesse caso, os eventos de drible sempre informam quem é o adversário. Na prática, o LLaMA apenas omitiu essa informação ou trouxe um nome genérico, errando somente em 15,39% dos casos como mostra a Figura 9.

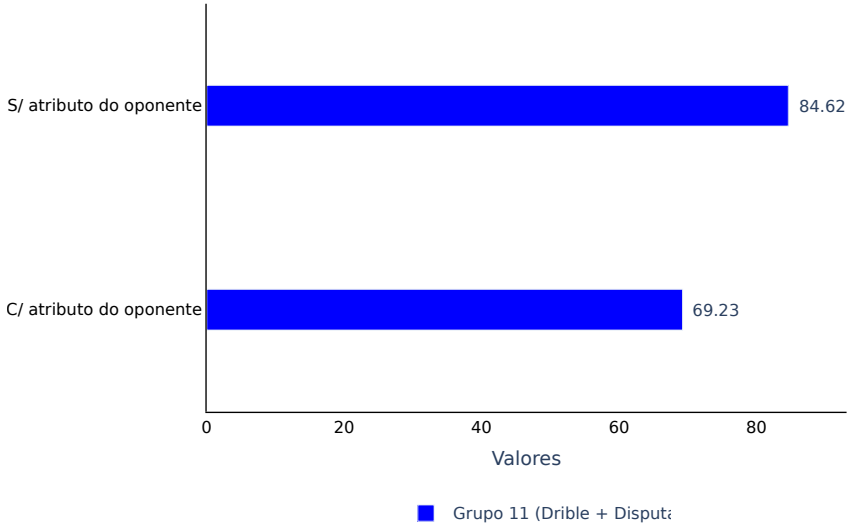

Figura 9: Gráfico de barras comparando o mesmo evento a partir de duas perspectivas: com e sem o atributo do oponente A adição do atributo não se mostrou tão importante com um acréscimo de apenas 15% na taxa de acerto **Fonte:** Autor (2023).

5.2 Análise comparativa entre os conjuntos

A distribuição de frequência do Índice de Facilidade de Leitura sob os conjuntos é evidenciada no histograma da Figura 10, que indica os primeiros contrastes entre os

conjuntos observados: 1) conjunto criativo e mais aleatório, e 2) conjunto controlado e determinístico. O Conjunto 2 segue uma curva com uma maior concentração entre 70-75 apresentando um bom ajuste a uma distribuição normal. Já o Conjunto 1 apresenta uma assimetria à direita demonstrando que esse conjunto de amostras dispõe de melhores níveis de compreensão de leitura. Essa visão pode ser justificada dado que o segundo conjunto tem parâmetros que trazem uma narrativa mais criativa e aleatório, tornando também o texto mais coeso e, por consequência mais fácil de ler. Em contrapartida, a taxa de acerto na narração foi afetada, visto que nesta coleção também foram identificados textos mais extensos, com uma maior inserção de informações fictícias e uma maior difusão da narrativa principal, caso semelhante ao enfrentando por Czaplicki (2023) com o GPT-3.5. Essa investigação traz à tona a necessidade de atentar-se a associação de que o modelo pode gerar respostas que são plausíveis do ponto de vista linguístico, porém que não refletem necessariamente a realidade.

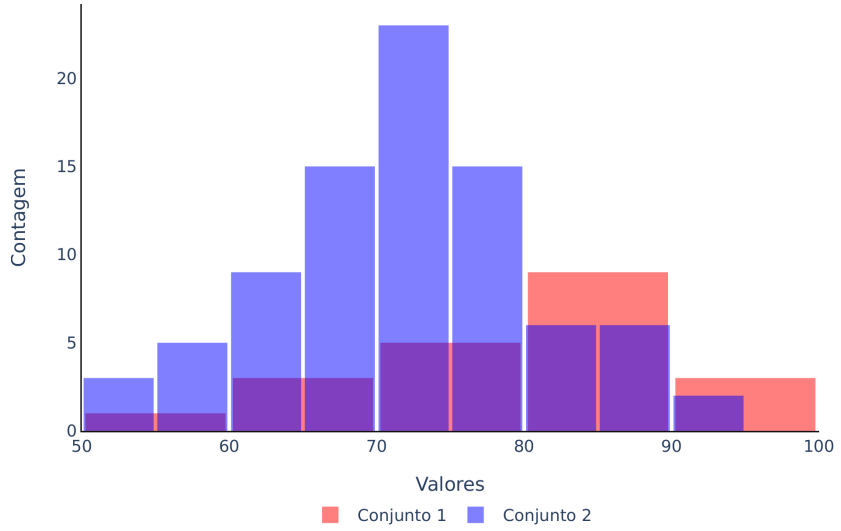

Figura 10: Histograma com a distribuição de frequência entre os conjuntos Revela-se um padrão para os dois conjuntos: o primeiro assumindo uma distribuição normal e o segundo com assimetria à direita. **Fonte:** Autor (2023).

A disposição dos índices de erro nas amostras é demonstrado na Tabela 5 evidenciando uma taxa de 71,43% para o Conjunto 1, comparado aos 55,95% do Conjunto 2. Isso reforça a necessidade de equilibrar os parâmetros iniciais do sistema, especialmente a temperatura de amostragem e métricas de top-k e top-p, podendo assim resultar em repercussões mais favoráveis.

Acerca das variações no uso da linguagem entre os conjuntos, pode-se verificar que o Conjunto 1 apresentou o uso de letras em caixa alta para expressar animação, reticências para gerar suspense na narração e uso de exclamação para dar destaque para àquele acontecimento. Já no Conjunto 2 é possível visualizar um texto que também traz um positivismo através das palavras, porém há uma atenção muito maior em detalhar as informações de entrada. Em ambos os conjuntos foram encontradas frases coloquiais e a

		Narração correta?				
		Sim		Não		
	F_{total}	F_i	$Fr_i(\%)$	F_i	$Fr_i(\%)$	
Grupo						**Total**
Conjunto 1	21	6	28,57	15	71,43	21 (100%)
Conjunto 2	84	37	44,05	47	55,95	84 (100%)
Total	105	43	40,95	62	59,05	105 (100%)

Tabela 5: Avaliação da taxa de acerto para os conjuntos. O conjunto 1 apresentou uma taxa maior de erro em comparação ao conjunto 2 **Fonte:** Autor (2023).

expressão *Oh my!*, para indicar surpresa. Essa forma reduzida da expressão é vista com maior frequência entre os falantes. Outra característica marcante em ambos os conjuntos foi a presença de *emoticons* para expressar sentimentos durante a narração, porém não foi possível verificar uma maior tendência no surgimento para um determinado grupo. A seguir é possível visualizar exemplos de narração para ambos os conjuntos.

> **Conjunto 1:** *"Oh my, what a beauty of a pass from Ivan Rakitić! The Barcelona midfielder, with the precision of a surgeon, finds Lionel Messi with a perfectly weighted ground pass, right on the money! Messi, with his quick feet and lightning speed, takes control of the ball and immediately starts to weave his magic! The away team, Real Sociedad, are left scratching their heads as Messi glides past their defenders with ease! What a start to the match, folks!"* [2]

> **Conjunto 2:** *"Oh my, what a fantastic opportunity for Barcelona's Gerard Piqué Bernabéu! He's got the ball from a corner kick and he's going for a half-volley shot with his right foot! Piqué sets himself up beautifully, using his right foot to shoot... BUT IT'S SAVED BY THE REAL SOCIEDAD GOALKEEPER, CARLOS DEL CERRO GRANDE!"* [3]

5.3 Caracterização do narrador esportivo

A narração esportiva de futebol se configura como um gênero discursivo que apresenta uma variedade grande no estilo de narrar e na estrutura textual (SANTOS, 2012),

[2] **Tradução:** Nossa, que lindo passe de Ivan Rakitić! O meio-campista do Barcelona, com a precisão de um cirurgião, encontra Lionel Messi com um passe rasteiro perfeitamente ponderado, acertando em cheio! Messi, com seus pés rápidos e velocidade relâmpago, assume o controle da bola e imediatamente começa a tecer sua magia! A equipe visitante, a Real Sociedad, fica coçando a cabeça enquanto Messi passa por seus defensores com facilidade! Que começo de partida, pessoal!

[3] **Tradução:** Nossa, que oportunidade fantástica para Gerard Piqué Bernabéu, do Barcelona! Ele pegou a bola na cobrança de escanteio e vai dar um chute de meio voleio com o pé direito! Piqué se posiciona lindamente, usando o pé direito para chutar... MAS É SALVO PELO GOLEIRO DO REAL SOCIEDAD, CARLOS DEL CERRO GRANDE!

Evento	Tempo	Comentário
Passe e rec. bola	30”	Enfim, de volta à ação, pessoal!
Passe e rec. bola	3’19”	...apenas um fio de cabelo acima do solo, você pode ver a bola beijando a grama ali!
Drible e Disputa	48”	Vai ser um ótimo jogo, pessoal, fiquem ligados!
Drible e Disputa	39”’	O mágico argentino é conhecido por seu incrível controle de bola
Disputa e desarme	20”46”’	Lembrem-se pessoal, não se trata apenas do placar, mas da paixão e determinação desses atletas
Disputa e desarme	26”23”’	...mas Alba Ramos está em cima dele como um terno barato!
Disputa e desarme	26”23”’	Fiquem ligados, pessoal, vai ser uma viagem louca!
Disputa e desarme	28”29”	A Real Sociedad não vai desistir - eles estão determinados a levar para casa a vitória hoje!
Substituição	27”’26	...você pode apostar seu último dólar nisso
Bloqueio	7”38”’	Fiquem de olho nisso, pessoal
Bloqueio	14”9”’	A multidão vai à loucura
Bloqueio	15”’11	...passou pela defesa do Barcelona como uma faca quente na manteiga
Interceptação	6”30”’	Será que a Real Sociedad conseguirá se recuperar e empatar o jogo ou o Barcelona aguentará a vitória?
Interceptação	23”24”’	...e eles não vão desistir tão cedo!
Interceptação	46”02”’	Eles serão capazes de capitalizar essa reviravolta?
Chute para fora	25”19”	A multidão na Reale Arena está enlouquecendo
Inicio/fim 1º e 2 T	7”12”’	e cara, eles estão dando um show!

Tabela 6: Principais frases característica da narração esportiva extraído da saída do LLaMA. As frases marcam o estilo de narração apropriada para o contexto, além de tentar transmitir emoções. **Fonte:** Autor (2023).

contudo há um padrão que se assemelha a narrativas tradicionais por conter fatos relacionados, personagens, além do objetivo principal que é a exposição de uma história (SANTOS, 2010). Essas características puderam ser observadas nas narrações geradas pelo modelo, revelando que os textos estão condizentes com o proposito esperado para a persona de um narrador. Além disso, marcadores no discurso como jargões, analogias e indagações também puderem ser visualizadas, trazendo ainda mais qualificação ao resultado. A Tabela 6 exibe algumas frases extraídas do experimento para atestar esse comportamento. As expressões foram traduzidas para o português a fim de facilitar o entendimento do leitor, porém não houve perda semântica no textos após esse processo.

5.4 Estilo de narrativa do modelo

Os comentários narrativos geradas pelo LLaMA-2 apresentaram características discutíveis e que se assemelham a outros modelos de linguagem. A principio foi observado que o modelo adicionava informações que não foram inseridos no comando de entrada.

Em algumas amostras têm-se a adição da nacionalidade dos jogadores, cidade em que está situado o estádio e as cidades de origem dos clubes. Para jogadores famosos, como o Lionel Messi, na época no Barcelona, o sistema trazia muitas qualidades positivos e pessoais para o jogador como "o grande", "mágico argentino", "a estrela do Barcelona", característica que não foi observada com frequência nas narrações para outros jogadores.

Outra propriedade encontrada foi a omissão de informações que já são padrões durante a partida, para parte das amostras, como o caso de uma partida regular ou um jogo aberto. A referência acontecia com maior frequência quando era um dado diferenciado como a ocorrência do evento durante um escanteio, tiro de meta, lançamento lateral entre outros. Existe também uma supressão de um dos nomes dos jogadores quando o evento era entra pessoas do mesmo time, substituindo por expressões como "colega" ou "companheiro de equipe". Mostrou-se interessante também a variação no modo como o tempo era exibido no texto, sendo em algumas vezes idêntico a entrada (hh:mm:ss.ms), em outras situações de forma suprimida, por exemplo 40 minutos, ou ainda mais reduzido como 40 min. Tal redução também ocorria quando era citada a posição do jogador reduzindo, por exemplo, meio-campista centro defensivo para meio-campista.

Determinadas características negativas também foram notadas, como falsear o placar para diversos tipos de eventos, principalmente o evento de chute ao gol. Por isso, torna-se interessante informar o placar atualizado em todos as amostras relacionadas a ações de chute ao gol, estratégia que não foi empregado nesse estudo pela ausência dessa informação na base de dados. Também ocorreu do modelo criar textos muito extensos, acrescentando informações dispensáveis para o entendimento geral do ocorrido, além de tornar o conteúdo entediante. O estudo de Kameko et al. (2015) traz resultados opostos, uma vez que no modelo proposto, a saída continha poucas informações, carecendo de mais detalhamento. Esse aspecto trouxe uma dificuldade relevante na análise de erros das narrações: a maioria dos eventos traziam informações corretas acerca dos atributos, porém o acréscimo de informação extra e incorreta ao contexto fez com que a amostra fosse qualificada como errada. Algumas falhas de ocorrências pouco comuns englobam amostras em língua espanhola, ao invés do inglês, e a falta de entendimento do modelo sobre alguns termos citados. Mostra-se interessante adicionar um glossário envolvendo termos mais complexos, contudo não foi vislumbrado nesse estudo pela limitação de *tokens* no *prompt* de entrada.

5.5 Principais falhas do modelo

As imprecisões verificadas nas 62 amostras com incoerências foram categorizadas na Tabela 7 a partir do tipo do erro e, após isso, avaliada a distribuição de frequência. Torna-se nítido que a ação do LLaMA-2 em trocar a posição do jogador pelo espaço do campo em que ocorreu a partida se apresentou como a falha de maior frequência. Esse equívoco pode

Tipo de erro	f_i	F_i	fr_i	Fr_i
Informou o placar errado	13	13	20,98	20,98
Trocou a posição do jogador com espaço do campo	15	28	24,19	45,17
Inventou um oponente que não foi citado	3	31	4,84	50,01
Errou o resultado do evento	11	42	17,74	67,75
Informou o evento errado	11	53	17,74	85,49
Adicionou informações extras erradas	7	60	11,29	96,78
Outros	2	62	3,22	100
Total	**62**		**100%**	

Tabela 7: Categorização dos erros em todas as amostras avaliadas como incorretas no experimento. As inconsistências foram avaliadas com base nas frequência absoluta, relativa e os respectivos acúmulos para ambas. **Fonte:** Autor (2023).

ter ocorrido visto que um jogador na posição de meio-campista naturalmente efetua ações nesse local, porém não é uma regra e essas amostras precisaram ser caracterizadas como erro. O autor Kameko et al. (2015) também relata esse tipo de inexatidão no experimento realizado com jogos de Shogi, porém a proporção desse erro no ensaio dele é menor. Uma técnica que pode ser aplicado para atenuar esse empasse é o mecanismo de espaço reservado utilizado por Taniguchi et al. (2019), uma vez que o trabalho desses autores sofreu com um problema parecido. Para o caso desse estudo, existiria a codificação da posição do jogador e do campo de forma distinta e com termos genéricos e, em sequência, a decodificação para o nome correto. Outro erro significativo estar relacionado a informar o saldo de gols de forma incorreta. Esses equívocos podem está associados a dificuldade que os modelos de LLM tem em lidar com aritmética simples, expressões matemáticas ou até mesmo lógica básica como relata Plevris et al. (2023) em uma investigação conduzida para aferir o desempenho desses sistema frente a problemas matemáticos. Uma dificuldade adicional encontrada no modelo foi a falta de associação entre frases do mesmo texto onde, em determinados momentos, o modelo gerou informações corretas e nas frases seguintes ocorria uma contradição. Esse fato mostra a dificuldade do LLaMA 2 em manter o contexto e associar as ideias no texto de saída, aspecto que é um desafio para diversos modelos de linguagem.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Esse estudo buscou conduzir experimentos para aferir a capacidade do modelo, comparar os desempenho com diferentes parâmetros, avaliar a qualidade das narrações esportivas geradas e identificando padrões e falhas do LLaMA-2. Constata-se que o objetivo geral e os específicos foram atendidos partindo da hipótese de compreender o desempenho do modelo de linguagem da Meta AI usando dados estruturados de partidas de futebol e constituindo um *prompt* de entrada adequado ao sistema.

Acerca do confronto para narrações com diferentes parâmetros, evidenciou-se que o conjunto com padrão de temperatura de amostragem, top-p e top-k, mais elevados demonstrou melhores valores de compreensão de leitura. Contudo, essa mesma coleção expressou baixos níveis de acerto pela criação de informações fictícias. Em contrapartida, o conjunto com valores mais baixos não apresentou com frequência tais limitações, porém o texto de saída transmitia um menor nível de emoção ao leitor. Assim, o ajuste equilibrado dos parâmetros é essencial para bons resultados. Sobre a qualidade narrativa em termos de legibilidade, grupos contendo frases mais curtas ganharam destaque. No entanto, os grupos de maneira geral apresentaram diversidade significativa, necessitando criar limitadores de palavras e/ou frases no texto. Os padrões de comportamento identificados na pesquisa contemplam o acréscimo de informações extras a narrativa, a omissão de dados redundantes e a capacidade de reduzir termos para facilitar e melhorar a entendimento do leitor. Mesmo sendo pontos positivos, essas características aumentaram a taxa de erro por falsear informações como a adição de um placar incorreto. Sobre a persona de narrador esportivo, o modelo cumpriu essa meta por trazer um estilo de narração próprio do gênero, com jargões, analogias e marcadores de indagação e surpresa.

Com base nos resultados apresentados, observou-se que o modelo de linguagem LLaMA-2 demonstrou resultados interessantes para narração automatizada de jogos esportivos, embora apresente algumas limitações. O problema, no entanto, necessita ainda de mais pesquisas e aperfeiçoamentos do sistema pelos desenvolvedores. Por consequência, trazendo incentivos para a automatização da narração esportiva, principalmente, em jogos de esportes que não recebem destaque na mídia ou investimento, atraindo mais facilmente narradores humanos e especialistas sobre o assunto.

6.1 Trabalhos futuros

1. **Ampliar as amostras do experimento:** Verificou-se que algumas testes não puderam ser aplicados devido a insuficiência na quantidade de dados analisados, como a exemplo a verificação da correlação entre a quantidade de atributos de entrada e a taxa de acerto do modelo. Assim, se faz necessário esse incremento.

2. **Experimento com mais grupos:** O teste para verificar se a omissão de uma informação importante afeta o desempenho do LLaMA pode ser expandido para mais grupos, a fim de identificar outros padrões.

3. **Adição de um glossário:** O modelo apresentou dificuldade com alguns termos, como o caso da troca do termo *block* por *blockbuster*, evidenciando a necessidade de explicar ao sistema esse vocabulário.

4. **Informar o placar:** Anunciar o placar atualizado a cada evento de chute ao gol pode se mostrar como um incremento interessante para evitar que ocorra o falseamento dessa informação.

5. **Correlacionar todos os eventos:** É interessante que o LLaMA conheça todas as ações anteriores para operar em tempo real. Todavia, o modelo possui uma limitação de *tokens* para entrada, necessitando investigar a melhor essa característica do modelo.

REFERÊNCIAS

BRAZ, Matheus Petroni et al. ChatGPT x portais de notícia: um estudo de representação sobre a comunidade LGBTQIAP+ em conteúdos digitais.

BROWN, Tom et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877-1901, 2020.

CANAVILHAS, João. Artificial intelligence and journalism: Current situation and expectations in the Portuguese sports media. **Journalism and media**, v. 3, n. 3, p. 510-520, 2022

CHATGPT, Assistente; DA SILVA MONTEIRO, Jean Carlos. Assistente ChatGPT no jornalismo. **Cadernos da Escola de Comunicação**, v. 19, p. 32-44, 2023.

CHOWDHURY, G. G. **Natural language processing**. Annual review of information science and technology, v. 37, p. 51-89, 2003.

CZAPLICKI, Michał. **Live commentary in a football video game generated by an AI**. 2023. Trabalho de Conclusão de Curso. University of Twente.

DAVIS, Ryan et al. Evaluating the effectiveness of artificial intelligence–powered large language models application in disseminating appropriate and readable health information in urology. **The Journal of urology**, v. 210, n. 4, p. 688-694, 2023.

DE MEIRA LEITE, Mauana Simas. Narração Audiodescritiva e a experiência de pessoas com deficiência visual em estádios de futebol.

DO NASCIMENTO, Thiago Gomes; CORTIZ, Diogo. Avaliação do senso comum em modelos de linguagem através de benchmarks: Desafio de Winograd aplicado ao ChatGPT em português brasileiro. In: **Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana.** SBC, 2023. p. 193-198.

FAN, Wenqi et al. Recommender systems in the era of large language models (llms). **arXiv preprint arXiv:2307.02046**, 2023.

FRANGANILLO, Jorge; LOPEZOSA, Carlos; SALSE, Marina. La inteligencia artificial generativa en la docencia universitaria. 2023.

FLESCH, Rudolf. How to write plain English. University of Canterbury. 1979

GAUTAM, Sushant et al. Soccer game summarization using audio commentary, metadata, and captions. In: **Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos**. 2022. p. 13-22.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. Ediitora Atlas SA, 2008.

GODOY, Arlida Schmidt. Introdução à pesquisa qualitativa e suas possibilidades. **Revista de administração de empresas**, v. 35, p. 57-63, 1995.

HUYNH, Linda My et al. New Artificial Intelligence ChatGPT Performs Poorly on the 2022 Self-assessment Study Program for Urology. **Urology Practice**, p. 10.1097/UPJ. 0000000000000406, 2023.

HOFFMANN, Jordan et al. Training compute-optimal large language models. **arXiv preprint arXiv:2203.15556**, 2022

JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing (2nd Edition). 2008.

KAMEKO, Hirotaka; MORI, Shinsuke; TSURUOKA, Yoshimasa. Learning a game commentary generator with grounded move expressions. In: **2015 IEEE Conference on Computational Intelligence and Games (CIG).** IEEE, 2015. p. 177-184.

LABONNE, Maxime. Quantize Llama models with GGML and llama.cpp. Towards Data Science. Disponível em: https://towardsdatascience.com/quantize-llama-models-with-ggml-and-llama-cpp-3612dfbcc172. Acesso em: 10 out. 2023.

LEE, Greg; BULITKO, Vadim. Automated storytelling in sports: A rich domain to be explored. In:**Interactive Storytelling: Third Joint Conference on Interactive Digital Storytelling, ICIDS 2010, Edinburgh, UK, November 1-3, 2010. Proceedings 3**. Springer Berlin Heidelberg, 2010. p. 252-255.

LEWIS, Seth C.; SANDERS, Amy Kristin; CARMODY, Casey. Libel by algorithm? Automated journalism and the threat of legal liability. **Journalism & mass communication quarterly**, v. 96, n. 1, p. 60-81, 2019.

LIU, Zechun et al. LLM-QAT: Data-Free Quantization Aware Training for Large Language Models. **arXiv preprint arXiv:2305.17888**, 2023.

LÓPEZ, Karla María Gutiérrez. Inteligencia artificial generativa: Irrupción y desafíos. **Enfoques**, v. 4, n. 2, p. 57-82, 2023.

NATALE, Simone. If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. **new media & society**, v. 21, n. 3, p. 712-728, 2019.

NEVES, José Luis. Pesquisa qualitativa: características, usos e possibilidades. **Caderno de pesquisas em administração**, São Paulo, v. 1, n. 3, p. 1-5, 1996

PAPINENI, Kishore et al. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th annual meeting on association for computational linguistics.** Association for Computational Linguistics, 2002. p. 311–318.

PEÑA FERNÁNDEZ, Simón et al. Without journalists, there is no journalism: the social dimension of generative artificial intelligence in the media. 2023.

PLEVRIS, Vagelis; PAPAZAFEIROPOULOS, George; RIOS, Alejandro Jiménez. Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. **arXiv preprint arXiv:2305.18618**, 2023.

PRODANOV, Cleber Cristiano; DE FREITAS, Ernani Cesar. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição**. Editora Feevale, 2013.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial-Uma abordagem moderna.** 4º edição. São Paulo: Grupo GEN LTC, 2022.

SANTOS, Cristiane Alvarenga Rocha. A NARRAÇÃO ESPORTIVA DE FUTEBOL: análise discursiva de um fenômeno midiático. 2010.

SANTOS, Cristiane Alvarenga Rocha. Narração esportiva de futebol e composicionalidade: uma proposta de estudo textual-discursiva das sequências textuais (Soccer narration and compositionality: a proposal for textual-discursive study of textual sequences). **Estudos da Língua (gem)**, v. 10, n. 2, p. 31-48, 2012.

SANTOS, Cristiane Alvarenga Rocha. Narração esportiva de futebol no rádio: entre convenções e intenções. **Fórum Linguístico**, v. 9, n. 3, p. 215-229, 2012.

SALVAGNO, Michele et al. Can artificial intelligence help for scientific writing?. **Critical care**, v. 27, n. 1, p. 1-5, 2023.

SETH, Ishith et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. **Aesthetic Surgery Journal**, v. 43, n. 10, p. 1126-1135, 2023.

SCHMID, Philipp et al. Llama 2 is here - get it on Hugging Face. Hugging Face. Disponível em: https://huggingface.co/blog/llama2#demo. Acesso em: 15 out. 2023.

TALONI, Andrea; SCORCIA, Vincenzo; GIANNACCARE, Giuseppe. Modern threats in academia: Evaluating plagiarism and artificial intelligence detection scores of ChatGPT. **Eye**, p. 1-4, 2023.

TANAKA, Kumiko et al. MIKE: An automatic commentary system for soccer. In: **Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)**. IEEE, 1998. p. 285-292.

TANIGUCHI, Yasufumi et al. Generating live soccer-match commentary from play data. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. 2019. p. 7096-7103.

TOUVRON, Hugo et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023

TURING, Alan. **Computing Machinery and Intelligence**. Mind, vol. LIX, n. 236, 1950

WHITTAKER, Edward WD; RAJ, Bhiksha. Quantization-based language model compression. In: **INTERSPEECH.** 2001. p. 33-36.

XIAO, Guangxuan et al. Smoothquant: Accurate and efficient post-training quantization for large language models. In: **International Conference on Machine Learning**. PMLR, 2023. p. 38087-38099.

ANEXO A – ANEXOS E APÊNDICES 1

Evento	Tempo	Comentário
Passe e rec. bola	30”	Anyways, back to the action, folks!
Passe e rec. bola	3”19”’	...just a hair above the ground, you can see the ball just kissing the turf there!
Drible e Disputa	48”	This is gonna be a great play, folks, stay tuned!
Drible e Disputa	39”’	The Argentine magician is known for his incredible ball control
Disputa e desarme	20”46”’	Remember, folks, it’s not just about the score, it’s about the passion and determination of these athletes!
Disputa e desarme	26”23”’	...but Alba Ramos is all over him like a cheap suit!
Disputa e desarme	26”23”’	Stay tuned, folks, it’s gonna be a wild ride!
Disputa e desarme	28”29”	Real Sociedad is not going to let up, though - they’re determined to take home the win today!
Substituição	27”’26	...you can bet your bottom dollar on that
Bloqueio	7”38”’	Stay tuned, folks
Bloqueio	14”9”’	The crowd goes wild
Bloqueio	15”’11	...weaved through the Barcelona defense like a hot knife through butter
Interceptação	6”30”’	Will Real Sociedad be able to recover and tie the game, or will Barcelona hold on for the win?
Interceptação	23”24”’	...and they’re not letting up anytime soon!
Interceptação	46”02”’	Will they be able to capitalize on this turn of events?
Chute para fora	25”19”	The crowd at the Reale Arena is going wild
Inicio/fim 1º e 2 T	46”04”’	and boy, have they been putting on a show!