



**UNIVERSIDADE FEDERAL DO CEARÁ**  
**CENTRO DE CIÊNCIAS**  
**DEPARTAMENTO DE COMPUTAÇÃO**  
**PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**  
**DOUTORADO EM COMPUTAÇÃO**

**LUIS GUSTAVO COUTINHO DO RÊGO**

**PERSONAREACT: PERSONALIZAÇÃO DE RESPOSTAS EM MODELOS DE  
LINGUAGEM COM BASE NO PARADIGMA REACT**

**FORTALEZA**

**2025**

LUIS GUSTAVO COUTINHO DO RÊGO

PERSONAREACT: PERSONALIZAÇÃO DE RESPOSTAS EM MODELOS DE  
LINGUAGEM COM BASE NO PARADIGMA REACT

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Computação. Área de Concentração: Ciência de Dados e Inteligência Artificial.

Orientador: Prof. Dr. José Antônio Fernandes de Macêdo.

Coorientadora: Prof<sup>a</sup>. Dr<sup>a</sup>. Ticiania Linhares Coelho da Silva.

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação  
Universidade Federal do Ceará  
Sistema de Bibliotecas  
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

---

R267p Rêgo, Luis Gustavo Coutinho do.  
PersonaReAct : Personalização de Respostas em Modelos de Linguagem com base no Paradigma ReAct / Luis Gustavo Coutinho do Rêgo. – 2025.  
166 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação , Fortaleza, 2025.

Orientação: Prof. Dr. José Antônio Fernandes de Macêdo.

Coorientação: Profa. Dra. Ticiania Linhares Coelho da Silva.

1. Grandes Modelos de Linguagem. 2. Personalização de Sistemas. 3. Agentes Inteligentes. 4. Engenharia de Prompts. I. Título.

CDD 005

---

LUIS GUSTAVO COUTINHO DO RÊGO

PERSONAREACT: PERSONALIZAÇÃO DE RESPOSTAS EM MODELOS DE  
LINGUAGEM COM BASE NO PARADIGMA REACT

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Computação. Área de Concentração: Ciência de Dados e Inteligência Artificial.

Aprovada em: 20 de agosto de 2025

BANCA EXAMINADORA

---

Prof. Dr. José Antônio Fernandes de Macêdo (Orientador)  
Universidade Federal do Ceará (UFC)

---

Prof<sup>a</sup>. Dr<sup>a</sup>. Ticiania Linhares Coelho da Silva (Coorientadora)  
Universidade Federal do Ceará (UFC)

---

Prof. Dr. João Paulo do Vale Madeiro  
Universidade Federal do Ceará (UFC)

---

Prof. Dr. Regis Pires Magalhães  
Universidade Federal do Ceará (UFC)

---

Prof. Dr. Marco Antônio Casanova  
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

À minha esposa, Fernanda Amâncio, que nunca deixou de estar ao meu lado nem por um segundo, e sei que nunca deixará de estar.

## AGRADECIMENTOS

Acredito que nosso caráter e nossa personalidade são moldados pelos exemplos que nos acompanham no cotidiano, nas situações comuns. Meus pais, José Tarcísio e Ludmila, sempre foram para mim esses exemplos de dedicação e integridade. Ao longo da vida, me disseram muitos “nãos”, mas nunca recusaram um pedido por um livro. Se cheguei até aqui, foi porque eles se esforçaram para construir esse caminho, com renúncia e incentivo. Sendo o caçula de quatro filhos, também encontrei inspiração nos meus irmãos Rafaela, Aníbal e Eduardo, que trilharam com seriedade e empenho seus próprios caminhos de sucesso. Suas conquistas me motivaram a seguir o mesmo percurso e me ensinaram, desde cedo, que o estudo vale a pena.

Mesmo não estando mais aqui, minha tia Maria José (*in memoriam*) continua presente nas minhas lembranças como uma das pessoas que mais me queriam bem e que mais vibravam pelas minhas conquistas. Aposto que ela estaria muito satisfeita em me ver doutor.

Durante esses seis anos de doutorado, foi com minha esposa, Fernanda Amâncio, que compartilhei minhas angústias e alegrias, minhas dúvidas e conquistas. Ela foi minha força diária, o apoio silencioso nos dias difíceis e o sorriso mais bonito nas minhas pequenas vitórias. Sou eternamente grato por todo o amor, carinho e paciência com os quais me acompanhou do início ao fim. Além disso, nossos dois gatinhos, Tom e Lua, cumpriram a missão de transformar qualquer dia ruim em um momento de leveza, fosse com uma espreguiçada dramática no meio da sala ou com um ronronado na hora certa. Do jeito deles, também me ajudaram a escrever esta tese.

Ao longo dos últimos anos, ouvi muitos relatos de colegas sobre relações difíceis com seus orientadores e orientadoras. Sempre me pareceram distantes da minha realidade, pois tive o privilégio de contar com a orientação do Prof. José Antônio e da Prof<sup>ª</sup>. Ticiania Linhares. Ambos sempre foram compreensivos, motivadores e, sobretudo, verdadeiros orientadores nos momentos em que mais precisei de direção. A Prof<sup>ª</sup>. Ticiania, por já ser minha amiga antes mesmo do doutorado começar, teve um papel ainda mais especial ao longo desta jornada. Mais do que contribuir para a publicação de artigos e para a escrita da tese, ambos foram fundamentais na formação do pesquisador que estou me tornando.

Sou grato aos amigos que o curso de Computação me deu e que levarei comigo para o resto da vida. Camila, Lucas, Toni e Victor, obrigado pelas dicas, pelos incentivos, pelos momentos divertidos e pelas boas fofocas. Não importa a distância, de alguma forma, vocês sempre estiveram ao meu lado. Peres, obrigado por todas as piadas ruins, pelos momentos de

desabafo e pelas dicas culinárias. Nobre Regis, obrigado por ser um exemplo de conhecimento e por conseguir contar piadas ainda piores que as do Peres.

Antes de chegar ao tema desta tese, trabalhei em diferentes linhas de pesquisa, desde segurança pública até dados biomédicos. Em todas elas, encontrei no Insight Data Science Lab colegas dispostos a ajudar e a compartilhar experiências. Andreza, Atslands, Babi, Carlos Freire, Felipe Marcel, Florêncio, Francesco Lettich, João Castelo, Lívia, Madson e Samir Chaves: aprendi muito com cada um de vocês. Espero, sinceramente, poder retribuir um dia tudo o que me ensinaram.

Nasci e me criei na cidade de Quixeramobim, onde tenho amigos que sempre torceram por mim, mesmo de longe. João Victor, Hélio Victor, Pedro Igor e Pedro Victor, peço desculpas pelas ausências e prometo ser mais presente. Além de vocês, agradeço especialmente ao Douglas, que foi um grande influenciador em vários momentos da minha vida. Foi ele, uns vinte anos atrás, quem me apresentou o curso de Computação. Hoje, estou entregando minha tese.

Desde que ingressei no Instituto Federal do Ceará (IFCE), em 2017, encontrei um ambiente que valorizou minha trajetória e me apoiou em diferentes etapas, tanto com os afastamentos necessários quanto com o incentivo constante à realização deste doutorado. Tive o privilégio de trabalhar ao lado de profissionais de excelência em suas áreas, tanto no campus Jaguaribe quanto no campus Quixadá. Agradeço a todos que, de alguma forma, contribuíram com minha formação e com o desenvolvimento deste trabalho. Em especial, registro minha gratidão ao colega e grande amigo Prof. Herleson Pontes, cuja postura profissional e humana é uma inspiração para mim.

Agradeço aos professores João Paulo Madeiro e Marco Antônio Casanova, membros da banca examinadora, pelo tempo dedicado, pelas valiosas contribuições e pelas sugestões que enriqueceram este trabalho.

O desenvolvimento desta pesquisa contou com o apoio do Fortaleza Esporte Clube, responsável pela cessão dos dados utilizados na tese, e da Samsung Eletrônica da Amazônia, que, por meio da Fundação CETREDE, financiou parte do período de doutorado.

Esta tese carrega um pouco de cada pessoa citada neste texto de agradecimentos e também de tantas outras que, mesmo não mencionadas, foram igualmente importantes. A todos, minha gratidão.

"We offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence."

(Noam Shazeer, 2020, p. 3.)

## RESUMO

Modelos de linguagem de grande porte (LLMs) têm se consolidado como ferramentas poderosas para tarefas de geração e recuperação de informações em linguagem natural. No entanto, a maioria dessas abordagens ainda carece de mecanismos robustos de personalização, o que limita sua aplicabilidade em contextos onde o perfil do usuário influencia diretamente a utilidade da resposta. Este trabalho propõe o PersonaReAct, uma especialização do paradigma ReAct que incorpora uma etapa final de adaptação baseada no perfil funcional do usuário. A proposta mantém o ciclo original de raciocínio e ação, introduzindo uma função de personalização (*Tailor*) responsável por ajustar a resposta final conforme as características e necessidades do usuário. O modelo foi avaliado por meio de dois estudos de caso em domínios distintos: o esportivo, com dados reais dos jogadores do Fortaleza E.C., e o médico, com dados de internação de pacientes com COVID-19. Em ambos os contextos, foram consideradas perguntas associadas a diferentes perfis e as respostas foram analisadas com base em critérios de linguagem, interpretação e coerência. Foram testados diversos modelos de linguagem com e sem raciocínio, como o GPT-5 mini e o GPT-4.1 mini. Os resultados confirmam que a abordagem proposta gera respostas mais adequadas, claras e contextualizadas quando comparada ao ReAct tradicional e com *prompt* enriquecido. Concluiu-se ainda que o PersonaReAct apresenta melhor desempenho em modelos sem raciocínio e também em modelos menores com raciocínio, como o GPT-5 nano. Além disso, observou-se que a separação modular da personalização contribui para maior controle e interpretabilidade do sistema. O trabalho também discute limitações e propõe caminhos futuros, como a transformação da etapa de personalização em um ciclo iterativo refinável e a aplicação prática em sistemas conversacionais.

**Palavras-chave:** Grandes Modelos de Linguagem; Personalização de Sistemas; Agentes Inteligentes; Engenharia de Prompts.

## ABSTRACT

Large Language Models (LLMs) have established themselves as powerful tools for natural language generation and information retrieval tasks. However, most of these approaches still lack robust personalization mechanisms, which limits their applicability in contexts where the user's profile directly influences the usefulness of the response. This work proposes PersonaReAct, a specialization of the ReAct paradigm that incorporates a final adaptation step based on the user's functional profile. The proposal preserves the original reasoning and action cycle, introducing a personalization function (Tailor) responsible for adjusting the final response according to the user's characteristics and needs. The model was evaluated through two case studies in distinct domains: the sports domain, using real data from Fortaleza E.C. players, and the medical domain, using hospitalization data from COVID-19 patients. In both contexts, questions associated with different user profiles were considered, and the answers were analyzed based on language, interpretation, and coherence criteria. Several language models with and without explicit reasoning were tested, such as GPT-5 mini and GPT-4.1 mini. The results confirm that the proposed approach generates more appropriate, clear, and contextualized answers when compared to traditional ReAct and ReAct with enriched prompts. It was further concluded that PersonaReAct performs better on models without reasoning and on smaller reasoning-capable models, such as GPT-5 nano. In addition, the modular separation of the personalization step was shown to contribute to greater control and interpretability of the system. The study also discusses limitations and proposes future directions, such as transforming the personalization step into a refinable iterative cycle and exploring practical applications in conversational systems.

**Keywords:** Large Language Models; System Personalization; Intelligent Agents; Prompt Engineering.

## LISTA DE FIGURAS

Figura 1 – Níveis de autonomia em arquiteturas baseadas em Grandes Modelos de Linguagem (LLMs). . . . .	38
Figura 2 – Comparação entre os paradigmas <i>Reason Only</i> , <i>Act Only</i> e ReAct no uso de agentes. . . . .	65
Figura 3 – Componentes da arquitetura PersonaReAct com ênfase na personalização via perfil. . . . .	66
Figura 4 – Visão geral do processo metodológico da tese, da coleta à avaliação dos agentes.	70
Figura 5 – Fluxograma detalhado do funcionamento do PersonaReAct implementado com LangGraph. . . . .	74
Figura 6 – Avaliação do critério “Coerência do Estilo ao Longo do Texto” nos modelos aplicados ao domínio esportivo. . . . .	93
Figura 7 – Avaliação do critério “Interpretação e Explicação dos Dados” nos modelos aplicados ao domínio esportivo. . . . .	94
Figura 8 – Avaliação do critério “Linguagem e Estilo Apropriados” nos modelos aplicados ao domínio esportivo. . . . .	94
Figura 9 – Comparação das médias por critério entre ReAct, ReAct usando <i>Prompt</i> Enriquecido e PersonaReAct para o modelo GPT-5 <i>mini</i> no domínio esportivo.	98
Figura 10 – Comparação das médias por critério entre ReAct, ReAct usando <i>Prompt</i> Enriquecido e PersonaReAct para o modelo GPT-5 <i>nano</i> no domínio esportivo.	99
Figura 11 – Comparação das médias por critério entre ReAct, ReAct usando <i>Prompt</i> Enriquecido e PersonaReAct para o modelo GPT-4o <i>mini</i> no domínio esportivo.	100
Figura 12 – Avaliação do critério “Coerência do Estilo ao Longo do Texto” nos modelos aplicados ao domínio médico. . . . .	103
Figura 13 – Avaliação do critério “Interpretação e Explicação dos Dados” nos modelos aplicados ao domínio médico. . . . .	104
Figura 14 – Avaliação do critério “Linguagem e Estilo Apropriados” nos modelos aplicados ao domínio médico. . . . .	104
Figura 15 – Comparação das médias por critério entre ReAct, ReAct usando <i>Prompt</i> Enriquecido e PersonaReAct para o modelo GPT-5 <i>mini</i> no domínio médico.	108
Figura 16 – Comparação das médias por critério entre ReAct, ReAct usando <i>Prompt</i> Enriquecido e PersonaReAct para o modelo GPT-5 <i>nano</i> no domínio médico.	109

Figura 17 – Comparação das médias por critério entre ReAct, ReAct usando <i>Prompt</i> Enriquecido e PersonaReAct para o modelo GPT-4o mini no domínio médico.	110
Figura 18 – Gráfico de barras das métricas de similaridade textual por pergunta para o modelo o3 mini no contexto do conjunto de dados do domínio esportivo.	116
Figura 19 – Gráfico de barras das métricas de similaridade textual por pergunta para o modelo gpt-4.1-mini no contexto do conjunto de dados do domínio esportivo.	116
Figura 20 – Gráfico de barras das métricas de similaridade textual por perfil para o modelo o3 mini no contexto do conjunto de dados do domínio esportivo.	118
Figura 21 – Gráfico de barras das métricas de similaridade textual por perfil para o modelo GPT-4.1 mini no contexto do conjunto de dados do domínio esportivo.	118
Figura 22 – Distribuição Geral das Métricas (ROUGE-L, BLEU, Distância de Edição) para os modelos o3 mini e GPT-4.1 mini no contexto do conjunto de dados do domínio esportivo.	121
Figura 23 – Clusterização das respostas com destaque para pares de respostas mais próximas (azul) e mais distantes (vermelho) em cada <i>cluster</i> (modelo GTP-4.1 mini) no domínio esportivo.	122
Figura 24 – Clusterização das respostas com destaque para pares de respostas mais próximas (azul) e mais distantes (vermelho) em cada <i>cluster</i> (modelo o3 mini) no domínio esportivo.	123
Figura 25 – Projeção das respostas com destaque para pares de respostas mais próximas (azul) e mais distantes (vermelho) (modelo GTP-4.1 mini) no domínio médico.	128
Figura 26 – Projeção das respostas com destaque para pares de respostas mais próximas (azul) e mais distantes (vermelho) (modelo o3 mini) no domínio médico.	129
Figura 27 – Gráfico de barras das métricas de similaridade textual por pergunta para o modelo o3 mini no contexto do conjunto de dados do domínio médico.	163
Figura 28 – Gráfico de barras das métricas de similaridade textual por pergunta para o modelo gpt-4.1-mini no contexto do conjunto de dados do domínio médico.	164
Figura 29 – Gráfico de barras das métricas de similaridade textual por perfil para o modelo o3 mini no contexto do conjunto de dados do domínio médico.	164
Figura 30 – Gráfico de barras das métricas de similaridade textual por perfil para o modelo GPT-4.1 mini no contexto do conjunto de dados do domínio médico.	165

Figura 31 – Distribuição Geral das Métricas (ROUGE-L, BLEU, Distância de Edição) para os modelos o3 mini e GPT-4.1 mini no contexto do conjunto de dados do domínio médico. . . . .	165
---	-----

## LISTA DE TABELAS

Tabela 1 – Médias e desvios padrão por modelo — ReAct (com raciocínio) - Domínio Esportivo. . . . .	90
Tabela 2 – Médias e desvios padrão por modelo — ReAct usando <i>Prompt</i> Enriquecido (com raciocínio) - Domínio Esportivo. . . . .	91
Tabela 3 – Médias e desvios padrão por modelo — PersonaReAct (com raciocínio) - Domínio Esportivo. . . . .	91
Tabela 4 – Médias e desvios padrão por modelo — ReAct (sem raciocínio) - Domínio Esportivo. . . . .	91
Tabela 5 – Médias e desvios padrão por modelo — ReAct usando <i>Prompt</i> Enriquecido (sem raciocínio) - Domínio Esportivo. . . . .	92
Tabela 6 – Médias e desvios padrão por modelo — PersonaReAct (sem raciocínio) - Domínio Esportivo. . . . .	92
Tabela 7 – Médias e desvios padrão por modelo — ReAct (com raciocínio) - Domínio Médico. . . . .	97
Tabela 8 – Médias e desvios padrão por modelo — ReAct usando <i>Prompt</i> Enriquecido (com raciocínio) - Domínio Médico. . . . .	97
Tabela 9 – Médias e desvios padrão por modelo — PersonaReAct (com raciocínio) - Domínio Médico. . . . .	98
Tabela 10 – Médias e desvios padrão por modelo — ReAct (sem raciocínio) - Domínio Médico. . . . .	101
Tabela 11 – Médias e desvios padrão por modelo — ReAct usando <i>Prompt</i> Enriquecido (sem raciocínio) - Domínio Médico. . . . .	101
Tabela 12 – Médias e desvios padrão por modelo — PersonaReAct (sem raciocínio) - Domínio Médico. . . . .	102
Tabela 13 – Médias e desvios padrão das métricas de similaridade textual por pergunta para o modelo o3 mini no contexto do conjunto de dados do domínio de futebol. . . . .	114
Tabela 14 – Médias e desvios padrão das métricas de similaridade textual por pergunta para o modelo gpt-4.1-mini no contexto do conjunto de dados do domínio de futebol. . . . .	115

Tabela 15 – Médias e desvios padrão das métricas de similaridade textual por perfil para o modelo o3 mini no contexto do conjunto de dados do domínio de futebol. . . . .	117
Tabela 16 – Médias e desvios padrão das métricas de similaridade textual por perfil para o modelo GPT-4.1 mini no contexto do conjunto de dados do domínio de futebol. . . . .	117
Tabela 17 – Distribuição geral das métricas de similaridade textual entre respostas personalizadas e padrão para o modelo o3 mini no domínio esportivo. . . . .	120
Tabela 18 – Distribuição geral das métricas de similaridade textual entre respostas personalizadas e padrão (modelo GPT-4.1 mini) no domínio esportivo. . . . .	120

## LISTA DE QUADROS

Quadro 1 – Rastreabilidade entre Objetivos, Questões de Pesquisa e Hipóteses . . . . .	26
Quadro 2 – Comparativo entre trabalhos recentes segundo características estruturais e de personalização. . . . .	57
Quadro 3 – Comparação entre ReAct e PersonaReAct . . . . .	68
Quadro 4 – Visão geral dos experimentos. . . . .	80
Quadro 5 – Exemplo de pergunta feita para o perfil “Torcedor do Time” com uma resposta padrão e uma resposta personalizada com o PersonaReAct. . . . .	83
Quadro 6 – Respostas personalizadas à pergunta. . . . .	86
Quadro 7 – Rastreabilidade entre Hipóteses, Experimentos e Contribuições . . . . .	87
Quadro 8 – Notas e justificativas do <i>framework</i> G-Eval para o domínio esportivo. Abordagem: ReAct. . . . .	95
Quadro 9 – Notas e justificativas do <i>framework</i> G-Eval para o domínio esportivo. Abordagem: ReAct usando <i>Prompt</i> Enriquecido. . . . .	95
Quadro 10 – Notas e justificativas do <i>framework</i> G-Eval para o domínio esportivo. Abordagem: PersonaReAct. . . . .	96
Quadro 11 – Notas e justificativas do <i>framework</i> G-Eval para o domínio médico. Abordagem: ReAct. . . . .	105
Quadro 12 – Notas e justificativas do <i>framework</i> G-Eval para o domínio médico. Abordagem: ReAct usando <i>Prompt</i> Enriquecido. . . . .	105
Quadro 13 – Notas e justificativas do <i>framework</i> G-Eval para o domínio médico. Abordagem: PersonaReAct. . . . .	106
Quadro 14 – Identificação do atleta e sessão. . . . .	147
Quadro 15 – Dados temporais da sessão. . . . .	147
Quadro 16 – Frequência cardíaca (HR). . . . .	147
Quadro 17 – Distância e velocidade. . . . .	148
Quadro 18 – Acelerações e desacelerações. . . . .	148
Quadro 19 – Carga fisiológica e esforço. . . . .	148
Quadro 20 – Variabilidade da frequência cardíaca (HRV) e intervalos RR. . . . .	148
Quadro 21 – Informações sobre os pacientes. . . . .	148
Quadro 22 – Dados dos exames coletados. . . . .	149
Quadro 23 – Informações gerais sobre os desfechos dos pacientes. . . . .	149

Quadro 24 – Variações das perguntas por perfil de usuário. . . . .	158
Quadro 25 – Respostas geradas pelo PersonaReAct para diferentes perfis a partir da mesma pergunta sobre sprints no jogo contra o Ceará. . . . .	159
Quadro 26 – Respostas geradas pelo PersonaReAct para diferentes perfis sobre o tempo de sprint de Bruno Pacheco no jogo contra o Ceará. . . . .	160
Quadro 27 – Respostas geradas pelo PersonaReAct para diferentes perfis sobre o maior coeficiente de variação da velocidade no jogo contra o Ceará. . . . .	161
Quadro 28 – Respostas geradas pelo PersonaReAct para diferentes perfis sobre as acelerações e desacelerações de Bruno Pacheco no jogo contra o Ceará. . . . .	162

## LISTA DE CÓDIGOS-FONTE

Código-fonte 1 – Critério “Linguagem e Estilo Apropriados”. . . . .	151
Código-fonte 2 – Critério “Interpretação e Explicação dos Dados”. . . . .	152
Código-fonte 3 – Critério “Coerência do Estilo ao Longo do Texto”. . . . .	153
Código-fonte 4 – Prompt do Agente de Personalização. . . . .	154

## LISTA DE ABREVIATURAS E SIGLAS

API	Interface de Programação de Aplicações, do inglês <i>Application Programming Interface</i>
GPT	Transformador Pré-treinado Generativo, do inglês <i>Generative Pre-trained Transformer</i>
GPU	Unidade de Processamento Gráfico, do inglês <i>Graphics Processing Unit</i>
IA	Inteligência Artificial
LLaMA	Grande Modelo de Linguagem da Meta AI, do inglês <i>Large Language Model Meta AI</i>
LLM	Grande Modelo de Linguagem, do inglês <i>Large Language Model</i>
LN	Linguagem Natural
LSTM	Memória de Longo Curto Prazo, do inglês <i>Long Short-Term Memory</i>
MoE	Mistura de Especialistas, do inglês <i>Mixture of Experts</i>
NLP	Processamento de Linguagem Natural, do inglês <i>Natural Language Processing</i>
PaLM	Modelo de Linguagem Pathways, do inglês <i>Pathways Language Model</i>
RAG	Geração Aumentada por Recuperação, do inglês <i>Retrieval-Augmented Generation</i>
RLHF	Aprendizado por Reforço com Feedback Humano, do inglês <i>Reinforcement Learning with Human Feedback</i>
RNN	Rede Neural Recorrente, do inglês <i>Recurrent Neural Network</i>
SQL	Linguagem de Consulta Estruturada, do inglês <i>Structured Query Language</i>
TPU	Unidade de Processamento Tensorial, do inglês <i>Tensor Processing Unit</i>

## LISTA DE SÍMBOLOS

$a_t$	Ação externa executada por um agente no instante $t$
$A$	Espaço de ações externas que produzem efeitos no ambiente
$\hat{a}_t$	Ação geral executada no instante $t$ , podendo ser externa ou interna
$\hat{A}$	Espaço total de ações, unindo ações externas ( $A$ ) e internas ( $L$ )
$c_t$	Contexto acumulado por um agente no instante $t$
$C$	Espaço de todos os contextos possíveis
$l_t$	Traço de raciocínio (ação interna) emitido por um agente no instante $t$
$L$	Espaço de linguagem contendo traços de raciocínio emitidos por um agente
$M$	Espaço das mensagens finais personalizadas
$o_t$	Observação recebida por um agente no instante de tempo $t$
$O$	Espaço de observações possíveis que um agente pode receber do ambiente
$O^*$	Conjunto (ou sequência) de observações acumuladas ao longo do ciclo de execução
$p$	Perfil do usuário em interação com um agente
$P$	Espaço dos perfis de usuários
$t$	Instante de tempo
$\pi$	Política de decisão de um agente
$\tau$	Função de personalização ( <i>Tailor</i> ) que adapta a resposta conforme o perfil

## SUMÁRIO

1	<b>INTRODUÇÃO</b>	21
1.1	<b>Justificativa</b>	22
1.2	<b>Objetivos</b>	23
1.3	<b>Questões de Pesquisa</b>	24
1.4	<b>Hipóteses</b>	25
1.5	<b>Contribuições do Trabalho</b>	26
1.6	<b>Organização da Tese</b>	28
2	<b>FUNDAMENTAÇÃO TEÓRICA</b>	29
2.1	<b>Grandes Modelos de Linguagem</b>	29
2.1.1	<i>Evolução e Arquitetura dos LLMs</i>	29
2.2	<b>Engenharia de Prompts</b>	31
2.2.1	<i>Prompts e Prompt Engineering</i>	31
2.2.2	<i>Paradigmas e Estratégias de Engenharia de Prompts</i>	33
2.2.3	<i>Personalização via Prompts</i>	35
2.3	<b>Agentes Baseados em Modelos de Linguagem</b>	37
2.3.1	<i>Fundamentos de Agentes com LLMs</i>	37
2.3.2	<i>Arquiteturas Multiagente</i>	39
2.4	<b>Ecosistema Tecnológico para Construção de Sistemas com LLMs</b>	41
2.4.1	<i>Principais Modelos de Linguagem</i>	41
2.4.2	<i>Frameworks e Bibliotecas para Agentes com LLMs</i>	43
2.4.3	<i>Recuperação de Informações Estruturadas</i>	44
3	<b>TRABALHOS RELACIONADOS</b>	47
3.1	<i>Prompts Contextuais</i>	48
3.2	<b>Prompts Baseados em Personas</b>	49
3.3	<b>Prompts com Enriquecimento por Perfil</b>	51
3.4	<b>Refinamento de Prompts</b>	53
3.5	<b>Análise Comparativa dos Trabalhos</b>	55
4	<b>PERSONAREACT: INTEGRAÇÃO DE RACIOCÍNIO, AÇÃO E PERSONALIZAÇÃO</b>	58
4.1	<b>Contextualização e Justificativa do Modelo</b>	58

4.2	Modelo Conceitual e Formalização . . . . .	62
4.3	Arquitetura e Funcionamento Interno . . . . .	64
4.4	Avanços em Relação ao Paradigma ReAct . . . . .	66
5	<b>METODOLOGIA . . . . .</b>	69
5.1	Coleta e Preparação dos Dados . . . . .	69
5.2	Arquitetura e Implementação dos padrões ReAct e PersonaReAct . . . . .	73
5.3	Avaliação de Respostas: <i>LLMs-as-a-Judge</i> e Métricas Tradicionais . . . . .	76
5.4	Construção do <i>Prompt</i> de Personalização . . . . .	78
5.5	Desenho Experimental . . . . .	80
5.6	Limitações e Considerações Metodológicas . . . . .	88
6	<b>RESULTADOS E DISCUSSÃO . . . . .</b>	90
6.1	Experimento 1 - Eficácia da Personalização . . . . .	90
6.2	Experimento 2 - Variação entre Perfis . . . . .	113
7	<b>CONCLUSÃO . . . . .</b>	133
	<b>REFERÊNCIAS . . . . .</b>	136
	<b>APÊNDICES . . . . .</b>	146
	<b>APÊNDICE A – CONJUNTO DE DADOS INICIAL . . . . .</b>	147
	<b>APÊNDICE B – CRITÉRIOS TEXTUAIS UTILIZADOS NA AVALIAÇÃO POR MODELOS DE LINGUAGEM . . . . .</b>	150
	<b>APÊNDICE C – PROMPTS DOS AGENTES . . . . .</b>	154
	<b>APÊNDICE D – EXEMPLOS DE PERGUNTAS E RESPOSTAS . . . . .</b>	157
	<b>APÊNDICE E – RESULTADOS DA ANÁLISE DE SIMILARIDADE TEXTUAL PARA ESTUDO DE CASO 2 . . . . .</b>	163

## 1 INTRODUÇÃO

A Inteligência Artificial (IA), especialmente por meio dos avanços em Processamento de Linguagem Natural (NLP), tem promovido transformações significativas em diversos setores da sociedade, ciência e tecnologia (Zhao *et al.*, 2025; Matarazzo; Torlone, 2025). LLMs, como o modelo GPT-5 da OpenAI e outros similares, são capazes de processar e gerar textos coerentes e contextualmente relevantes a partir de grandes volumes de dados textuais (Ziems *et al.*, 2024; Raza *et al.*, 2025; Minaee *et al.*, 2025). Esses modelos já se consolidam como ferramentas centrais em aplicações práticas como assistentes virtuais, sistemas de recomendação, atendimento automatizado, análise de dados e ambientes educacionais (Shao *et al.*, 2024).

Uma limitação recorrente desses modelos diz respeito à sua baixa sensibilidade ao perfil, contexto ou papel funcional do interlocutor (Wozniak *et al.*, 2024; Zhang *et al.*, 2025). Ainda que ofereçam respostas coerentes com a pergunta formulada, tendem a produzir conteúdos genéricos, muitas vezes descolados das necessidades específicas de quem realiza a consulta. Por exemplo, ao perguntar “*Qual foi o desempenho do jogador Rafael no último jogo?*”, um torcedor pode esperar destaques emocionantes e estatísticas sobre gols ou dribles, enquanto um fisioterapeuta busca informações mais técnicas, como carga de esforço, tempo de recuperação ou número de *sprints*. Ao invés disso, modelos sem personalização fornecem respostas em um tom padrão, como “*O jogador Rafael teve um desempenho acima da média.*”. Essa ausência de personalização fragiliza o impacto da informação, especialmente em domínios especializados que exigem respostas ajustadas a diferentes objetivos, níveis de conhecimento ou funções institucionais.

A personalização das respostas geradas por LLMs emerge, portanto, como uma demanda estratégica (Mok *et al.*, 2025). Não se trata apenas de tornar a comunicação mais natural, mas de garantir que a informação transmitida seja verdadeiramente útil para quem a recebe. Sistemas interativos centrados no usuário precisam reconhecer que diferentes perfis (como técnicos, analistas, educadores, pacientes ou gestores) acessam os mesmos dados com intenções distintas. A mesma pergunta pode representar diferentes objetivos, exigindo variações no conteúdo, na profundidade e na linguagem da resposta.

Embora existam avanços em técnicas de personalização baseadas em histórico de uso, preferências anteriores ou padrões estatísticos, esses métodos nem sempre capturam adequadamente o papel desempenhado pelo usuário no momento da interação (Wozniak *et al.*, 2024). A ausência de uma estrutura capaz de ajustar sistematicamente a resposta com base no

perfil funcional limita a aplicabilidade dos modelos em contextos profissionais, educacionais e institucionais. Além disso, a eficácia de diferentes estratégias de personalização (como o enriquecimento direto do *prompt* ou a adoção de módulos específicos) pode variar de acordo com a arquitetura e as capacidades do LLM, especialmente em modelos com mecanismos avançados de raciocínio.

Diante desse cenário, esta tese propõe o PersonaReAct, um padrão de interação que estende o paradigma ReAct (Yao *et al.*, 2023) ao incorporar personalização explícita com base em perfis funcionais de usuário. O ReAct é uma abordagem para agentes baseados em modelos de linguagem que combina, de forma intercalada, etapas de raciocínio (*reasoning*) e de execução de ações (*acting*), permitindo que o agente justifique suas decisões e consulte ferramentas externas antes de formular a resposta final. O modelo aqui proposto mantém esses princípios, mas os amplia ao considerar o contexto funcional do interlocutor como variável determinante na formatação do conteúdo gerado. Dessa forma, uma mesma pergunta pode dar origem a respostas distintas, adaptadas às necessidades, linguagem e prioridades de diferentes perfis de usuários.

A proposta foi validada por meio de dois estudos de caso: um no domínio esportivo, com foco no futebol profissional, e outro no domínio médico, com foco em casos de COVID-19. Utilizando dados reais de desempenho de atletas e contextos de jogo e dados clínicos e laboratoriais de pacientes, os experimentos demonstram como o PersonaReAct pode produzir respostas mais alinhadas às expectativas de técnicos, fisioterapeutas, médicos ou pacientes. Essas aplicações práticas permitem avaliar não apenas a eficácia técnica do modelo, mas também sua relevância comunicacional e eficácia relativa frente a abordagens mais simples de personalização. Os resultados indicam que o PersonaReAct apresenta ganhos mais evidentes em modelos sem capacidades avançadas de raciocínio, enquanto, em modelos com esse recurso, seu desempenho se assemelha ao obtido com estratégias de enriquecimento direto do *prompt*.

## 1.1 Justificativa

O uso de IA e, em especial, de LLMs, tem crescido rapidamente em áreas como saúde, educação, esportes e gestão pública. Com isso, surge uma necessidade que vai além de obter respostas corretas: é preciso que essas respostas sejam úteis para quem pergunta, levando em conta o seu papel e o contexto da interação. Essa personalização orientada pelo perfil do usuário ainda é pouco explorada e, muitas vezes, negligenciada nas soluções atuais (Wozniak *et*

*al.*, 2024; Zhang *et al.*, 2025).

A maioria das abordagens existentes de personalização foca em históricos de uso ou preferências passadas. Embora úteis, essas estratégias não captam bem a situação do momento nem a função específica de quem interage com o sistema. Em cenários especializados, como um diagnóstico médico, uma análise de desempenho esportivo ou uma recomendação técnica, essa limitação pode fazer a diferença entre uma resposta realmente relevante e uma informação genérica, sem aplicabilidade prática.

O PersonaReAct foi pensado para preencher essa lacuna. Ele insere o perfil funcional do usuário diretamente no fluxo de geração de respostas, aproveitando a estrutura do paradigma ReAct, mas ampliando sua capacidade de adaptação. Isso significa que a mesma pergunta pode resultar em respostas diferentes, mais adequadas ao que cada perfil realmente precisa.

Embora esta tese valide a proposta em dois estudos de caso (um no futebol profissional e outro no contexto médico da COVID-19) o modelo pode ser aplicado a qualquer domínio onde o perfil de quem pergunta influencia o que deve ser respondido. Essa versatilidade amplia seu potencial de uso prático e, ao mesmo tempo, oferece à comunidade científica uma nova perspectiva para pensar e comparar diferentes formas de personalização.

Assim, o trabalho se justifica tanto pela utilidade concreta da solução, que pode melhorar a comunicação em diversos setores, quanto pela contribuição acadêmica, que ajuda a entender melhor até onde modelos de linguagem podem ir quando o foco é adaptar a resposta ao usuário final.

## **1.2 Objetivos**

### ***Objetivo Geral***

Propor um novo paradigma baseado no paradigma ReAct, denominado PersonaReAct, que incorpora mecanismos explícitos de personalização em fluxos de raciocínio e ação, permitindo a adaptação dinâmica das respostas ao perfil do usuário.

### ***Objetivos Específicos***

1. Definir a arquitetura geral do PersonaReAct como uma especialização do modelo ReAct, descrevendo seus módulos de raciocínio, execução e personalização, bem como o fluxo iterativo entre eles.

2. Projetar e implementar, dentro dessa arquitetura, o módulo de personalização textual (*Tailor*), responsável por transformar observações brutas em respostas ajustadas ao conteúdo, estilo e nível de detalhamento esperados por diferentes perfis de usuário (como técnico, fisioterapeuta e torcedor).
3. Avaliar a eficácia do PersonaReAct em dois estudos de caso: um aplicado ao domínio esportivo (utilizando dados reais de jogadores e partidas de futebol) e outro aplicado ao domínio médico (utilizando dados clínicos e laboratoriais de pacientes).
4. Comparar o desempenho do PersonaReAct com o ReAct original, avaliando os efeitos da personalização modular na consistência, alinhamento e personalização das respostas geradas.
5. Investigar o impacto do uso de diferentes modelos de linguagem no desempenho do PersonaReAct e do ReAct original, analisando como variações na arquitetura e nas capacidades de raciocínio influenciam a efetividade da personalização.

### 1.3 Questões de Pesquisa

A pesquisa proposta busca responder às seguintes questões:

1. Como a inclusão de um estágio de personalização (*Tailor*) em um agente baseado no paradigma ReAct afeta a consistência, alinhamento e personalização das respostas para diferentes perfis de usuários?  
*Esta questão investiga se a adição de uma etapa dedicada à personalização, capaz de adaptar o conteúdo gerado ao perfil do usuário, contribui efetivamente para tornar as respostas mais personalizadas de forma consistente ao longo do texto e alinhadas às necessidades comunicacionais dos diferentes tipos de usuários.*
2. A personalização com base no papel do usuário modifica significativamente o conteúdo, o estilo e o nível de detalhamento das respostas, mesmo para perguntas idênticas?  
*Busca-se verificar se o mesmo prompt inicial pode resultar em respostas distintas a depender do perfil especificado, evidenciando que o sistema é sensível não apenas à pergunta, mas também ao papel do usuário na adaptação da resposta.*
3. O PersonaReAct oferece ganhos significativos em relação ao uso direto do ReAct usando um *prompt* personalizado que incorpore as informações de perfil? Ou seja, seria possível obter resultados semelhantes sem a separação explícita do componente de personalização?  
*Avalia-se se o uso de uma arquitetura modular, que separa a personalização em uma etapa*

*específica, resulta em respostas superiores às obtidas por abordagens que incorporam informações de perfil diretamente nos prompts do ReAct tradicional.*

4. O desempenho relativo do PersonaReAct em comparação a outras estratégias de personalização varia de acordo com o tipo de modelo de linguagem utilizado, especialmente entre modelos com e sem capacidades avançadas de raciocínio?

*Busca-se investigar se a arquitetura modular do PersonaReAct apresenta ganhos consistentes em diferentes tipos de modelos, ou se sua vantagem é mais pronunciada em modelos sem recursos robustos de raciocínio, enquanto modelos com raciocínio integrado se beneficiam de forma semelhante com abordagens mais simples, como o enriquecimento direto do prompt.*

#### **1.4 Hipóteses**

1. A personalização das respostas com base no perfil do usuário aumenta a consistência, alinhamento e personalização das respostas, tornando-as mais adequadas para diferentes tipos de usuários finais.

*Parte-se da hipótese de que perfis distintos possuem expectativas comunicacionais diferentes e que adaptar a linguagem e o conteúdo da resposta a essas expectativas torna a interação mais eficiente e compreensível.*

2. A personalização baseada em perfis gera modificações substanciais no conteúdo, na linguagem e no foco das respostas, mesmo quando a pergunta de entrada é idêntica.

*Espera-se que o perfil do usuário atue como um parâmetro determinante na construção da resposta final, levando a transformações estruturais e semânticas que diferenciem respostas originadas da mesma consulta.*

3. A separação explícita do componente de personalização no PersonaReAct pode influenciar a qualidade e a organização das respostas em relação a abordagens que utilizam apenas o ReAct com um *prompt* enriquecido.

*Espera-se que a modularização permita maior controle sobre a etapa de personalização e possibilite ajustes direcionados sem interferir no processo de raciocínio e ação, ainda que a magnitude desse impacto possa variar conforme o modelo utilizado.*

4. O tipo de modelo de linguagem utilizado (com ou sem capacidades avançadas de raciocínio) influencia o desempenho relativo das diferentes abordagens de personalização avaliadas.

*Parte-se da hipótese de que a arquitetura e as capacidades internas do modelo afetam a*

*forma como instruções de personalização são processadas e integradas ao conteúdo final gerado.*

O Quadro 1 apresenta um mapeamento entre os objetivos específicos da pesquisa, as questões de pesquisa investigadas e as hipóteses formuladas. Essa estrutura permite visualizar claramente como cada etapa do trabalho está relacionada a aspectos investigativos e testáveis, garantindo coerência entre o desenvolvimento do sistema, sua avaliação e os critérios científicos adotados.

Quadro 1 – Rastreabilidade entre Objetivos, Questões de Pesquisa e Hipóteses

<b>Objetivo Específico</b>	<b>Questões de Pesquisa Associadas</b>	<b>Hipóteses Relacionadas</b>
OE1 – Definir a arquitetura geral do PersonaReAct como uma especialização do modelo ReAct	QP3 – O PersonaReAct oferece ganhos em relação ao ReAct com <i>prompt</i> personalizado?	H3 – A separação da personalização pode influenciar a qualidade e organização das respostas em relação ao ReAct com <i>prompt</i> enriquecido
OE2 – Projetar e implementar o módulo de personalização textual ( <i>Tailor</i> )	QP1 – A personalização melhora qualidade e utilidade? QP2 – A resposta muda conforme o perfil?	H1 – Personalização aumenta utilidade e clareza H2 – Perfis diferentes geram respostas diferentes
OE3 – Avaliar a eficácia do PersonaReAct no domínio esportivo com dados reais	QP1, QP2	H1, H2
OE4 – Comparar o desempenho do PersonaReAct com o ReAct original	QP3 – O PersonaReAct oferece ganhos em relação ao ReAct com <i>prompt</i> personalizado?	H3
OE5 – Investigar o impacto do uso de diferentes modelos de linguagem no desempenho do PersonaReAct e do ReAct original	QP4 – O tipo de modelo (com ou sem <i>reasoning</i> ) influencia o desempenho relativo das abordagens?	H4 – O tipo de modelo influencia o desempenho relativo das diferentes abordagens de personalização

Fonte: elaborado pelo autor.

## 1.5 Contribuições do Trabalho

Nesta tese, apresentamos as seguintes contribuições:

- 1. Proposição do PersonaReAct:** Foi definido formalmente o PersonaReAct, um paradigma de interação com modelos de linguagem que estende o ReAct ao incorporar uma etapa adicional de personalização textual (*Tailor*) com base no perfil do usuário. O modelo permite que agentes raciocinem sobre a tarefa (*Reason*), atuem sobre o ambiente (*Act*) e adaptem a resposta textual (*Tailor*) conforme características comunicacionais e cognitivas específicas do interlocutor.
- 2. Implementação de um *framework* modular baseado no PersonaReAct:** Implementou-se um sistema funcional que materializa o PersonaReAct como um fluxo multiagente

estruturado com base no LangGraph (LangChain Team, 2024), no qual distintos nós do grafo representam etapas de raciocínio, tomada de decisão, execução de ações e personalização, operando de forma dinâmica e condicional.

3. **Projeto e validação de um mecanismo de personalização textual (*Tailor*):** Foi proposto e testado um agente especializado na adaptação das respostas finais geradas pelo sistema com base em diferentes perfis de usuários. Esse componente considera o histórico da interação e aplica técnicas de filtragem, ajuste de tom e seleção de conteúdo, resultando em respostas comunicacionalmente mais adequadas a públicos diversos, como técnicos, fisioterapeutas e torcedores de um time de futebol.
4. **Avaliação do PersonaReAct no domínio esportivo e médico:** Foram conduzidos dois estudos de caso baseados em dados reais de partidas de futebol profissional, incluindo trajetórias dos jogadores e métricas fisiológicas, e em dados reais de pacientes hospitalares, incluindo dados clínicos e laboratoriais, com o objetivo de avaliar a eficácia do PersonaReAct na produção de respostas personalizadas e embasadas.
5. **Comparação sistemática com o ReAct:** Comparou-se o desempenho do PersonaReAct com o ReAct original, evidenciando os ganhos qualitativos e estruturais decorrentes da introdução da personalização modular.
6. **Análise do impacto do tipo de modelo na personalização:** Foi avaliado o desempenho do PersonaReAct e do ReAct com *prompt* enriquecido em diferentes modelos de linguagem, incluindo variantes com e sem capacidades avançadas de raciocínio, investigando como a arquitetura e os recursos internos do modelo influenciam a eficácia das abordagens de personalização.
7. **Criação de um corpus estruturado de dados esportivos:** Uma base de dados relacional foi construída, com informações espaciais e fisiológicas reais de jogadores de futebol, utilizada como fonte para raciocínio, consultas e personalização das respostas.
8. **Definição e aplicação de critérios para avaliação da personalização:** Elaborou-se um conjunto de três critérios específicos para mensurar a qualidade da personalização em respostas geradas por modelos de linguagem: *Linguagem e Estilo Apropriados*, *Interpretação e Explicação dos Dados* e *Coerência do Estilo ao Longo do Texto*.
9. **Disponibilização pública do código-fonte:** Foi tornado público o código-fonte do sistema PersonaReAct, com o objetivo de garantir a reprodutibilidade dos experimentos e fomentar o avanço de pesquisas futuras na área.

## **1.6 Organização da Tese**

O presente trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta a Fundamentação Teórica, abordando os principais conceitos que fundamentam o trabalho. O Capítulo 3 discute os Trabalhos Relacionados, evidenciando as lacunas existentes na literatura e situando a proposta do PersonaReAct. O Capítulo 4 apresenta a formalização do PersonaReAct. No Capítulo 5, é detalhada a Metodologia utilizada, incluindo as etapas de implementação e testes do sistema proposto. O Capítulo 6 traz os Resultados e a Discussão, com análises das validações realizadas. Por fim, o Capítulo 7 apresenta as Conclusões e aponta direções para Trabalhos Futuros, destacando contribuições e possibilidades de continuidade da pesquisa.

## 2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos conceituais e técnicos necessários para a compreensão e o desenvolvimento do PersonaReAct. A proposta integra múltiplas frentes de pesquisa recentes em NLP, com destaque para LLMs, Engenharia de *Prompts*, arquiteturas baseadas em agentes e ecossistemas de ferramentas voltadas à construção de sistemas inteligentes.

A Seção 2.1 introduz os LLMs, abordando seu funcionamento geral. Em seguida, a Seção 2.2 explora o campo da Engenharia de *Prompts*, com ênfase em estratégias de controle, personalização e adaptação do comportamento de modelos por meio de instruções formuladas.

A Seção 2.3 discute o uso de LLMs como componentes centrais em arquiteturas multiagentes, destacando abordagens orientadas ao raciocínio e à execução de ações (como o paradigma ReAct). Por fim, a Seção 2.4 apresenta o conjunto de ferramentas utilizadas para implementação do sistema proposto, incluindo *frameworks* como LangChain e LangGraph, bem como estratégias de recuperação estruturada de informações, como consultas via *Text-to-SQL*, que se alinham ao conceito de Geração Aumentada por Recuperação (RAG).

### 2.1 Grandes Modelos de Linguagem

#### 2.1.1 Evolução e Arquitetura dos LLMs

Os LLMs representam uma evolução significativa em relação aos modelos estatísticos e neurais clássicos de NLP (Zhao *et al.*, 2025). Essa transformação foi possibilitada principalmente pela adoção da arquitetura Transformer, por avanços na disponibilidade computacional e pelo uso de conjuntos de dados massivos para pré-treinamento. Nesta seção, discutimos a origem e os fundamentos arquiteturais dos LLMs, enfatizando o papel do Transformer e os desafios e estratégias envolvidos na sua escalabilidade e treinamento em larga escala.

A arquitetura Transformer (Vaswani *et al.*, 2017) revolucionou o campo ao eliminar o uso de mecanismos recorrentes ou convolucionais. Em vez disso, os Transformers utilizam mecanismos de autoatenção, que permitem ao modelo ponderar diferentes partes da entrada com base em sua relevância contextual (característica especialmente importante para capturar dependências de longo alcance em textos).

A principal inovação do Transformer está no mecanismo de autoatenção, que calcula, para cada palavra da entrada, uma representação contextualizada em relação às demais palavras

da sequência. Essa abordagem permite paralelismo eficiente durante o treinamento e supera limitações das Redes Neurais Recorrentes (RNNs) e Memórias de Longo Curto Prazo (LSTMs), que sofriam com dificuldades de paralelização e degradação de gradientes em sequências longas. Além disso, a introdução de camadas residuais, normalização de camada e codificações posicionais aumentou a estabilidade e a capacidade de generalização dos modelos.

A partir da introdução dos Transformers, diversos modelos foram desenvolvidos com variações no número de camadas, cabeças de atenção e parâmetros, visando à ampliação da capacidade de aprendizado dos modelos. Essa tendência ficou conhecida como *scaling laws*, ou leis de escalabilidade, que demonstram empiricamente que o desempenho de LLMs melhora sistematicamente com o aumento da quantidade de dados, do tamanho do modelo e da duração do treinamento (Kaplan *et al.*, 2020).

Modelos como BERT (Devlin *et al.*, 2019), a família de Transformadores Pré-treinados Generativos (GPTs), incluindo GPT-2 (Radford *et al.*, 2019), GPT-3 (Brown *et al.*, 2020) e GPT-4 (OpenAI *et al.*, 2024) e o Grande Modelo de Linguagem da Meta AI (LLaMA) (Touvron *et al.*, 2023) são exemplos dessa trajetória. O GPT-3, por exemplo, foi treinado com 175 bilhões de parâmetros e 300 bilhões de *tokens*, enquanto o GPT-4, embora tenha detalhes ainda não públicos, é estimado em alguns trilhões de parâmetros com capacidade multimodal.

Essa escalabilidade impõe, no entanto, desafios técnicos significativos. O treinamento de LLMs exige *clusters* de Unidades de Processamento Gráfico (GPUs) ou Unidades de Processamento Tensorial (TPUs) altamente especializados, estratégias avançadas de paralelismo de dados e modelos (como *tensor parallelism*, *pipeline parallelism* e *ZeRO optimization*) (Narayanan *et al.*, 2021), além de gestão rigorosa de memória e energia. O custo de treinamento de um modelo como o GPT-3 é estimado em milhões de dólares, limitando sua produção a empresas ou instituições com grande poder computacional.

Outro componente essencial dos LLMs é o paradigma de pré-treinamento seguido por adaptação. Em geral, esses modelos são treinados inicialmente com tarefas de previsão de palavras (como *masked language modeling*, no caso do BERT, ou *causal language modeling*, no caso da família GPT), utilizando grandes volumes de textos extraídos da internet, livros, artigos e outros domínios (Charpentier; Samuel, 2024).

Após o pré-treinamento, os modelos podem ser ajustados com diferentes técnicas:

- *Fine-tuning* supervisionado, em que o modelo é refinado com conjuntos de dados anotados para tarefas específicas;

- *Few-shot learning*, em que o modelo generaliza para novas tarefas com poucos exemplos no próprio *prompt*;
- *In-context learning*, em que a tarefa é exemplificada durante a geração, sem atualizar os pesos do modelo;
- *Aprendizado por Reforço com Feedback Humano (RLHF)*, técnica usada em modelos como ChatGPT (Ouyang *et al.*, 2022), que refina o comportamento do modelo com base em preferências humanas sobre a qualidade das respostas.

Essa última estratégia é particularmente importante para alinhar os modelos com valores humanos e evitar respostas ofensivas, enganosas ou enviesadas.

## 2.2 Engenharia de *Prompts*

### 2.2.1 *Prompts e Prompt Engineering*

A engenharia de *prompts* (*prompt engineering*) se consolidou como uma das práticas mais relevantes na utilização de LLMs, atuando como principal meio de controle e condicionamento das saídas desses modelos. Ao contrário de abordagens que exigem ajuste dos pesos do modelo, a engenharia de *prompts* opera no nível da entrada textual, utilizando linguagem natural estruturada para induzir comportamentos desejados. Antes de explorar as estratégias e paradigmas envolvidos, é fundamental estabelecer definições formais para os conceitos centrais.

**Definição 2.2.1 (Prompt)** *Um prompt é uma sequência textual de entrada fornecida a um LLM com o objetivo de condicionar a geração de texto subsequente. Formalmente, é o prefixo  $x$  de uma sequência, a partir do qual o modelo estima uma distribuição de probabilidade  $P(y|x)$  sobre as possíveis continuções  $y$ . O prompt pode incluir instruções, perguntas, exemplos, dados contextuais ou qualquer outra informação textual relevante para guiar a resposta.*

**Definição 2.2.2 (Engenharia de *Prompts*)** *A engenharia de prompts é o processo sistemático de concepção, elaboração, adaptação e avaliação de prompts com o objetivo de otimizar o desempenho, a utilidade e a interpretabilidade das respostas geradas por LLMs. Essa prática envolve o uso intencional de estruturas linguísticas, estratégias de condicionamento e técnicas de personalização para induzir comportamentos específicos no modelo, sem modificar seus parâmetros internos.*

Estudos recentes demonstram que variações aparentemente sutis na formulação de *prompts* podem provocar diferenças significativas nas saídas dos modelos, tanto em termos de conteúdo quanto de qualidade (Brown *et al.*, 2020; Wei *et al.*, 2022). Essa sensibilidade reforça a importância da engenharia de *prompts* como prática estruturada, especialmente em contextos que exigem precisão, rastreabilidade e controle interpretável.

O tipo dos *prompts* varia conforme a tarefa, o modelo utilizado e os objetivos de interação. Algumas das principais categorias incluem:

- *Prompts* instrutivos: comandos diretos, como “Explique o que é aprendizado supervisionado”.
- *Prompts* exemplificados: apresentam pares de entrada e saída como exemplos (típico do *few-shot prompting*).
- *Prompts* contextuais: incorporam dados históricos ou informações do usuário para personalização.
- *Prompts* multi-etapa: estruturam a geração em fases (ex.: *chain-of-thought*, ReAct).

Essas categorias se desdobram em diferentes estratégias, que serão discutidas na Seção 2.2.2.

Diversos termos recorrentes na literatura sobre LLMs e *prompting* são essenciais para a compreensão da engenharia de *prompts*:

- *In-context learning*: aprendizagem induzida por exemplos contidos no próprio *prompt*, sem ajuste de pesos do modelo.
  - *Zero-shot prompting*: realização da tarefa com base em instrução textual pura, sem exemplos.
  - *Few-shot prompting*: o *prompt* apresenta alguns exemplos para guiar a resposta do modelo.
- *Prompt template*: estrutura genérica parametrizável de *prompt*, usada para padronização.
- *Prompt injection*: técnica de inserção deliberada de instruções no *prompt*, podendo ser usada tanto para personalização quanto para ataques adversariais.
- *Instruction tuning*: processo de ajuste fino do modelo com um grande conjunto de instruções e respostas, promovendo maior aderência a comandos textuais (Ouyang *et al.*, 2022).

### 2.2.2 Paradigmas e Estratégias de Engenharia de Prompts

A evolução da engenharia de *prompts* tem sido marcada pelo surgimento de diversos paradigmas que buscam explorar a capacidade dos LLMs de generalizar comportamentos a partir de instruções textuais. Esses paradigmas refletem diferentes estratégias para instruir, guiar e modular o raciocínio dos modelos, com variados graus de complexidade e exigência cognitiva. Nesta seção, analisa-se os principais paradigmas consolidados na literatura, com foco nas abordagens de *zero-shot prompting*, *few-shot prompting*, *chain-of-thought prompting* e *ReAct prompting*.

O paradigma de *zero-shot prompting* consiste em solicitar que o modelo realize uma tarefa específica a partir de uma única instrução textual, sem fornecer exemplos prévios. Essa abordagem explora a capacidade latente dos LLMs de interpretar comandos em linguagem natural e realizar inferências generalizadas com base em seus conhecimentos adquiridos durante o pré-treinamento.

Por exemplo, ao apresentar o *prompt* “Traduza para o inglês: ‘O céu está azul’”, espera-se que o modelo compreenda tanto a tarefa quanto o conteúdo sem necessidade de exemplos adicionais. O trabalho de Brown *et al.* (2020), ao introduzir o GPT-3, destacou a surpreendente competência dos modelos em tarefas *zero-shot*, mesmo em domínios complexos, desde que a instrução seja formulada de forma clara e específica. A principal vantagem dessa abordagem reside em sua simplicidade e economia de *tokens*, sendo particularmente útil em aplicações de uso geral. Contudo, sua performance pode ser limitada em tarefas que exigem raciocínio abstrato, conhecimento de formato ou sensibilidade semântica específica.

No *few-shot prompting*, o *prompt* inclui uma pequena quantidade de exemplos de entrada e saída, que servem como referência para orientar o comportamento do modelo. Essa técnica simula um mecanismo de *in-context learning*, no qual o modelo infere padrões a partir dos exemplos apresentados, sem atualização explícita de seus parâmetros. Um *prompt few-shot* típico para uma tarefa de classificação de sentimentos pode ter o seguinte formato:

**Texto:** “O filme foi incrível.” Sentimento: positivo.

**Texto:** “A comida estava horrível.” Sentimento: negativo.

**Texto:** “O atendimento foi excelente.” Sentimento:

O modelo, então, completa a sequência com “positivo”, inferindo a lógica a partir dos pares anteriores. Essa estratégia foi amplamente demonstrada por Brown *et al.* (2020) como

uma das mais eficazes para generalização rápida em novos domínios, especialmente quando não se dispõe de dados suficientes para *fine-tuning* supervisionado.

Embora ofereça melhor desempenho em comparação ao *zero-shot*, o *few-shot prompting* é sensível à escolha e ordem dos exemplos, bem como à distribuição semântica dos dados incluídos no *prompt* (Lu *et al.*, 2022). Além disso, o limite de *tokens* nas janelas de contexto dos modelos impõe restrições à quantidade de exemplos que podem ser incluídos.

*Chain-of-thought prompting* é uma estratégia desenvolvida para estimular raciocínio explícito em tarefas que exigem múltiplas etapas cognitivas. Em vez de apenas fornecer uma resposta final, o modelo é induzido a gerar uma sequência de passos intermediários que explicam sua linha de pensamento. Essa abordagem foi formalizada por Wei *et al.* (2022), que demonstraram ganhos significativos em tarefas matemáticas, lógicas e de compreensão verbal.

Um exemplo clássico é:

**Pergunta:** Se João tem cinco maçãs e dá duas para Maria, quantas maçãs João tem agora?

**Resposta:** João começou com cinco maçãs. Ele deu duas para Maria. Então ele tem cinco menos duas, que são três.

**Resposta final:** três.

Essa técnica se mostra especialmente útil para mitigar a superficialidade de raciocínio que frequentemente caracteriza respostas de LLMs. Ao estruturar o caminho lógico, o modelo reduz o risco de respostas erradas por “atalhos estatísticos” e permite maior transparência no processo de inferência.

No entanto, nem todos os modelos são igualmente responsivos a essa técnica. Modelos menos treinados ou sem instrução específica para raciocínio estruturado podem gerar respostas confusas ou truncadas. Além disso, a necessidade de espaço extra para as etapas intermediárias impacta negativamente o uso em contextos com restrição de gastos de *tokens*.

O paradigma ReAct (*Reasoning + Acting*), proposto por Yao *et al.* (2023), combina raciocínio verbal explícito com ações de consulta a ferramentas externas. Esse modelo propõe uma estrutura iterativa em que o LLM raciocina sobre uma pergunta, decide uma ação (como chamar uma ferramenta de busca ou uma base de dados), observa os resultados, reflete novamente e repete o ciclo até chegar à resposta final.

A estrutura básica de uma interação ReAct segue o padrão:

**Pergunta:** Qual é a capital da França?

**Pensamento:** Preciso verificar qual cidade é a capital da França.

**Ação:** pesquisar(“capital da França”)

**Observação:** Resultado encontrado: Paris.

**Pensamento:** Com base na observação, a capital é Paris.

**Resposta final:** Paris.

Esse paradigma representa um avanço importante por permitir a interação de LLMs com ambientes externos de forma controlada, promovendo respostas mais precisas, auditáveis e contextualizadas. Sua aplicação é particularmente promissora em sistemas de RAG, onde o modelo pode iterar entre raciocínio e consultas estruturadas (por exemplo, Linguagem de Consulta Estruturada (SQL) ou grafos de conhecimento).

O ReAct, no entanto, exige infraestrutura adicional para suporte à execução de ações, gestão do histórico de raciocínios e controle de fluxo. Além disso, modelos precisam ser adaptados para entender e seguir esse ciclo de raciocínio e ação, o que nem sempre é garantido em implementações genéricas.

Cada um dos paradigmas apresentados oferece vantagens específicas, moldando a forma como os LLMs interagem com o conteúdo, o usuário e o ambiente. A escolha do paradigma ideal depende do tipo de tarefa, da necessidade de explicabilidade, da disponibilidade de dados e do contexto de aplicação. Estratégias mais avançadas, como *chain-of-thought* e ReAct, são mais indicadas quando o raciocínio transparente e o acesso a fontes externas são cruciais.

Na presente tese, esses paradigmas fornecem a base para compreender as escolhas de design e a estrutura do modelo PersonaReAct, que se apoia explicitamente no ciclo ReAct para promover personalização iterativa e contextualizada das respostas geradas por LLMs.

### 2.2.3 Personalização via Prompts

A personalização via *prompts* refere-se ao uso de estratégias específicas para adaptar a geração de respostas de modelos de linguagem de acordo com características, preferências ou necessidades particulares de um usuário ou grupo de usuários. Essa abordagem busca tornar a interação com modelos mais eficaz, contextualizada e relevante, especialmente em aplicações sensíveis como educação, saúde, suporte técnico e comunicação especializada.

As estratégias de personalização por *prompting* podem ser agrupadas em quatro cate-

gorias principais: injeção de contexto, uso de personas, enriquecimento de perfil e refinamento iterativo.

A injeção de contexto consiste em fornecer ao modelo informações históricas ou situacionais diretamente no texto do *prompt*. Isso pode incluir interações passadas, objetivos da tarefa, limitações impostas ou metadados associados ao usuário. Trabalhos como Liu *et al.* (2023b) demonstram que a inserção de contexto melhora significativamente a relevância e a coerência das respostas, especialmente em tarefas conversacionais contínuas ou orientadas a objetivos. No entanto, essa abordagem impõe desafios relacionados ao orçamento de *tokens*, podendo prejudicar o desempenho em tarefas complexas com janelas de contexto limitadas.

A segunda estratégia é o uso de personas, que envolve instruir o modelo a assumir papéis específicos com base em características demográficas, sociais, culturais ou até fictícias (Jiang *et al.*, 2023). Essa técnica tem sido amplamente utilizada em aplicações de *chatbots* e agentes conversacionais, com o objetivo de manter consistência de estilo, tom e perspectiva ao longo da interação. Personas também têm sido utilizadas para simular usuários em sistemas de recomendação ou perfis específicos de público em testes de usabilidade.

A terceira abordagem é o enriquecimento de perfil, que busca complementar os dados explícitos do usuário com inferências obtidas a partir do próprio modelo. Em vez de fornecer diretamente todas as características do usuário no *prompt*, o sistema pode consultar interações anteriores ou aplicar técnicas de compressão semântica para sintetizar um perfil representativo. Trabalhos como LLM-Rec (Lyu *et al.*, 2024) mostram que modelos podem ser usados como geradores de perfis latentes, capazes de recuperar preferências implícitas e características comportamentais a partir de dados mínimos de entrada. Essa abordagem melhora a escalabilidade e reduz a sobrecarga cognitiva no *prompt*, mas depende da capacidade de generalização do modelo e da precisão das inferências realizadas.

Por fim, o refinamento iterativo de *prompts* constitui uma abordagem em que a personalização é ajustada ao longo da interação com base no desempenho anterior do modelo, em *feedbacks* do usuário ou em heurísticas de melhoria. Técnicas como RePrompt (Chen *et al.*, 2025) e AutoPrompt (Shin *et al.*, 2020) representam esse paradigma, permitindo que *prompts* sejam ajustados automaticamente com base em critérios de qualidade ou sucesso. O refinamento pode incluir reformulações linguísticas, reordenação de informações ou até reescrita completa com objetivos mais específicos. Embora poderosa, essa abordagem exige mecanismos de avaliação confiáveis e, em alguns casos, um ciclo de experimentação que pode ser computacionalmente

custoso.

Essas quatro estratégias não são mutuamente exclusivas e podem ser combinadas em arquiteturas híbridas de personalização, especialmente em sistemas orientados a tarefas complexas ou a múltiplos perfis de usuários. A escolha da estratégia ideal depende do domínio de aplicação, da disponibilidade de dados do usuário, da janela de contexto disponível e da capacidade de processamento do sistema.

Apesar dos avanços recentes, a personalização via *prompts* ainda enfrenta limitações estruturais, como a falta de memória persistente, dificuldades de adaptação a longo prazo e sensibilidade à formulação textual. Esses desafios têm motivado a combinação de *prompting* com outras técnicas como *finetuning* leve, recuperação de informações e sistemas multiagentes, conforme será discutido nas seções subsequentes.

## 2.3 Agentes Baseados em Modelos de Linguagem

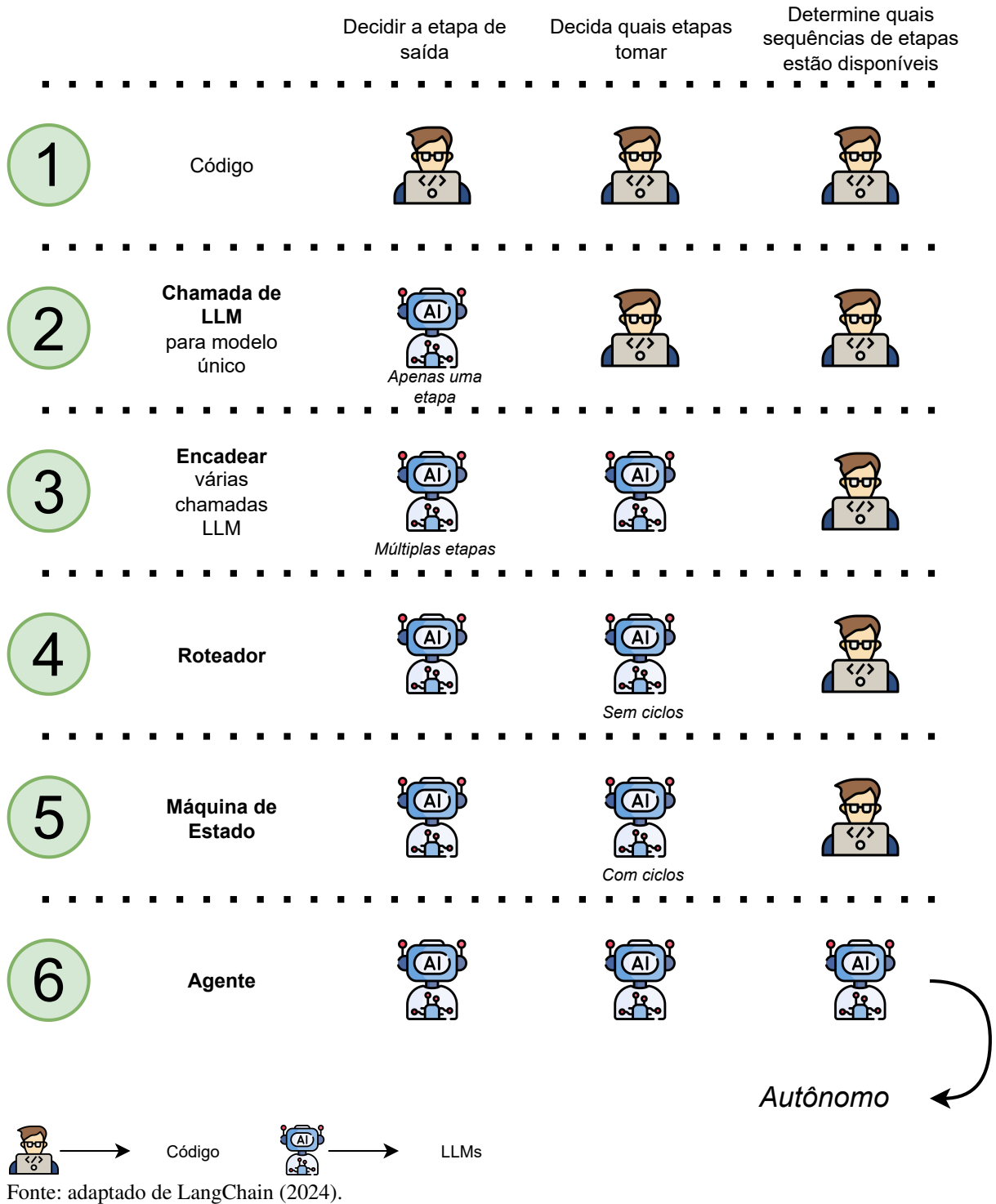
### 2.3.1 Fundamentos de Agentes com LLMs

O conceito de agente tem sido amplamente utilizado em diversas áreas da inteligência artificial para descrever entidades capazes de realizar ações de forma autônoma com base em percepções do ambiente. No contexto de LLMs, agentes são estruturas computacionais que integram raciocínio e atuação, orquestrando chamadas a modelos de linguagem e outras ferramentas para atingir objetivos definidos de forma dinâmica e adaptativa. Tais agentes são concebidos segundo o ciclo percepção-decisão-ação, no qual uma entrada (por exemplo, uma pergunta ou observação) é processada, uma decisão é tomada e uma ação é executada, possivelmente gerando novas percepções e reiniciando o ciclo.

A evolução da autonomia em sistemas baseados em LLMs pode ser representada como uma escala progressiva de sofisticação arquitetural, que vai desde chamadas simples a modelos de linguagem até agentes completos com capacidade de planejamento e execução autônoma. A Figura 1 ilustra de forma didática essa progressão, destacando o grau de controle exercido por código convencional versus decisões tomadas pelo próprio modelo de linguagem em cada nível da arquitetura.

No nível mais básico, temos o uso de código estático, em que todas as decisões e saídas são rigidamente programadas por humanos. Em seguida, surge a Chamada de LLM, onde um modelo de linguagem é consultado uma única vez, sendo ainda o código responsável

Figura 1 – Níveis de autonomia em arquiteturas baseadas em LLMs.



por estruturar a consulta e interpretar a resposta. O próximo estágio é o de Cadeias, nos quais múltiplas chamadas a LLMs são encadeadas em etapas predefinidas, ainda com lógica controlada por desenvolvedores.

A complexidade aumenta com os Roteadores, que introduzem lógica de decisão delegada ao modelo de linguagem, permitindo escolher entre múltiplos caminhos possíveis, mas

ainda sem retroalimentação ou ciclos. O passo seguinte são as Máquinas de Estado, que incorporam a capacidade de ciclos e reavaliação de estado, aproximando-se de um comportamento reativo. Finalmente, os Agentes são sistemas capazes de decidir a saída, qual passo executar e quais sequências estão disponíveis, utilizando exclusivamente o modelo de linguagem para tomada de decisão em tempo de execução. Isso os caracteriza como sistemas verdadeiramente autônomos, com comportamento emergente e adaptativo.

A formalização do ciclo percepção-decisão-ação em agentes com LLMs tem sido explorada por diversos autores. Shinn *et al.* (2023) propõe o paradigma *Reflexion*, no qual agentes são capazes de refletir sobre suas falhas e ajustar estratégias em iterações posteriores. Yao *et al.* (2023) introduz o ReAct, uma abordagem em que o raciocínio e a ação são intercalados, com o LLM decidindo tanto os passos quanto as ferramentas a serem acionadas.

Um ponto importante sobre os agentes baseados em LLMs é que eles não são estruturas monolíticas. Em vez disso, tendem a ser compostos por diversos componentes especializados: módulos de percepção que extraem informações do ambiente ou do histórico de diálogo, módulos de decisão que raciocinam sobre os próximos passos, e módulos de ação que executam comandos ou interagem com ferramentas externas. Em alguns casos, esses módulos podem ser implementados como subagentes especializados, compondo sistemas multiagente mais amplos, como é discutido na Subseção 2.3.2.

A crescente sofisticação dos agentes com LLMs impõe desafios de controle, confiabilidade e interpretabilidade. Quanto maior o grau de autonomia concedido ao modelo, mais difícil se torna prever e validar seu comportamento. Por isso, estratégias como *self-monitoring*, *prompt critic* e *LLM-as-a-Judge* têm sido exploradas para mitigar riscos e aprimorar a robustez dos agentes (Tong *et al.*, 2024).

Dessa forma, os agentes com LLMs representam o fim de um espectro de arquiteturas que transitam do controle humano explícito para a autonomia baseada em linguagem. Eles constituem uma nova fronteira para o desenvolvimento de sistemas inteligentes, capazes de raciocinar, planejar e agir de forma contextualizada e iterativa.

### **2.3.2 Arquiteturas Multiagente**

À medida que crescem em complexidade e ambição, os sistemas baseados em LLMs têm migrado de arquiteturas monolíticas para composições mais sofisticadas envolvendo múltiplos agentes especializados. Esses sistemas multiagente se beneficiam do princípio da

divisão de responsabilidades, no qual diferentes subagentes são projetados para desempenhar papéis específicos, promovendo modularidade, reuso e controle refinado sobre os processos de raciocínio e atuação.

Em um sistema multiagente, cada agente pode ser definido como uma entidade autônoma capaz de perceber, raciocinar e agir dentro de um ambiente compartilhado. A diferença em relação a agentes únicos é que, em vez de centralizar toda a lógica em um único LLM ou estrutura, as decisões são distribuídas entre agentes com objetivos locais, que cooperam ou competem para atingir metas globais.

Uma tipologia básica dos papéis desempenhados por agentes em arquiteturas multiagente com LLMs inclui:

- **Agentes de Raciocínio:** responsáveis por planejar ações ou analisar situações complexas, normalmente empregando técnicas como *chain-of-thought* (Wei *et al.*, 2022), *toolformer* (Schick *et al.*, 2023) ou *self-reflection* (Wang *et al.*, 2023).
- **Agentes de Execução:** especializados em interagir com ferramentas externas, Interfaces de Programação de Aplicações (APIs), bases de dados ou ambientes virtuais, geralmente após receber instruções claras.
- **Agentes de Recuperação:** encarregados de buscar informações relevantes em fontes específicas, como vetores semânticos, bancos de dados relacionais ou grafos de conhecimento.
- **Agente Orquestrador:** módulo central que coordena os demais, decidindo qual agente deve ser acionado com base no contexto e na pergunta, podendo ele mesmo ser um LLM ou uma combinação de lógica programada e chamadas a modelos de linguagem.

Esse tipo de arquitetura, por vezes denominado *agentic composition*, tem sido explorado em plataformas como LangGraph<sup>1</sup>, AutoGen<sup>2</sup>, CrewAI<sup>3</sup> e CAMEL<sup>4</sup>, entre outras. O trabalho de Li *et al.* (2023a) apresenta o *framework* CAMEL, onde dois agentes especializados (um como “assistente” e outro como “usuário”) interagem para resolver tarefas complexas, demonstrando capacidades emergentes de negociação, correção de erros e divisão de trabalho. Já Wu *et al.* (2023) propõem o AutoGen, um *framework* que permite compor agentes LLMs interativos com habilidades diferenciadas, facilitando o design de fluxos complexos com *feedback* interno entre os agentes.

Uma das vantagens fundamentais de arquiteturas multiagente é a capacidade de

<sup>1</sup> <https://www.langchain.com/langgraph>

<sup>2</sup> <https://microsoft.github.io/autogen/stable//index.html>

<sup>3</sup> <https://www.crewai.com/>

<sup>4</sup> <https://www.camel-ai.org/>

lidar com tarefas que exigem múltiplas competências simultaneamente, como consultas que envolvem raciocínio lógico, busca de dados numéricos, adaptação ao perfil do usuário e geração de linguagem natural interpretável. Em vez de sobrecarregar um único modelo com todos esses requisitos, a divisão em papéis facilita tanto o controle quanto a extensibilidade. Além disso, permite a inserção de validações cruzadas, crítica de respostas (por meio de agentes *critics*) e até mesmo comportamentos de meta-aprendizado, onde um agente aprende a ajustar o comportamento de outros com base em *feedback*.

Entretanto, arquiteturas multiagente também trazem desafios significativos. A coordenação entre agentes pode gerar *overhead* computacional, risco de inconsistência nas mensagens trocadas e dificuldade de depuração. Além disso, há preocupações sobre a opacidade e a rastreabilidade das decisões quando múltiplos LLMs interagem em ciclos recursivos. Soma-se a isso o aumento do custo operacional, já que cada troca de mensagem entre agentes implica consumo de *tokens*, o que pode tornar o sistema economicamente inviável em larga escala. Por isso, abordagens modernas vêm integrando mecanismos de *logging*, visualização gráfica dos fluxos de decisão e restrições de escopo para evitar efeitos indesejados.

O futuro das arquiteturas multiagente com LLMs parece promissor, sobretudo em domínios como automação de tarefas complexas, agentes pedagógicos personalizados, simulações econômicas e sistemas de apoio à decisão em tempo real. O aumento da capacidade de contexto dos LLMs e o avanço em técnicas de controle sobre esses modelos deverão permitir arquiteturas mais robustas e explicáveis, conciliando flexibilidade linguística com segurança operacional.

## **2.4 Ecossistema Tecnológico para Construção de Sistemas com LLMs**

### **2.4.1 Principais Modelos de Linguagem**

A escolha do modelo de linguagem a ser utilizado em um sistema baseado em LLMs impacta diretamente sua capacidade de raciocínio, geração textual, compreensão contextual e integração com ferramentas externas. Nos últimos anos, diversas arquiteturas têm sido desenvolvidas por empresas, universidades e comunidades de código aberto, cada uma com suas particularidades em termos de desempenho, escalabilidade, custo e licenciamento. Esta subseção apresenta os principais modelos disponíveis no ecossistema atual, destacando suas características, diferenças técnicas e aplicações mais comuns.

GPT é uma das famílias mais influentes de modelos de linguagem, desenvolvida pela

OpenAI. Seu marco inicial foi o GPT-2, mas a popularização da arquitetura se consolidou com o GPT-3, que apresentou capacidades impressionantes de geração de texto e raciocínio. A evolução para o GPT-3.5 e, posteriormente, o GPT-4 representou avanços substanciais em robustez, contextualização e capacidade multimodal. Os modelos da linha GPT seguem a abordagem de pré-treinamento não supervisionado em larga escala, com *fine-tuning* supervisionado e RLHF para alinhamento. O GPT-4, por exemplo, já opera de forma multimodal, aceitando entradas de texto e imagem, e serve de base para o sistema ChatGPT da OpenAI (OpenAI *et al.*, 2024).

Modelos de Linguagem Pathways (PaLMs) são uma série de modelos desenvolvidos pelo Google DeepMind, cuja arquitetura baseia-se no uso da infraestrutura Pathways, projetada para treinar modelos de forma mais eficiente e especializada. O PaLM 2, em particular, destaca-se pelo uso de dados em múltiplos idiomas e múltiplas modalidades, otimizando sua performance em tarefas de tradução, codificação e raciocínio lógico. A integração do PaLM com o ecossistema da Google, como Bard e ferramentas empresariais do Google Cloud, reforça sua adoção em ambientes corporativos (Chowdhery *et al.*, 2023).

LLaMAs são uma família de modelos lançada pela Meta com foco na comunidade acadêmica e de código aberto. Os modelos LLaMAs foram projetados para serem eficientes e acessíveis, com variações como LLaMA 2 oferecendo múltiplos tamanhos (por exemplo, 7B, 13B, 65B e 70B parâmetros) e LLaMA 3, permitindo treinamento e inferência em ambientes com recursos computacionais mais limitados. Sua popularidade entre a comunidade de pesquisa levou ao desenvolvimento de variantes ajustadas como Vicuna, Alpaca e Orca. Os modelos LLaMAs utilizam apenas texto durante o treinamento e têm foco em instruções genéricas, o que os torna flexíveis para múltiplas tarefas (Touvron *et al.*, 2023).

Claude, desenvolvido pela Anthropic, representa uma abordagem diferenciada ao enfatizar alinhamento com valores humanos e segurança. Utilizando um processo chamado Constitutional AI (Bai *et al.*, 2022), os modelos Claude (Claude 1, Claude 2 e Claude 3) são treinados para seguir princípios éticos predefinidos e evitar comportamentos indesejados. Claude é particularmente utilizado em aplicações onde a segurança do conteúdo gerado e a conformidade com diretrizes normativas são prioritárias. Além disso, Claude demonstra forte desempenho em tarefas de raciocínio complexo, explicação e síntese de documentos.

Mistral é um modelo de código aberto que se destaca pela eficiência de treinamento e inferência. A arquitetura Mistral 7B foi lançada como um modelo denso, competitivo com modelos maiores em *benchmarks* de linguagem geral. Posteriormente, o lançamento do

Mixtral, um modelo do tipo Mistura de Especialistas (MoE), elevou o desempenho ao combinar especialistas distintos ativados dinamicamente em cada inferência. Essa abordagem permite alta capacidade com baixo custo de inferência. Os modelos da Mistral vêm sendo amplamente utilizados em *frameworks* de agentes LLM abertos, oferecendo uma alternativa acessível a modelos proprietários (Jiang *et al.*, 2024).

Além desses, o cenário atual inclui uma diversidade de modelos adaptados a contextos específicos. O Gemini, também do Google DeepMind, integra capacidades multimodais avançadas. O Command R da Cohere foca em RAG. O Yi da 01.AI apresenta desempenho notável em tarefas de linguagem em chinês e inglês. A existência desses modelos permite aos pesquisadores e desenvolvedores selecionarem soluções conforme os requisitos técnicos, legais e financeiros de cada aplicação (AI *et al.*, 2025).

#### **2.4.2 Frameworks e Bibliotecas para Agentes com LLMs**

A construção de sistemas inteligentes baseados em modelos de linguagem de grande escala demanda não apenas o acesso a LLMs de alto desempenho, mas também o uso de *frameworks* que facilitem a orquestração de fluxos de interação, integração com ferramentas externas e controle sobre o comportamento do sistema. Tais *frameworks* abstraem as complexidades envolvidas na criação de agentes autônomos, permitindo que desenvolvedores implementem raciocínio iterativo, tomada de decisão baseada em observações, uso de memória, roteamento entre modelos e recuperação de informações externas. Esta subseção descreve os principais *frameworks* atualmente utilizados para a construção de agentes com LLMs, destacando suas características, arquitetura e aplicações.

LangChain é atualmente um dos *frameworks* mais amplamente adotados para desenvolvimento de aplicações com LLMs. Ele fornece uma estrutura modular que permite combinar diferentes componentes, como modelos de linguagem, ferramentas externas (por exemplo, mecanismos de busca, APIs e bancos de dados), memória de contexto, e lógica de encadeamento de *prompts*. O LangChain introduz o conceito de cadeias, que representam sequências de chamadas a modelos com controle explícito sobre a entrada, saída e transformações intermediárias. Além disso, oferece suporte a ferramentas que permitem ao agente realizar ações externas, e agentes que tomam decisões sobre quais ferramentas utilizar com base nas respostas dos LLMs. Sua compatibilidade com múltiplos *backends* de LLMs (OpenAI, Cohere, Anthropic, Hugging Face, entre outros) e com diversas fontes de dados, como Pinecone, FAISS e bases SQL, tornou-o uma

escolha versátil para aplicações em múltiplos domínios.

LangGraph surge como uma extensão do LangChain voltada à modelagem explícita de fluxos de controle por meio de grafos de estados. Diferentemente das cadeias tradicionais que seguem fluxo linear, o LangGraph permite que o desenvolvedor defina máquinas de estado e grafos de decisão, com ciclos e transições condicionais controladas por LLMs ou por regras programáticas. Essa abordagem permite maior expressividade no controle da execução de agentes, especialmente em tarefas iterativas ou que demandam raciocínio deliberativo. O LangGraph combina o paradigma ReAct (*Reasoning + Acting*) com a capacidade de persistência de estado, o que favorece o desenvolvimento de arquiteturas multiagente, sistemas de criticismo interno e fluxos baseados em metas.

Além desses, diversos outros *frameworks* vêm sendo explorados para a construção de agentes baseados em LLMs. O Haystack<sup>5</sup>, desenvolvido pela Deepset, é voltado para aplicações de Pergunta-Resposta com RAG e permite encadeamento de componentes como pré-processadores, pipelines de busca, e módulos de geração. O CrewAI propõe uma abordagem inspirada em equipes de agentes colaborativos, com papéis bem definidos e protocolos de comunicação. Já o AutoGen, da Microsoft Research, implementa agentes interativos com ciclos de reflexão e coordenação, integrando múltiplos LLMs e ferramentas em ciclos de raciocínio conjunto. O DSPy, proposto por Khattab *et al.* (2023) adota uma perspectiva declarativa sobre o raciocínio, permitindo compilar fluxos de tarefas supervisionadas como grafos semânticos ajustáveis com base em *feedback* de desempenho.

A escolha do *framework* mais adequado depende de múltiplos fatores, incluindo a complexidade do raciocínio desejado, o grau de autonomia do agente, a necessidade de controle sobre o fluxo de execução e os requisitos de integração com dados externos.

### **2.4.3 Recuperação de Informações Estruturadas**

A recuperação de informações estruturadas desempenha um papel central na construção de sistemas baseados em modelos de linguagem que necessitam acessar dados precisos e atualizados. Diferentemente das abordagens tradicionais de recuperação de documentos, que se baseiam em texto livre, a recuperação estruturada envolve o acesso direto a bases de dados relacionais, APIs e outros repositórios que organizam a informação em formatos tabulares, hierárquicos ou semânticos. Esse acesso permite que modelos de linguagem forneçam respostas

---

<sup>5</sup> <https://haystack.deepset.ai/>

exatas, auditáveis e verificáveis, especialmente em domínios como governo digital, medicina, finanças, esportes e ciências sociais.

Um dos avanços mais significativos nesse contexto é o desenvolvimento de técnicas de *Text-to-SQL* (Zhong *et al.*, 2017), que consistem na tradução automática de perguntas em linguagem natural para comandos SQL válidos. Essa tarefa de mapeamento semântico-natural requer que o modelo compreenda a intenção do usuário, identifique as tabelas relevantes, selecione os atributos corretos e construa cláusulas apropriadas para filtragem, ordenação e agregação. Modelos como SQLNet (Xu *et al.*, 2017) e mais recentemente T5-SQL e Codex (Chen *et al.*, 2021) demonstraram capacidade notável na conversão de linguagem natural para SQL, alcançando bons resultados em *benchmarks* como o Spider. O uso de esquemas de banco como contexto adicional, muitas vezes representado via *embeddings* ou descrições naturais, também tem sido essencial para melhorar a precisão desses modelos.

Em sistemas com agentes baseados em LLMs, o uso de *Text-to-SQL* é frequentemente integrado por meio de ferramentas especializadas que conectam o modelo de linguagem a bancos relacionais. *Frameworks* como LangChain e Semantic Kernel permitem definir interfaces para consultas parametrizadas, onde o agente pode gerar comandos SQL, executá-los e incorporar as respostas estruturadas ao seu raciocínio. Essa abordagem é vantajosa em relação à recuperação textual tradicional, pois permite consultas mais precisas, com menor ambiguidade, além de garantir rastreabilidade dos dados utilizados para gerar uma resposta.

Além dos bancos relacionais, o uso de APIs estruturadas é uma alternativa complementar. APIs RESTful, GraphQL ou específicas de domínio oferecem *endpoints* bem definidos para acesso a informações de forma programática. Modelos de linguagem podem ser treinados ou ajustados para compreender a documentação dessas APIs e gerar chamadas adequadas. Por exemplo, em tarefas como obtenção de dados meteorológicos, resultados esportivos ou status logístico, agentes baseados em LLMs podem consultar APIs externas, interpretar a resposta JSON e integrar essas informações ao fluxo de diálogo ou à resposta final. A geração de chamadas de API por linguagem natural é um subcampo emergente, que explora técnicas semelhantes ao *Text-to-SQL*, com desafios adicionais de autenticação, versionamento e interpretação da semântica dos parâmetros.

Outro ponto relevante é a validação das informações estruturadas recuperadas. Uma vez que os LLMs podem sofrer de alucinação ou inconsistência, é fundamental implementar mecanismos que assegurem que a resposta final ao usuário esteja ancorada nos dados estruturados

obtidos. Isso pode ser feito por meio de *prompt templates* que forçam a referência explícita às variáveis recuperadas, ou por meio de arquiteturas que separam claramente a geração da resposta e a execução de comandos estruturados, conforme sugerido por paradigmas como o ReAct e o AutoCoT (Zhou *et al.*, 2023).

A utilização de informações estruturadas também viabiliza aplicações de maior complexidade, como recomendações personalizadas, análise de dados multimodais, simulações baseadas em regras de negócio, e interações com grafos de conhecimento. O acoplamento entre linguagem natural e dados formais oferece um caminho promissor para a criação de agentes explicáveis, auditáveis e confiáveis, essenciais em contextos de alta responsabilidade.

### 3 TRABALHOS RELACIONADOS

A personalização tem se consolidado como uma das estratégias mais eficazes para melhorar o desempenho de LLMs em tarefas específicas (Lyu *et al.*, 2024). Dentre as diversas abordagens existentes, a personalização via *prompting* tem se destacado pela simplicidade de implementação e pela flexibilidade em adaptar o comportamento dos modelos sem a necessidade de reconfiguração de seus parâmetros internos. Essa técnica consiste na elaboração e modificação de instruções fornecidas ao modelo, moldando sua geração textual conforme características contextuais, perfis ou personas do usuário.

No escopo desta tese, o PersonaReAct posiciona-se no campo das arquiteturas que combinam raciocínio e atuação (ReAct), utilizando estratégias de personalização via *prompts* como principal mecanismo para adaptar as respostas ao perfil funcional do usuário. Para contextualizar e comparar a proposta desta tese com o estado da arte, adotamos as quatro categorias de personalização por *prompting* sistematizadas por Zhang *et al.* (2025), que organizam a literatura recente segundo diferentes estratégias de adaptação das respostas de modelos de linguagem aos usuários.

Este capítulo apresenta essas quatro categorias, que servem como referencial para a comparação das contribuições do PersonaReAct em relação a abordagens consolidadas.

- A Seção 3.1 explora os **Prompts Contextuais**, que consistem na inserção direta de informações históricas do usuário no texto de entrada do modelo, fornecendo contexto personalizado para a geração de respostas.
- A Seção 3.2 discute os **Prompts baseados em Personas**, nos quais o modelo é instruído a assumir papéis específicos com base em características demográficas, comportamentais ou fictícias do usuário, com o objetivo de simular estilos ou perspectivas.
- A Seção 3.3 trata dos **Prompts com Enriquecimento por Perfil**, que utilizam o conhecimento interno dos próprios modelos para complementar ou condensar perfis de usuários a partir de interações passadas.
- Por fim, a Seção 3.4 apresenta técnicas de **Refinamento de Prompts**, cujo foco é a melhoria iterativa de instruções com base em desempenho anterior, *feedbacks* ou heurísticas de otimização, tornando os *prompts* cada vez mais eficazes e personalizados.

### 3.1 Prompts Contextuais

*Prompts* contextuais representam uma estratégia de personalização baseada na injeção direta de informações relevantes sobre o usuário no próprio *prompt*, como histórico de interações, preferências ou comportamentos anteriores. Em vez de modificar a arquitetura do modelo ou realizar ajuste fino (*fine-tuning*), essa abordagem explora a capacidade dos LLMs de interpretar instruções complexas e gerar respostas condicionadas ao conteúdo textual inserido no contexto da entrada.

A literatura recente apresenta diversos estudos que avaliam a eficácia dessa abordagem, especialmente em sistemas de recomendação.

No trabalho de Palma (2023), os autores propõem o *Retrieval-Augmented Recommender System*, uma arquitetura que combina um mecanismo de recuperação de documentos com LLMs para realizar recomendações personalizadas. A ideia central é a construção de um *prompt* que incorpora não apenas o histórico de interações do usuário, mas também informações de contexto relevantes recuperadas de uma base documental. O modelo utilizado é o GPT-3.5 Turbo, com contexto construído dinamicamente a partir de consultas em um índice vetorial. Esse processo torna possível a adaptação das respostas sem a necessidade de ajustes finos, destacando o potencial da abordagem para cenários de *cold start*, ou seja, situações em que o sistema precisa operar com usuários ou itens para os quais há pouca ou nenhuma informação histórica disponível. No entanto, o sistema ainda opera com um modelo único e linear, o que pode limitar a gestão da complexidade crescente dos perfis de usuários e das nuances de personalização requeridas.

Wang e Lim (2023) utilizam *prompts* contendo listas de itens previamente consumidos pelo usuário para prever o próximo item em um cenário de recomendação. A abordagem mostrou resultados promissores mesmo em situações de *zero-shot*, ou seja, sem treinamento adicional.

Sanner *et al.* (2023) demonstram que LLMs podem superar abordagens tradicionais de recomendação em cenários de *cold start*, desde que o *prompt* inclua informações linguísticas e baseadas em itens que caracterizam os interesses do usuário.

Estudos como os de Li *et al.* (2023b) e Liu *et al.* (2023a) também mostram que a performance de LLMs melhora significativamente quando há um histórico textual bem formatado no *prompt*. Esses estudos reforçam o papel do contexto na personalização de respostas.

Outros trabalhos, como o de Christakopoulou *et al.* (2023), destacam a importância da progressão temporal dos interesses do usuário. Os autores propõem *prompts* que simulam

uma “jornada de interesses”, organizando eventos em sequência para gerar recomendações mais coerentes com o momento atual do usuário.

Liu *et al.* (2023b) destaca uma limitação importante dos *prompts* contextuais. Os autores observam a tendência dos LLMs a darem menos atenção às informações que estão no meio de contextos longos. Esse efeito limita a eficácia da personalização quando o histórico do usuário ultrapassa a capacidade de atenção do modelo.

Por fim, o trabalho de Kang *et al.* (2023) mostra que, mesmo quando expostos a informações detalhadas sobre preferências anteriores, os LLMs podem falhar em inferir corretamente as preferências implícitas. Isso indica que a simples presença de contexto no *prompt* não garante personalização eficaz.

O PersonaReAct também se beneficia do uso de contexto no *prompt*, especialmente por incluir, quando disponível, o histórico de raciocínio do ReAct. No entanto, seu diferencial está na combinação desse contexto com a orientação explícita de personalização baseada em perfil funcional, característica que o posiciona além da abordagem puramente contextual.

Enquanto os trabalhos acima se concentram em preferências inferidas a partir de interações anteriores, o PersonaReAct atua de forma mais direta. O perfil funcional é informado ao modelo como uma variável explícita (*profile*). A atribuição desse valor não é feita pelo PersonaReAct em si, mas pela aplicação que o utiliza, podendo ocorrer de diferentes maneiras: por exemplo, por meio de um menu de seleção, de uma configuração inicial fornecida no início da sessão, ou ainda a partir de informações declaradas pelo usuário em um *chat*. Uma vez definido, o *prompt* instrui o LLM a simular “mentalmente” o ponto de vista associado a esse perfil. Esse modelo híbrido, que combina contexto e simulação de persona, visa mitigar problemas como os descritos por Liu *et al.* (2023b) relacionados à perda de atenção em contextos longos. Isso é feito ao dividir a tarefa de personalização em etapas e tornar explícitas as instruções no próprio *prompt*.

### 3.2 Prompts Baseados em Personas

Os *prompts* baseados em *personas* representam uma estratégia de personalização em que os LLMs são instruídos a adotar ou simular características específicas de um papel, identidade ou traço psicológico durante a geração de texto. Em vez de apenas considerar o histórico de interações ou informações contextuais, o modelo é explicitamente orientado a assumir uma perspectiva ou comportamento de um agente com identidade definida, seja um

especialista, um personagem fictício ou um usuário com traços de personalidade específicos.

Essa abordagem tem ganhado relevância em pesquisas que investigam a capacidade dos LLMs de simular seres humanos, tanto em contextos de estudos de comportamento quanto para fins de personalização interativa. Por exemplo, Aher *et al.* (2023) demonstram que LLMs podem simular múltiplos participantes humanos em experimentos de ciências sociais, atribuindo a cada um características de personalidade e históricos comportamentais distintos. Ao utilizar descrições ricas de *personas*, os modelos geram respostas com variações que se aproximam de amostras humanas reais. No entanto, o trabalho é focado em ambientes de simulação e não em aplicações interativas com usuários reais. Além disso, o controle sobre a continuidade da persona ao longo de múltiplas interações ainda é limitado.

Argyle *et al.* (2023) reforçam a ideia de que variações em *personas* afetam significativamente a geração textual. Os autores mostram que LLMs são sensíveis às descrições de identidade presentes nos *prompts*, o que pode ser explorado tanto para simular grupos populacionais quanto para identificar vieses gerados a partir dessas descrições. O estudo, porém, utiliza personas genéricas e pouco contextualizadas, o que reduz sua aplicabilidade em cenários funcionais e personalizados.

O trabalho de Chen *et al.* (2024) oferece uma revisão sistemática sobre agentes de linguagem baseados em papéis, categorizando as abordagens em simulação de papéis sociais, emulação de traços de personalidade e imitação de personagens. A revisão aponta que a personalização via persona é particularmente eficaz quando o papel é bem definido e o *prompt* fornece instruções explícitas para que o modelo incorpore essa identidade de forma consistente.

Outro estudo relevante é o de Wang *et al.* (2024a), que propõe avaliar a fidelidade da simulação de personalidade por meio de entrevistas psicológicas. O trabalho mostra que LLMs podem sustentar traços de personalidade consistentes ao longo de múltiplas respostas, desde que as instruções no *prompt* sejam claras e reiteradas.

Entretanto, essas abordagens também enfrentam desafios. Gupta *et al.* (2024) destaca que atribuir personas pode amplificar vieses implícitos nos modelos, o que levanta preocupações éticas e metodológicas. Além disso, estudos como o de Deshpande *et al.* (2023) sugerem que personas podem intensificar conteúdos tóxicos ou polarizados, caso as instruções não sejam cuidadosamente elaboradas.

O PersonaReAct se insere diretamente no escopo dos *prompts* baseados em personas, mas com características distintas. Em sua arquitetura, o perfil funcional do usuário é incorporado

explicitamente como variável no estado de execução do sistema (`profile`) e interpretado por um nó específico (`persona_node`) responsável por simular o ponto de vista associado a esse perfil. Essa simulação é orientada por *prompts* cuidadosamente elaborados, inspirados por técnicas de *expert prompting* e *role alignment*, como as discutidas por Xu *et al.* (2024) e Wang *et al.* (2024b).

Na prática, a atribuição desse perfil depende da aplicação que utiliza o PersonaReAct. Por exemplo, em um *chatbot* no domínio esportivo, o usuário pode selecionar no início da interação se deseja receber respostas como “Técnico”, “Torcedor” ou “Jornalista Esportivo”, e essa escolha é armazenada na variável `profile`. Em um sistema médico, por outro lado, o perfil poderia ser definido previamente de acordo com as credenciais de acesso do usuário (ex.: “Médico”, “Paciente” ou “Pesquisador”). Assim, o valor de `profile` é fornecido externamente ao *framework* e utilizado pelo `persona_node` para orientar a personalização das respostas.

Diferentemente de trabalhos que aplicam personas genéricas ou personagens fictícios, o PersonaReAct define personas com base em papéis funcionais dentro de um contexto específico. Essa orientação funcional permite que o sistema ajuste não apenas o tom e o estilo da resposta, mas também o foco informacional e o tipo de linguagem utilizada, promovendo uma adequação mais direta às necessidades do usuário.

Outro diferencial é o uso de um fluxo multiagente que isola a etapa de simulação da persona das demais etapas da geração, permitindo maior controle sobre o comportamento do modelo e facilitando testes comparativos. Essa modularidade evita também o acúmulo de instruções dentro de um único *prompt* extenso.

Em resumo, embora compartilhe fundamentos com abordagens contemporâneas de simulação baseada em personas, o PersonaReAct avança ao estruturar a personalização com separação de responsabilidades e controle explícito de entrada e saída, integrando os princípios de *role-based prompting* a um fluxo de raciocínio orientado.

### 3.3 Prompts com Enriquecimento por Perfil

O uso de *prompts* com enriquecimento por perfil surge como uma resposta à limitação de dados escassos, irrelevantes ou incompletos em sistemas de personalização. Essa abordagem busca complementar os perfis de usuários com conhecimento interno dos LLMs, enriquecendo os dados disponíveis e potencializando a performance dos modelos em tarefas personalizadas. A seguir, discutimos os principais trabalhos dessa categoria, evidenciando suas estratégias e limitações, e ao final comparamos com o PersonaReAct.

Shi *et al.* (2023) e Liu *et al.* (2024a) chamam atenção para um problema recorrente: mesmo perfis extensos podem conter ruído ou informações irrelevantes, prejudicando o desempenho dos LLMs. O trabalho de Liu *et al.* (2024b), ONCE, propõe a geração de perfis enriquecidos a partir do histórico de navegação do usuário, com resumos sobre tópicos e regiões de interesse extraídos via *prompting*. Essa contextualização ajuda os LLMs a inferirem preferências implícitas, superando limitações de dados frios.

Lyu *et al.* (2024), por sua vez, com o LLM-REC, define quatro estratégias de *prompting* para enriquecer descrições de itens que são insuficientes, permitindo uma melhor associação com as preferências dos usuários. Xi *et al.* (2024) introduz o conceito de *factorization prompting*, que extrai características latentes dos usuários e itens e as converte em vetores contextualizados através de um adaptador especializado. Isso permite maior compatibilidade com modelos preexistentes de recomendação.

Sun *et al.* (2025) apresentam o *Persona-DB*, uma abordagem hierárquica que constrói personas a partir de históricos de interação usando LLMs. Seu diferencial está no refinamento colaborativo, em que dados de usuários semelhantes são utilizados para ajustar e aprimorar o perfil individual, melhorando a previsão de respostas personalizadas com menos dados diretos.

Complementarmente, Perez *et al.* (2007) e Dumitru *et al.* (2011) discutem a construção de perfis mesmo quando não se conhece a totalidade das preferências dos usuários (um cenário descrito como relação de preferência incompleta). Nesses casos, nem todos os itens podem ser comparados entre si, devido à ausência de dados explícitos. Além disso, os autores exploram a extração de características com base em descrições públicas de produtos. Essas abordagens fundamentam a motivação para que os LLMs possam não apenas gerar perfis, mas também completá-los por meio de inferência semântica e generalização, lidando com lacunas de informação de forma mais flexível e inteligente.

O trabalho de Shi *et al.* (2024) propõe um sistema de recuperação e geração que incorpora mecanismos dinâmicos de personalização, chamado ERAGent, com base em dados históricos de interação e sinais de preferência do usuário. A arquitetura do ERAGent realiza o enriquecimento do *prompt* com informações derivadas do perfil do usuário, adaptando o comportamento do modelo de acordo com interações anteriores.

Já o trabalho de Zerhoudi e Granitzer (2024) introduz o sistema PersonaRAG, que propõe uma arquitetura com agentes centrados no usuário, baseando-se em vetores de perfil que influenciam tanto a recuperação quanto a geração de conteúdo. A personalização se dá por meio

de *embeddings* que representam atributos do usuário. No entanto, o modelo depende de um vetor vetorial fixo para caracterizar o usuário, e a adaptação textual é conduzida por meio de filtros e ajustes indiretos sobre o conteúdo, não necessariamente instruindo o modelo a simular a persona no sentido clássico.

O PersonaReAct compartilha com os trabalhos desta seção a premissa de que perfis de usuários enriquecidos são fundamentais para gerar respostas mais alinhadas com as necessidades e expectativas individuais. No entanto, o PersonaReAct avança ao introduzir um agente especializado de personalização final, que atua exclusivamente na etapa de formatação da resposta gerada, após a fase de raciocínio e ação ter sido concluída.

Diferentemente de abordagens como ONCE ou LLM-REC, que utilizam modelos de linguagem para enriquecer preferências ou gerar representações latentes de perfil, o PersonaReAct se baseia no paradigma ReAct para conduzir o raciocínio e a recuperação das informações relevantes, delegando ao Agente de Personalização a tarefa de adaptar a resposta ao perfil do usuário. Essa personalização final considera o contexto completo da interação, filtrando as informações mais relevantes e ajustando o tom e a linguagem conforme o papel funcional (como torcedor, treinador ou jornalista), sem modificar o fluxo decisório da cadeia de raciocínio.

Enquanto abordagens como o Persona-DB realizam refinamentos colaborativos com base em perfis previamente estruturados ou agrupamentos de usuários similares, o PersonaReAct adota uma abordagem distinta: o enriquecimento do perfil é conduzido diretamente pela LLM, que infere, em tempo de execução, as informações mais relevantes a partir do perfil fornecido. Em vez de depender de um banco externo de preferências ou regras fixas, o próprio LLM analisa o contexto da pergunta e o papel do usuário para destacar, adaptar e priorizar conteúdos pertinentes ao perfil informado. Essa estratégia resulta em uma personalização dinâmica, que afeta tanto o conteúdo quanto a forma da resposta gerada.

### **3.4 Refinamento de Prompts**

O refinamento de *prompts* representa uma estratégia de personalização baseada na adaptação iterativa e direcionada das instruções fornecidas aos modelos de linguagem. Ao invés de simplesmente injetar contexto ou perfis, essa abordagem propõe melhorar a qualidade da geração textual personalizando o formato, a estrutura ou a semântica dos *prompts* com base no comportamento anterior do modelo ou em objetivos específicos do usuário. Assim, o refinamento atua como uma camada de otimização comunicativa entre o usuário e o LLM.

Li *et al.* (2024) propõem um modelo que aprende a reescrever *prompts* para alinhar melhor as gerações do modelo com as preferências do usuário. Utilizando uma rede neural supervisionada com *feedback* de preferências, o sistema ajusta o *prompt* original em tempo real para gerar saídas mais personalizadas, mostrando que a performance melhora mesmo sem ajuste nos pesos do LLM.

Salemi *et al.* (2024) introduzem um pipeline, chamado LAMP, que inclui tanto o enriquecimento de contexto quanto a reescrita dinâmica dos *prompts* com base em atributos de usuários. O diferencial da abordagem LAMP está em sua arquitetura modular, que permite refinar a entrada com base em atributos explícitos ou latentes, como estilo de escrita, tom emocional ou nível de formalidade esperado.

Kim e Yang (2025), por sua vez, abordam o desafio da personalização em cenários de *few-shot learning*, onde há desalinhamento entre as respostas geradas e os exemplos fornecidos. Eles demonstram que o refinamento estratégico do *prompt*, por meio da seleção ou reformulação dos exemplos, melhora substancialmente a adequação das respostas aos perfis esperados.

Além desses trabalhos voltados à personalização direta, Santurkar *et al.* (2023) e Durmus *et al.* (2024) revelam que os LLMs tendem a refletir visões de mundo enviesadas, muitas vezes alinhadas a certos padrões socioculturais. Isso destaca a importância de intervenções explícitas via *prompt engineering* para corrigir ou balancear essas representações (um tipo de refinamento voltado à imparcialidade e alinhamento com múltiplas perspectivas subjetivas).

O modelo T5 (Raffel *et al.*, 2020), apesar de não focar exclusivamente em personalização, reforça a ideia de que um paradigma unificado baseado em *Text-to-Text prompting* favorece reescritas mais sistemáticas, o que serve de base conceitual para muitas estratégias de refinamento aplicadas em trabalhos posteriores.

O PersonaReAct adota uma estratégia distinta das abordagens tradicionais de refinamento de *prompts*, ao delegar a personalização da resposta final a um agente especializado, cuja função é adaptar a linguagem e o conteúdo da resposta ao perfil previamente informado pelo usuário. Esse agente atua apenas após a conclusão do raciocínio e das ações realizadas no ciclo ReAct, e não interfere nas fases intermediárias de pensamento ou execução.

Diferentemente de sistemas como LAMP ou métodos de reescrita automática de *prompts* com aprendizado supervisionado ou reforçado, o PersonaReAct não realiza ajustes nos *prompts* iniciais nem busca corrigir trajetórias incorretas. Sua personalização ocorre exclusivamente na resposta final, com base no contexto completo da interação, utilizando o raciocínio

anterior e inferências realizadas internamente pelo modelo de linguagem para identificar quais informações são mais relevantes para o perfil funcional do usuário (como treinador, torcedor ou jornalista) e adaptar o tom da mensagem conforme esse papel.

Além disso, enquanto a maioria dos trabalhos dessa linha se concentra na personalização textual baseada em estilo ou preferências subjetivas, o PersonaReAct foca na adaptação semântica e comunicacional, moldando a resposta final de forma que ela seja não apenas coerente com o conteúdo, mas também adequada ao propósito funcional do usuário. Essa personalização inclui mudanças no vocabulário, na ênfase e na organização da informação, promovendo maior utilidade e alinhamento com as expectativas típicas de cada perfil.

### 3.5 Análise Comparativa dos Trabalhos

Ao longo deste capítulo, foram discutidas diferentes estratégias consolidadas de personalização aplicadas a modelos de linguagem, como os *prompts* contextuais, os *prompts* baseados em personas, os *prompts* com enriquecimento por perfil e o refinamento iterativo de *prompts*. Diversos trabalhos representativos de cada uma dessas categorias foram mencionados, destacando algumas de suas contribuições.

A presente seção complementa essa análise por meio de uma sistematização comparativa. O Quadro 2 a seguir não busca incluir todos os trabalhos discutidos anteriormente, mas apresenta um subconjunto representativo, selecionado por sua relevância e diversidade de soluções técnicas. A tabela resume os principais elementos estruturais e metodológicos de cada abordagem, permitindo uma visão panorâmica de como essas estratégias têm sido implementadas na prática recente.

Cada coluna do Quadro 2 representa um elemento-chave para análise comparativa:

- **Objetivo:** Finalidade do trabalho (ex.: melhorar recomendação, personalização de respostas).
- **Tipo de Personalização:** Nível da personalização (ex.: individual, por perfil funcional).
- **Fonte de Dados:** Tipo de dados utilizados (ex.: histórico de interações, descrições de itens, perfil aprendido).
- **LLM usada:** Modelos de linguagem utilizados.
- **Técnica de Personalização:** Estratégias adotadas para personalizar a saída.
- **Avaliação:** Métricas de desempenho utilizadas.
- **Cold Start?:** Indica se o método lida com a ausência de histórico do usuário.

- **Arquitetura:** Estrutura geral do sistema (ex.: modular, multiagente).
- **Recuperação:** Técnicas de recuperação utilizadas (ex.: vetorial, textual, *Text-to-SQL*).
- **Domínio:** Área de aplicação (ex.: recomendação, Q&A, mídia, esporte).

O PersonaReAct, apresentado na última linha do quadro, propõe uma arquitetura multiagente que integra personalização baseada em perfil como etapa final, após a geração do raciocínio e das ações. Diferentemente das abordagens que injetam informações do perfil já no início do processo, o PersonaReAct preserva a lógica da interação e aplica a personalização apenas no momento da resposta, o que permite adaptar linguagem, foco e estilo ao papel funcional do usuário (como técnico, torcedor ou analista), mantendo a consistência e a rastreabilidade do raciocínio.

Quadro 2 – Comparativo entre trabalhos recentes segundo características estruturais e de personalização.

Trabalho	Objetivo	Tipo de Personalização	Fonte de Dados	LLM usada	Técnica de Personalização	Avaliação	Cold Start?	Arquitetura	Recuperação	Domínio
Palma (2023)	Melhorar sistemas de recomendação	Nível de Usuário	Histórico de Interações do Usuário	GPT-3.5	<i>Prompts</i> Contextuais	nDCG; MAP; HR;	Não	<i>framework</i> Modular	Baseados em Filtragem de Dados	Recomendação de Filmes e Livros
Liu <i>et al.</i> (2024b)	Melhorar sistemas de recomendação	Nível de Usuário	Conteúdo dos Itens; Histórico de Navegação do Usuário	GPT-3.5; LLaMa-7B; LLaMA-13B	<i>Prompts</i> com Enriquecimento por Perfil	AUC; MRR; nDCG	Sim	<i>framework</i> Modular	Não mencionado	Recomendação de Notícias e Livros
Lyu <i>et al.</i> (2024)	Melhorar sistemas de recomendação	Nível de Usuário	Descrições de itens; Histórico de Interações do Usuário com Itens	GPT-3; LLaMa-2	<i>Prompts</i> Contextuais	Precision@K; Recall@K; NDCG@K	Não	<i>framework</i> Modular	<i>Personalized PageRank</i>	Recomendação de Filmes e Receitas
Sun <i>et al.</i> (2025)	Melhorar personalização de LLMs	Nível de Usuário	Histórico de Interações do Usuário	GPT-3.5	<i>Prompts</i> com Enriquecimento por Perfil	Acurácia; F1 Score; MSE; Análise Qualitativa	Sim	<i>framework</i> Modular	Vetorial	Previsão de Respostas em Mídias de Notícias; P&R
Kim e Yang (2025)	Melhorar personalização de LLMs	Nível de Usuário	Perfil do Usuário; Opiniões do Usuário	GPT-3.5; GPT-4; Mistral-7B; LLaMa-70B	Refinamento de <i>Prompts</i>	Acurácia; MAE; ROUGE-L	Sim	<i>framework</i> Modular	<i>Retrieval-of-Prompt</i>	P&R; Classificação; Regressão; Geração de Texto
Zerhoubi e Granitzer (2024)	Personalizar e Adaptar Respostas	Nível de Usuário	Histórico de Interações; Preferências do Usuário	GPT-3.5; LLaMa-3; MoE (8x7b)	<i>Prompts</i> Contextuais; <i>Prompts</i> com Enriquecimento por Perfil	Acurácia; BLEU-2; Tamanho da Sentença; Tamanho Médio das Sílabas	Não	Arquitetura Multiagente	BM25	Q&A em domínios diversos
Shi <i>et al.</i> (2024)	Personalizar e Adaptar Respostas	Nível de Usuário	Histórico de Interações; Perfil Aprendido	GPT-3.5	<i>Prompts</i> com Enriquecimento por Perfil	<i>Exact Match</i> (EM); Precisão; Recall; Hit Rate; <i>LLM-as-a-Judge</i>	Não	<i>Framework</i> Modular	BM25	Q&A no domínio médico
<b>PersonaReAct</b>	<b>Personalizar e Adaptar Respostas</b>	<b>Nível de Perfil</b>	<b>Perfil do Usuário</b>	o3-mini, o4-mini, GPT-5 mini, GPT-5 nano, GPT-4o mini, GPT-4.1 mini e GPT-4.1 nano	<b><i>Prompts</i> Contextuais; <i>Prompts</i> com Enriquecimento por Perfil</b>	<b>ROUGE-L; BLEU; <i>LLM-as-a-Judge</i></b>	<b>Sim</b>	<b>Arquitetura Multiagente</b>	<b><i>Text-to-SQL</i></b>	<b>Q&amp;A no domínio esportivo e médico</b>

Fonte: elaborado pelo autor.

## 4 PERSONAREACT: INTEGRAÇÃO DE RACIOCÍNIO, AÇÃO E PERSONALIZAÇÃO

O paradigma ReAct (*Reasoning + Acting*) (Yao *et al.*, 2023) propõe uma abordagem inovadora para agentes interagirem com ambientes complexos ao integrar explicitamente a capacidade de raciocínio em linguagem natural com a execução prática de ações. Embora eficaz, o paradigma ReAct não contempla a dimensão de personalização das respostas, o que limita sua aplicabilidade em cenários onde o contexto específico do usuário influencia significativamente na relevância das respostas fornecidas. Por exemplo, ao consultar dados sobre o desempenho de um jogador de futebol, um torcedor pode esperar uma resposta com ênfase em momentos decisivos e estatísticas de destaque. Já um fisioterapeuta busca informações mais técnicas, como carga de esforço e tempo de recuperação. Essas são demandas distintas que o ReAct original não é capaz de diferenciar.

Diante dessa limitação, o presente capítulo propõe uma especialização do paradigma ReAct, denominada PersonaReAct. O PersonaReAct mantém a integração original de raciocínio e ação proposta pelo paradigma ReAct, mas incorpora explicitamente um componente adicional de personalização das respostas, o módulo de adaptação final (*Tailor*). Este módulo adapta as informações recuperadas das ações executadas ao contexto específico do usuário, filtrando ou destacando dados relevantes e ajustando o tom da mensagem final.

Este capítulo está estruturado da seguinte forma: inicialmente, é apresentada uma fundamentação conceitual detalhada do paradigma ReAct original, evidenciando suas limitações e motivando a necessidade do PersonaReAct. Em seguida, é apresentada uma formulação matemática clara e detalhada da especialização proposta. Na sequência, a arquitetura geral do PersonaReAct é apresentada, explicitando seus componentes e a interação multiagente envolvida no processo. Os ciclos internos do *framework* (*Reason*, *Act* e o módulo de adaptação final *Tailor*) são então detalhados, seguidos por exemplos ilustrativos práticos que demonstram o funcionamento do PersonaReAct em diferentes contextos. Finalmente, são destacadas as diferenças e contribuições dessa proposta frente ao padrão ReAct original, e algumas considerações técnicas sobre sua implementação são discutidas.

### 4.1 Contextualização e Justificativa do Modelo

O paradigma ReAct original propõe integrar duas capacidades fundamentais dos agentes inteligentes:

- **Reason (Raciocínio)**: refere-se à habilidade do agente de formular pensamentos estruturados, em linguagem natural, com o objetivo de planejar, analisar ou justificar suas próximas ações;
- **Act (Ação)**: diz respeito à execução prática dessas decisões por meio de ações que interagem com o ambiente externo, geralmente por meio do uso de ferramentas ou APIs.

A ideia central é permitir que agentes baseados em modelos de Linguagem Natural (LN) realizem tarefas complexas por meio da alternância entre raciocinar e agir, intercalando deliberação interna com interações externas. Essa abordagem promove maior interpretabilidade, flexibilidade e robustez nas decisões do agente.

Para dar suporte à integração entre pensamento e execução, o paradigma ReAct propõe que agentes não se limitem apenas à realização de ações externas no ambiente, mas também expressem explicitamente seus raciocínios em linguagem natural. Dessa forma, o agente alterna entre refletir sobre o problema e interagir com o ambiente, estruturando sua tomada de decisão como uma sequência intercalada de pensamentos e ações.

A seguir, são apresentadas as definições formais dos elementos que compõem essa dinâmica, incluindo observações, ações, contexto e política de decisão.

**Definição 4.1.1 (Espaço de observações)** *O espaço de observações, denotado por  $O$ , é o conjunto de todas as observações possíveis que o agente pode receber a partir da interação com o ambiente. Cada observação representa uma evidência concreta sobre o estado do ambiente em um instante específico. Um elemento  $o_t \in O$  é denominado observação e corresponde à informação percebida pelo agente no tempo  $t$ .*

**Definição 4.1.2 (Espaço de ações externas)** *O espaço de ações externas, denotado por  $A$ , é o conjunto de todas as ações que o agente pode executar com efeitos observáveis no ambiente. Cada elemento  $a_t \in A$  é denominado ação externa e representa uma operação realizada pelo agente no instante  $t$ , cuja execução pode alterar o estado do ambiente e gerar uma nova observação.*

**Definição 4.1.3 (Espaço de linguagem)** *O espaço de linguagem, denotado por  $L$ , é o conjunto de todas as ações internas que o agente pode emitir utilizando linguagem natural. Essas ações não modificam diretamente o ambiente externo, mas têm a função de explicitar o processo deliberativo do agente. Cada elemento  $l_t \in L$  é denominado traço de raciocínio, sendo utilizado*

para registrar hipóteses, estruturar planos ou organizar pensamentos ao longo da resolução da tarefa.

**Definição 4.1.4 (Espaço total de ações)** O espaço total de ações, denotado por  $\hat{A}$ , é definido como a união entre o espaço de ações externas e o espaço de linguagem:

$$\hat{A} = A \cup L.$$

Cada elemento  $\hat{a}_t \in \hat{A}$  representa uma ação que pode ser tanto externa ( $a_t \in A$ ) quanto interna ( $l_t \in L$ ), e corresponde à decisão tomada pelo agente no instante  $t$ .

**Definição 4.1.5 (Contexto)** Seja  $c_t$  o contexto do agente no instante  $t$ , definido como a sequência ordenada de eventos até o momento atual:

$$c_t = (o_1, \hat{a}_1, o_2, \hat{a}_2, \dots, o_t),$$

onde  $o_i \in O$  são observações e  $\hat{a}_i \in \hat{A}$  são ações (externas ou traços de raciocínio). O contexto representa o histórico acumulado de percepções e decisões que guiam o comportamento futuro do agente.

**Definição 4.1.6 (Política)** A política de decisão, denotada por  $\pi$ , é uma função

$$\pi : C \rightarrow \hat{A}$$

que associa a cada contexto  $c_t \in C$  uma ação  $\hat{a}_t \in \hat{A}$ . Essa política orienta o agente sobre qual ação tomar com base no histórico de interações acumulado, podendo ser determinística ou estocástica.

Com base nessas definições, é possível compreender como um agente operando sob o paradigma ReAct alterna entre ações externas e traços de raciocínio ao longo do tempo. A cada instante  $t$ , o agente recebe uma observação  $o_t \in O$ , atualiza seu contexto  $c_t$ , e escolhe uma ação  $\hat{a}_t \in \hat{A}$ , que pode ser tanto uma ação externa  $a_t \in A$ , quanto um traço de raciocínio  $l_t \in L$ , conforme orientado pela política  $\pi$ .

O contexto é continuamente atualizado para incorporar as observações e ações realizadas, permitindo que o agente desenvolva estratégias mais robustas para lidar com tarefas complexas, como decompor objetivos, adaptar planos ou reagir a situações inesperadas.

A seguir, apresenta-se um exemplo prático que ilustra esse processo de forma detalhada.

### Exemplo ilustrativo do funcionamento do ReAct:

Considere um agente de linguagem treinado para auxiliar em tarefas domésticas. Suponha que ele receba a seguinte pergunta:

*Usuário: “Onde está o sal?”*

Assuma que o contexto inicial do agente está vazio. A sequência abaixo descreve a evolução do processo, com base no paradigma ReAct:

1. **Observação**  $o_1 \in O$ : o agente recebe a pergunta do usuário, que constitui sua primeira observação.
2. **Contexto**  $c_1$ : atualizado para incluir  $o_1$ , resultando em  $c_1 = (o_1)$ .
3. **Traço de raciocínio**  $l_1 \in L$ : o agente decide raciocinar antes de agir e emite o pensamento: “O sal geralmente fica na prateleira da cozinha.”. Essa ação interna é registrada no contexto.
4. **Contexto**  $c_2$ : atualizado para  $c_2 = (o_1, l_1)$ .
5. **Ação externa**  $a_2 \in A$ : com base no novo contexto, o agente executa uma ação que aciona uma ferramenta de busca visual (por exemplo, uma câmera na cozinha).
6. **Observação**  $o_2 \in O$ : a ferramenta retorna a informação “Sal não encontrado na prateleira.”.
7. **Contexto**  $c_3$ : atualizado com os novos eventos:  $c_3 = (o_1, l_1, a_2, o_2)$ .
8. **Traço de raciocínio**  $l_3 \in L$ : o agente reflete: “Talvez o sal esteja dentro do armário.”. Esse novo pensamento é registrado.
9. **Ação externa**  $a_4 \in A$ : o agente decide abrir o armário, onde finalmente o sal é encontrado.

Esse exemplo demonstra como o agente intercala traços de raciocínio ( $l_t \in L$ ) com ações externas ( $a_t \in A$ ), utilizando o contexto acumulado para tomar decisões progressivamente mais informadas. A inclusão explícita do raciocínio no fluxo de interação permite que o processo seja não apenas mais eficaz, mas também mais interpretável.

Embora o paradigma ReAct represente um avanço significativo ao combinar raciocínio e ação no processo decisório dos agentes, ele não considera explicitamente as particularidades individuais dos usuários com os quais interage. Todas as decisões são tomadas com base apenas no contexto da tarefa, sem levar em conta aspectos como preferências, perfil técnico ou objetivos específicos de um usuário.

Essa limitação restringe a aplicabilidade do modelo em cenários que exigem personalização, como sistemas educacionais, suporte especializado, assistência médica ou aplicações esportivas, nos quais diferentes perfis de usuários demandam diferentes formas de resposta e interpretação.

Diante disso, o PersonaReAct surge como uma especialização conceitual e prática do paradigma original, ao incorporar de forma explícita uma dimensão de personalização baseada no perfil do usuário. Essa especialização permite que as ações e os traços de raciocínio gerados pelos agentes sejam modulados conforme as características individuais e contextuais do perfil de quem interage, resultando em respostas finais adaptadas ao perfil, tanto no tom e estilo de comunicação quanto no conteúdo informacional propriamente dito, que pode variar de acordo com as necessidades e expectativas de cada perfil.

## 4.2 Modelo Conceitual e Formalização

Diferentemente do paradigma original ReAct, que toma decisões unicamente com base no contexto da tarefa, o PersonaReAct incorpora de forma explícita o perfil do usuário como variável central. Essa personalização não altera o processo de decisão interna do agente, mas atua na etapa final, permitindo que a resposta gerada seja adaptada ao perfil do usuário, tanto em termos de linguagem e estilo quanto no foco das informações apresentadas.

Considere um agente operando em um ambiente de perguntas e respostas (*question answering*) no qual é necessário adaptar a resposta final conforme o perfil do usuário.

**Definição 4.2.1 (Perfil de Usuário)** *O perfil de usuário é representado por  $p \in P$ , onde  $P$  é o espaço de possíveis perfis. Cada perfil contém informações contextuais, preferências, objetivos ou características que orientam a personalização das ações e respostas do agente.*

### **Ciclo de funcionamento do PersonaReAct**

O funcionamento do PersonaReAct ocorre em um ciclo iterativo composto por três fases principais:

$\underbrace{\text{Reason} \rightarrow \text{Act}}_{\text{repetição múltipla}} \rightarrow \text{Tailor}$

- **Reason:** o agente emite traços de raciocínio internos  $l_t \in L$ , atualizando seu contexto com pensamentos que estruturam planos ou hipóteses.
- **Act:** o agente executa ações externas  $a_t \in A$ , interagindo com o ambiente para obter novas observações  $o_t \in O$ .

- **Tailor**: após os ciclos de *Reason* e *Act*, o agente aplica um mecanismo de adaptação sobre as informações coletadas, gerando uma resposta final personalizada com base no perfil do usuário.

### **Função de Personalização (Tailor)**

O operador *Tailor* é formalizado como uma função de transformação:

$$\tau : P \times O^* \rightarrow M \quad (4.1)$$

onde:

- $P$  é o espaço de perfis de usuários;
- $O^*$  é a sequência de observações obtidas ao longo dos passos *Act*;
- $M$  é o espaço das mensagens finais geradas pelo agente, já adaptadas ao perfil informado.

A aplicação da função  $\tau(p, o_1, o_2, \dots, o_n)$  permite que a resposta final entregue pelo agente seja ajustada conforme o nível de detalhe, linguagem, conteúdo ou foco mais adequado ao perfil  $p$ , promovendo personalização sem comprometer a base lógica construída nos passos anteriores.

### **Exemplo ilustrativo do funcionamento do PersonaReAct com seleção de ações adaptadas ao perfil:**

Considere um agente de linguagem integrado a um sistema de manutenção predial assistida. O usuário solicita ajuda com o seguinte pedido:

*Usuário: “O ar-condicionado não está funcionando. O que eu faço?”*

O sistema, ao receber a requisição, identifica o perfil do usuário como:

- $p =$  “usuário leigo”: alguém sem conhecimento técnico, com permissão apenas para ações básicas.

A seguir, apresenta-se a sequência de interações do agente sob o paradigma Persona-ReAct:

1. **Observação**  $o_1 \in O$ : o agente recebe a pergunta. Inicializa  $c_1 = (o_1)$ .
2. **Traço de raciocínio**  $l_1$ : “Verificar se o aparelho está conectado à energia.”.
3. **Ação**  $a_2$ : consulta remota ao status da tomada inteligente.
4. **Observação**  $o_2$ : “Tomada ligada corretamente.”

5. **Traço de raciocínio**  $l_3$ : “Pode ser um problema no controle remoto.”.
6. **Ação**  $a_4$ : consulta à carga das pilhas no controle.
7. **Observação**  $o_3$ : “Nível de bateria baixo.”.
8. **Traço de raciocínio**  $l_5$ : “Verificar se há erros no condensador ou falhas internas.”.
9. **Ação**  $a_6$ : diagnóstico técnico da unidade externa via API de manutenção.
10. **Observação**  $o_4$ : “Erro E07: falha no sensor de temperatura.”.
11. **Contexto final**  $c_{t+1} = (o_1, l_1, a_2, o_2, l_3, a_4, o_3, l_5, a_6, o_4)$ .
12. **Personalização via**  $\tau(p, o_1, \dots, o_4)$ : o agente seleciona apenas as ações e observações relevantes ao perfil do usuário leigo.

**Resposta final gerada para o perfil “usuário leigo”:**

*“O ar-condicionado está ligado corretamente, mas o controle remoto parece com pouca bateria. Troque as pilhas e tente novamente. Se o problema continuar, chame um técnico.”.*

Caso o sistema, ao receber a requisição, identifique o perfil do usuário como:

- $p = \text{“técnico de manutenção”}$ : alguém com conhecimento técnico avançado, com permissão para todo tipo de ações relacionadas ao aparelho.

**Resposta final gerada para o perfil “técnico de manutenção”:**

*“A tomada está energizada e o controle remoto apresenta bateria fraca. Além disso, foi detectado o erro E07, indicando falha no sensor de temperatura do condensador. Verifique o sensor e substitua se necessário.”.*

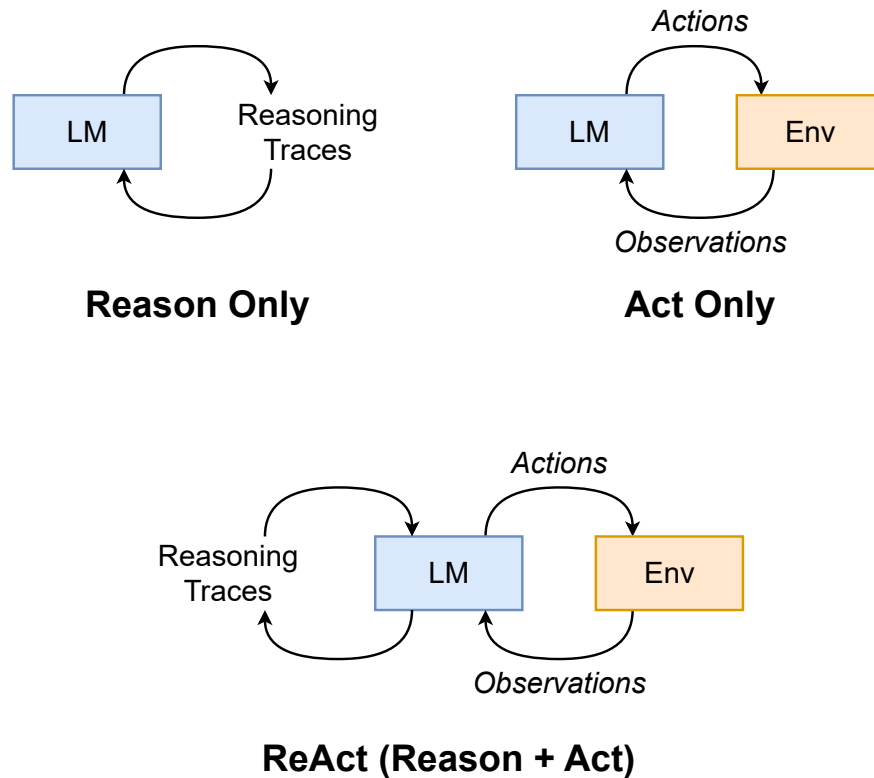
Este exemplo demonstra como o PersonaReAct estende o fluxo tradicional do ReAct ao incorporar o perfil do usuário na política de decisão e, principalmente, na personalização da resposta final. A função  $\tau$  atua como um filtro adaptativo que traduz os dados brutos em uma comunicação adequada ao público-alvo.

### 4.3 Arquitetura e Funcionamento Interno

A arquitetura do PersonaReAct foi concebida como uma especialização do paradigma ReAct, incorporando uma etapa adicional de personalização que permite adaptar as respostas ao perfil do usuário. Para compreender essa evolução, é útil observar a trajetória dos paradigmas

anteriores. A Figura 2 apresenta os três modelos que antecedem o PersonaReAct: *Reason Only*, *Act Only* e ReAct.

Figura 2 – Comparação entre os paradigmas *Reason Only*, *Act Only* e ReAct no uso de agentes.



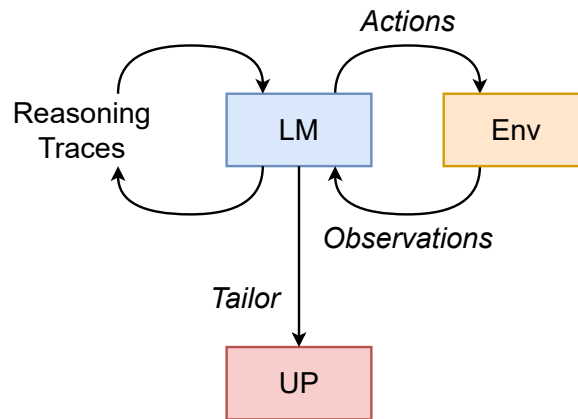
Fonte: adaptado de Yao *et al.* (2023).

Os modelos representados podem ser descritos da seguinte forma:

- ***Reason Only***: o modelo de linguagem gera apenas raciocínios em linguagem natural (*traces*), sem realizar ações externas;
- ***Act Only***: o modelo interage com o ambiente por meio de ações e observações, sem explicitar seu processo de raciocínio;
- ***ReAct (Reason + Act)***: integra as duas dimensões anteriores, promovendo raciocínio explícito e ações coordenadas, com maior transparência e capacidade de controle.

Avançando além do ReAct, propomos o PersonaReAct, mostrado na Figura 3, que mantém a estrutura de raciocínio e ação, mas incorpora um módulo dedicado à personalização da resposta com base em perfis de usuários.

Figura 3 – Componentes da arquitetura PersonaReAct com ênfase na personalização via perfil.



### PersonaReAct (Personalization + Reason + Act)

Fonte: elaborado pelo autor.

#### *Componentes principais da arquitetura*

O PersonaReAct é composto pelos seguintes blocos principais:

- **LM (*Language Model*)**: responsável por gerar raciocínios ( $l_t \in L$ ) e determinar ações ( $a_t \in A$ ), com base no contexto atual;
- **Ambiente (Env)**: responde com observações ( $o_t \in O$ ) de acordo com as ações executadas;
- **Contexto ( $c_t$ )**: memória temporária contendo a sequência de raciocínios, ações e observações anteriores;
- **Perfil do Usuário (UP)**: estrutura que representa características, preferências e estilo comunicacional de diferentes tipos de usuários;
- **Função de Personalização (*Tailor*)**: transforma o conteúdo gerado em uma resposta textual final, adaptada ao perfil indicado.

Essa arquitetura modular permite maior adaptabilidade e controle sobre as respostas, garantindo que o conteúdo final não apenas seja logicamente correto e relevante, mas também adequado ao perfil de quem interage com o sistema.

#### **4.4 Avanços em Relação ao Paradigma ReAct**

O PersonaReAct foi concebido como uma especialização estruturada do paradigma ReAct, mantendo seus princípios fundamentais, como a alternância entre raciocínio e ação,

mas incorporando um novo eixo de atuação: a personalização orientada por perfil de usuário. Essa seção resume as principais contribuições conceituais, funcionais e arquiteturais do modelo proposto.

### ***Integração explícita de perfis de usuário***

Diferentemente do ReAct tradicional, que executa ações e raciocínios com base unicamente no contexto da tarefa, o PersonaReAct introduz uma fase adicional de personalização ao final do processo, em que o perfil do usuário  $p \in P$  é incorporado para ajustar a resposta final. Essa adaptação permite que o mesmo conjunto de observações e raciocínios gere respostas distintas, mais adequadas a diferentes perfis de usuários, como leigos, especialistas, técnicos ou crianças, considerando suas necessidades, níveis de conhecimento ou interesses. Assim, a diferenciação de comportamento não ocorre nas decisões intermediárias, mas na forma como o conteúdo final é estruturado e apresentado.

### ***Separação entre informação coletada e informação comunicada***

Uma contribuição chave do PersonaReAct é a dissociação entre o que o agente observa e raciocina durante o ciclo *Reason–Act* e o que é efetivamente comunicado ao usuário ao final do processo. A função de personalização  $\tau$  atua como um filtro condicional aplicado pela própria LLM, que, com base no perfil funcional  $p$  fornecido, julga quais observações são mais relevantes ( $O^*$ ) para aquele perfil e devem compor a resposta final. Assim, ainda que as ações e raciocínios internos do agente sejam idênticos para diferentes usuários, a resposta comunicada é adaptada de forma sensível ao papel de quem pergunta.

### ***Generalização do ciclo Reason–Act com pós-processamento adaptativo***

O modelo PersonaReAct generaliza o ciclo *Reason–Act* ao adicionar uma etapa final de pós-processamento (*Tailor*), que sistematiza a personalização com base em função explícita. Essa etapa não é um mero ajuste de linguagem, mas pode influenciar a forma, o conteúdo, o nível de detalhamento e até a ordenação das informações apresentadas.

### ***Compatibilidade retroativa com o ReAct***

Apesar de seu novo módulo, o PersonaReAct mantém compatibilidade conceitual com o ReAct original. Se nenhum perfil for fornecido ou considerado, o modelo pode ser simplificado naturalmente para o ReAct tradicional, operando unicamente com base no contexto  $c_t$ . Essa característica facilita a transição entre os paradigmas e permite aplicações híbridas.

### ***Síntese Comparativa***

Ao longo deste capítulo, apresentamos os componentes, o funcionamento e a motivação para o PersonaReAct como uma especialização do paradigma ReAct, com ênfase em mecanismos de personalização textual. Para consolidar as diferenças fundamentais entre os dois modelos, o Quadro 3 apresenta uma comparação direta entre suas principais características arquiteturais e funcionais.

Quadro 3 – Comparação entre ReAct e PersonaReAct

<b>Aspecto</b>	<b>ReAct</b>	<b>PersonaReAct</b>
Perfil do usuário	Não considerado	Considerado explicitamente
Resposta final	Baseada apenas em contexto e observações	Adaptada conforme o perfil do usuário
Seleção de conteúdo	Fixa (utiliza o histórico completo ou parcial)	Condicional, via função $\tau(p, O^*)$ que considera o perfil
Compatibilidade retroativa	Não aplicável	Sim (funciona como ReAct se nenhum perfil for informado)

Fonte: elaborado pelo autor.

Essa comparação ressalta como o PersonaReAct amplia as capacidades do ReAct ao incluir, de forma explícita, o perfil do usuário como variável de entrada no processo decisório. Isso permite não apenas respostas mais alinhadas às expectativas do interlocutor, mas também maior controle sobre o estilo, a profundidade e a relevância da comunicação final.

## 5 METODOLOGIA

Este capítulo apresenta a metodologia adotada para avaliar a proposta do PersonaReAct, por meio de dois estudos de caso aplicados a um sistema de perguntas e respostas no domínio do futebol e a um sistema de perguntas e respostas no domínio médico. Todas as etapas metodológicas, desde a preparação dos dados até a avaliação experimental, foram estruturadas para verificar a eficácia da personalização baseada em perfis de usuários nesses contextos práticos, garantindo controle e reprodutibilidade na análise comparativa com o padrão ReAct.

A metodologia está organizada em seis seções. Inicialmente, são apresentados os processos de coleta e preparação dos dados, incluindo a construção da base relacional, definição dos perfis funcionais e elaboração das perguntas. Em seguida, detalha-se a implementação dos agentes ReAct e PersonaReAct, a criação dos *prompts* de personalização e os mecanismos utilizados para avaliação automatizada das respostas, com apoio de métricas tradicionais e do paradigma *LLM-as-a-Judge*. Por fim, também são descritos o desenho experimental e as principais limitações metodológicas dos estudos de caso.

Para facilitar a compreensão do fluxo metodológico, a Figura 4 apresenta uma visão geral dividida em três etapas principais. A primeira etapa, *Coleta de Dados e Geração de Perguntas*, envolve a criação de uma base de dados relacional alimentada por trajetórias de jogadores, a definição de perfis funcionais (como técnico, jornalista, torcedor, etc.) e a geração de perguntas personalizadas com base nesses perfis. A segunda etapa, *Implementação do Sistema*, contempla a construção dos agentes baseados nos paradigmas ReAct e PersonaReAct, incluindo a definição dos fluxos de raciocínio e escrita dos *prompts* específicos para cada abordagem. Por fim, a terceira etapa, *Avaliação Experimental*, compreende a formulação de critérios avaliativos (como coerência, linguagem e alinhamento com o perfil), a análise das respostas geradas e a aplicação de duas estratégias de avaliação: métricas tradicionais e a abordagem *LLM-as-a-Judge*.

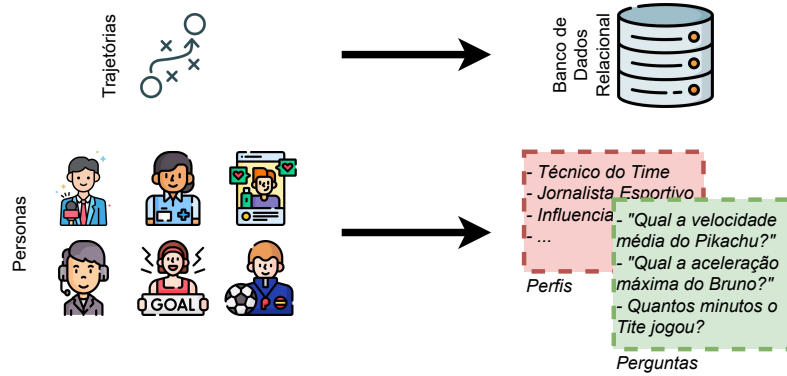
### 5.1 Coleta e Preparação dos Dados

#### *Estudo de Caso 1: Domínio Esportivo*

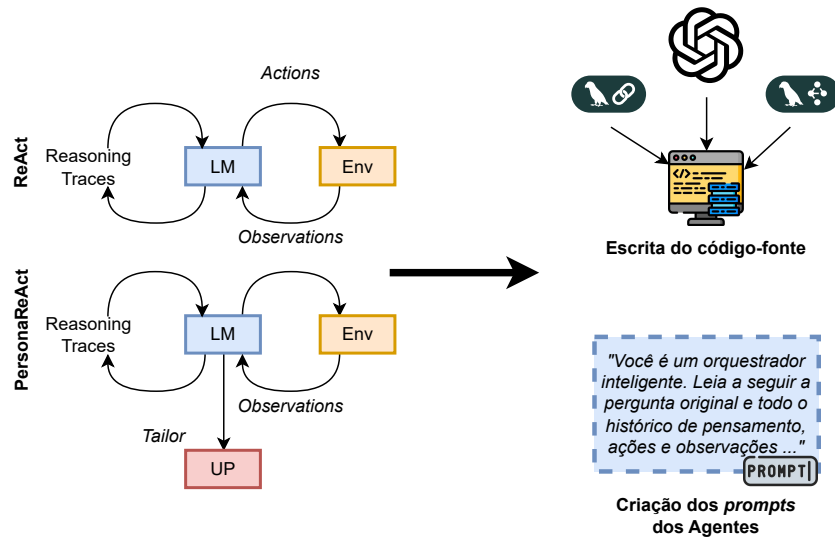
Os dados utilizados para este estudo de caso referem-se às informações de desempenho físico e fisiológico de todos os jogadores do Fortaleza Esporte Clube durante os jogos

Figura 4 – Visão geral do processo metodológico da tese, da coleta à avaliação dos agentes.

**1. Coleta de Dados e Geração de Perguntas**



**2. Implementação do Sistema**

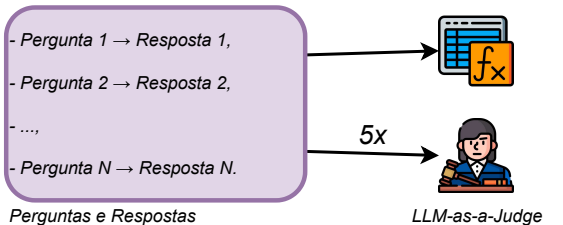


**3. Avaliação Experimental**

**3.1. Criação dos Critérios**

- Linguagem e Estilo Apropriados,
  - Interpretação e Estilo Apropriados,
  - Coerência do Estilo ao Longo do Texto.
- Critérios

**3.2. Avaliação das Respostas**



Fonte: elaborado pelo autor.

realizados ao longo da temporada de 2023. Esses dados foram obtidos graças a uma parceria formal com o clube, que autorizou o uso dos registros capturados pelos dispositivos de monitoramento utilizados pelos atletas em cada partida oficial. Por se tratarem de dados sensíveis relacionados à saúde e desempenho dos jogadores, o conjunto completo de dados não pode ser disponibilizado publicamente.

A coleta foi realizada por meio do sensor Polar Team Pro<sup>1</sup>, um dispositivo vestível que cada jogador utilizou durante os jogos. Esse sistema fornece registros contínuos de localização e métricas fisiológicas, exportados em dois formatos principais: arquivos no formato CSV, contendo informações agregadas sobre o tempo, duração das sessões e esforço cardíaco; e arquivos no formato GPX, que registram a trajetória georreferenciada de cada jogador ao longo de toda a partida. A lista completa dos atributos extraídos desses dispositivos encontra-se apresentada no Apêndice A.

O arquivo consolidado com todas as trajetórias da temporada possui aproximadamente 10 milhões de registros, onde cada linha representa um ponto da trajetória (com latitude e longitude no sistema WGS84) capturado com frequência de 1Hz (um ponto por segundo). Essa granularidade permite a reconstrução detalhada da movimentação dos atletas, viabilizando análises de padrão espacial e variações de posicionamento ao longo do tempo.

Os dados obtidos por meio do dispositivo Polar Team Pro, exportados em formato CSV, forneceram atributos fisiológicos e físicos agregados por jogador para cada jogo da temporada. Esses atributos representam médias ou totais referentes ao desempenho do atleta durante toda a partida, como a frequência cardíaca média, a frequência mínima ou a velocidade média em um jogo específico.

No entanto, para possibilitar análises mais granulares e contextualizadas ao longo da partida, foram utilizados também os dados de trajetória dos jogadores, obtidos a partir dos arquivos GPX. Com esses dados, foi possível recalculer diversos atributos, como aceleração, velocidade, distância e direção, em segmentos menores de trajetória compostos por 30 pontos consecutivos. Essa segmentação permitiu, por exemplo, identificar quantas acelerações um jogador realizou entre os minutos 30 e 35 de uma partida específica ou avaliar padrões de movimentação em trechos específicos do jogo.

Essas análises foram realizadas com o auxílio de bibliotecas consolidadas da linguagem Python voltadas à ciência de dados, como Pandas, NumPy, SciPy, Scikit-learn e Seaborn.

---

<sup>1</sup> <https://support.polar.com/br/polar-team-pro>

Os atributos extraídos foram então organizados e armazenados em um banco de dados relacional SQLite, que passou a compor a principal fonte estruturada de dados utilizada pelo sistema proposto. Tal organização permitiu a implementação de técnicas de *Text-to-SQL* no contexto de RAG, possibilitando ao modelo de linguagem natural gerar consultas automáticas em SQL a partir das perguntas dos usuários. Além disso, o sistema foi projetado para levar em consideração o perfil funcional de quem realiza a consulta, personalizando as respostas conforme o papel atribuído ao usuário, como torcedor, treinador ou fisioterapeuta.

Além da construção da base relacional com os atributos refinados, foi necessário estruturar um conjunto de perguntas que permitissem avaliar a capacidade do sistema em gerar respostas relevantes e personalizadas com base nos perfis dos usuários. Para isso, foram inicialmente elaboradas 20 perguntas em tom neutro, todas projetadas para serem plenamente respondidas a partir dos dados disponíveis no banco relacional construído. Em seguida, foram definidos seis perfis funcionais representativos do contexto do futebol: Torcedor do Time, Técnico do Time, Narrador de Futebol, Influenciador Digital, Fisioterapeuta do Time e Jornalista Esportivo. A partir da definição desses perfis, cada uma das 10 perguntas neutras foi adaptada para refletir o estilo, a linguagem e o foco informacional de cada perfil. Isso resultou em mais 120 perguntas, totalizando 140 questões (20 neutras e 120 adaptadas). Cada conjunto de perguntas foi codificado de P1 a P20, mantendo a correspondência entre a versão neutra e suas variações por perfil. Alguns exemplos das perguntas criadas e suas respectivas adaptações estão listadas no Apêndice D desta tese.

### ***Estudo de Caso 2: Domínio Médico***

Os dados utilizados neste estudo de caso foram obtidos a partir do repositório COVID-19 Data Sharing/BR<sup>2</sup>, uma iniciativa da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) em cooperação com a Universidade de São Paulo, com participação inicial do Instituto Fleury, Hospital Sírio-Libanês e Hospital Israelita Albert Einstein. O objetivo dessa iniciativa é disponibilizar dados relacionados à COVID-19 para fomentar pesquisas científicas sobre a doença (Mello *et al.*, 2020). Para este trabalho, foram utilizados exclusivamente os dados provenientes do Hospital Sírio-Libanês.

O conjunto de dados apresenta informações anonimizadas de pacientes que realizaram testes para COVID-19 (por sorologia ou RT-PCR) a partir de 1º de novembro de 2019.

<sup>2</sup> <https://repositoriodatasharingfapesp.uspdigital.usp.br/>

Ele inclui dados demográficos (gênero, ano de nascimento e região de residência), resultados de exames clínicos e laboratoriais, registros de movimentação do paciente (como internações) e, quando disponíveis, desfechos dos casos (como recuperação ou óbito).

Os arquivos originais estavam compactados em formato .zip e continham três arquivos .csv e um arquivo .xlsx: (i) Pacientes: identificador anonimizado, informações demográficas e localização de residência; (ii) Exames: resultados de exames laboratoriais, incluindo o identificador anonimizado do paciente e um identificador de atendimento; (iii) Desfechos: registros que descrevem cada atendimento, com o identificador anonimizado do paciente, identificador de atendimento e o respectivo desfecho.

As planilhas Pacientes, Exames e Desfechos podem ser interligadas pelo identificador anonimizado do paciente. Adicionalmente, as tabelas Exames e Desfechos são conectadas pelo par (identificador do paciente, identificador do atendimento). No total, o dataset inclui mais de 90 mil registros de desfechos, cerca de 3 milhões de resultados de exames e quase 15 mil pacientes.

Como os dados já estavam organizados em formato tabular, foi possível estruturá-los diretamente em um banco de dados relacional SQLite, preservando as chaves de ligação entre as tabelas e facilitando a execução de consultas complexas. Esse banco passou a ser a principal fonte de dados estruturados para a implementação e teste do sistema no domínio médico.

Assim como no estudo de caso do domínio esportivo, foi elaborado um conjunto de 20 perguntas em tom neutro, formuladas de modo a serem integralmente respondidas com base nos dados disponíveis no banco relacional. Foram definidos seis perfis funcionais representativos do contexto hospitalar e de saúde pública: Enfermeiro, Familiar de um Paciente, Gestor Hospitalar, Médico, Paciente e Pesquisador em Saúde Pública. Cada pergunta neutra foi adaptada para refletir o estilo, o vocabulário e o foco informacional característico de cada perfil, resultando em 120 variações adicionais. Dessa forma, o conjunto total atingiu 140 perguntas (20 neutras e 120 adaptadas), codificadas de P1 a P20 para manter a correspondência entre cada versão neutra e suas respectivas adaptações. Exemplos dessas perguntas e suas variações encontram-se listados no Apêndice D.

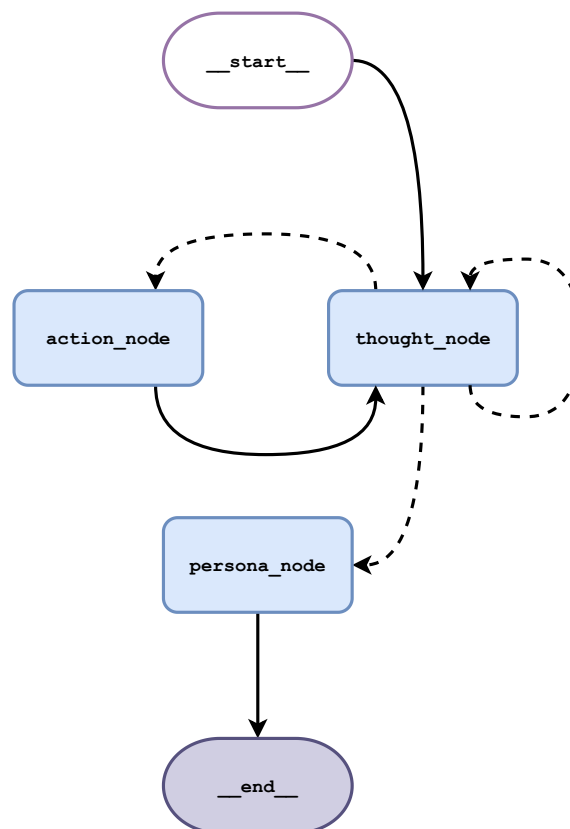
## 5.2 Arquitetura e Implementação dos padrões ReAct e PersonaReAct

Esta seção descreve a implementação da arquitetura do PersonaReAct. A implementação foi conduzida com o auxílio da biblioteca LangGraph, que possibilita a definição de fluxos

de execução com controle de estado em agentes baseados em LLMs. Todo o código desenvolvido está disponível em um repositório online criado especificamente para o PersonaReAct<sup>3</sup>.

A Figura 5 ilustra o grafo de execução do agente PersonaReAct, conforme gerado automaticamente pelo método `get_graph().draw_mermaid()` da biblioteca LangGraph. O fluxo é iniciado no nó `thought_node`, que tem como responsabilidade analisar a pergunta original e todo o histórico de pensamentos, observações e ações anteriores. Com base nessas informações, esse nó decide o próximo passo, podendo acionar uma ferramenta, emitir um novo pensamento ou finalizar a resposta.

Figura 5 – Fluxograma detalhado do funcionamento do PersonaReAct implementado com LangGraph.



Fonte: elaborado pelo autor.

Na maioria dos casos do estudo de caso, o `thought_node` ativa o `action_node`, responsável por executar consultas em uma base relacional a partir de um sistema de *Text-to-SQL*. Após recuperar as informações estruturadas necessárias, o controle retorna ao `thought_node`, que avalia se a resposta já é suficiente para concluir o raciocínio. Quando isso ocorre, o fluxo é transferido para o `persona_node`, responsável por aplicar a personalização da resposta conforme

<sup>3</sup> <https://github.com/gustavolgcr/personareact>

o perfil do usuário. O `persona_node` utiliza a sequência completa de observações acumuladas no contexto e adapta a resposta final em termos de linguagem, ênfase e conteúdo, de forma alinhada ao perfil funcional (por exemplo, Técnico do Time, Torcedor do Time, Fisioterapeuta do Time, etc.).

O fluxo é encerrado no nó especial `_end_`. A arquitetura também contempla casos em que, após o número máximo de iterações (limitado a 5 ciclos), o controle é redirecionado para o `persona_node`, mesmo que a resposta não tenha sido completamente construída. Nesse cenário, o sistema emite uma resposta padrão de falha, respeitando ainda assim o estilo linguístico correspondente ao perfil.

Todos os nós do grafo foram implementados utilizando a interface de `StateGraph` do `LangGraph`, sem extensões ou personalizações adicionais. Os modelos de linguagem utilizados foram invocados por meio da biblioteca `ChatOpenAI` do `LangChain`, com uso da versão `GPT-4o mini` da `OpenAI`, temperatura configurada em 1 e sem modificação de outros parâmetros.

O `thought_node` foi implementado com base na proposta original do `ReAct`, utilizando um *prompt* adaptado para este projeto. O modelo recebe a pergunta original e o histórico completo em formato textual, sendo instruído a decidir o próximo passo. As possíveis saídas são:

- um `thought` com raciocínio interno;
- um `action` com respectivo `action_input` textual, quando necessário acionar uma ferramenta;
- ou um `standard_answer`, que indica que a resposta já pode ser entregue ao usuário.

O `action_node` ativa a ferramenta de *Text-to-SQL* implementada no sistema, capaz de gerar e executar comandos `SQL` sobre o banco relacional contendo os atributos extraídos dos segmentos de trajetória dos jogadores.

Já o `persona_node` é o nó responsável pela transformação da resposta padrão (`standard_answer`) em uma versão personalizada. Esse nó recebe como entrada o perfil funcional do usuário (passado como variável no *prompt*) e a resposta obtida do modelo anteriormente.

O controle de fluxo do grafo garante que, mesmo em situações em que o modelo falha em retornar uma resposta ou excede o número máximo de passos, o sistema segue para o nó de personalização e finaliza a execução de forma controlada. Essa abordagem evita *loops* infinitos e preserva a coerência da resposta.

Todos os pensamentos, observações e decisões tomadas em cada etapa são armazenados no contexto global (*scratchpad*), permitindo ao `persona_node` ter acesso a uma visão

completa do raciocínio e das informações recuperadas. Esse mecanismo assegura que a personalização ocorra sobre uma base sólida de dados, sem comprometer a lógica construída nos passos anteriores.

### 5.3 Avaliação de Respostas: *LLMs-as-a-Judge* e Métricas Tradicionais

A avaliação das respostas geradas pelo PersonaReAct ao longo dos experimentos foi conduzida por meio de duas abordagens complementares: (i) o uso de modelos de linguagem como avaliadores automáticos (*LLMs-as-a-Judge*), e (ii) a aplicação de métricas tradicionais de similaridade textual e semântica. O objetivo foi garantir uma análise completa, unindo avaliação contextual e medição precisa das diferenças entre as respostas.

#### *Avaliação com LLMs-as-a-Judge*

Para avaliar qualitativamente a adequação das respostas geradas, foi adotada a abordagem automatizada baseada em modelos de linguagem como juízes, mais conhecida como *LLMs-as-a-Judge* (Gu *et al.*, 2025). O modelo utilizado foi o GPT-4o da OpenAI.

As avaliações foram conduzidas com o auxílio da biblioteca *G-Eval* (Liu *et al.*, 2023), um componente do *framework* DeepEval (Ip; Vongthongsri, 2025), que permite a execução estruturada de avaliações baseadas em critérios definidos pelo usuário. Em cada avaliação, o modelo recebe a pergunta feita, o perfil associado e a resposta gerada pelo sistema. Com base nessas informações, o LLM avaliador atribui notas a três critérios principais:

- **Linguagem e Estilo Adequados:** o quanto o estilo da resposta soa natural para o perfil informado, com palavras, tom e jeito de falar que combinem com esse tipo de perfil;
- **Interpretação e Explicação dos Dados:** o quanto a resposta interpreta corretamente os dados apresentados e fornece explicações úteis e contextualizadas para o perfil informado;
- **Coerência do Estilo ao Longo do Texto:** o quanto o estilo adotado é mantido de forma consistente ao longo da resposta, sem mudanças abruptas ou incoerentes no tom, vocabulário ou estrutura.

Cada critério recebeu uma nota em escala ordinal (entre 0 e 10), acompanhada de uma justificativa em linguagem natural fornecida pelo próprio modelo. Para minimizar variações entre execuções, cada avaliação foi repetida 5 vezes e a média das notas foi utilizada como resultado final. As rubricas detalhadas que orientaram o modelo na atribuição de notas, bem

como exemplos das justificativas geradas, encontram-se no Apêndice B e nos Quadros 8, 9, 10, 11, 12 e 13 do Capítulo 6.

A utilização de LLMs como juizes foi motivada não apenas pela sua capacidade de realizar julgamentos contextualizados e escaláveis, mas também pela dificuldade prática de reunir avaliadores humanos com experiência nos diferentes perfis definidos (Torcedor do Time, Técnico do Time, Pacientes, Médicos, etc.).

### ***Métricas Tradicionais de Similaridade***

Além da avaliação subjetiva por meio do *LLM-as-a-Judge*, também foram utilizadas métricas tradicionais de similaridade textual e semântica na Seção 5.5. As comparações foram sempre realizadas entre a resposta gerada para um perfil específico e a resposta padrão (neutra), tratando esta última como referência.

As seguintes métricas foram utilizadas:

- **Distância de Edição** (Levenshtein *et al.*, 1966): mede o número mínimo de operações (inserções, deleções ou substituições) necessárias para transformar uma *string* em outra. Seja  $s$  e  $t$  duas *strings* de tamanhos  $|s|$  e  $|t|$ , respectivamente. Define-se  $D(i, j)$  como a distância de edição entre os prefixos  $s_1^i$  e  $t_1^j$ , isto é, os primeiros  $i$  caracteres de  $s$  e os primeiros  $j$  caracteres de  $t$ . A função é definida recursivamente por:

$$D(i, j) = \begin{cases} \max(i, j) & \text{se } \min(i, j) = 0 \\ \min \begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \delta \end{cases} & \text{caso contrário} \end{cases}$$

onde  $\delta = 0$  se  $s_i = t_j$ , e  $\delta = 1$  caso contrário.

- **BLEU** (*Bilingual Evaluation Understudy*) (Papineni *et al.*, 2002): calcula a sobreposição de n-gramas entre a resposta gerada e a referência. Para uma resposta  $r$  e uma referência  $R$ , o score BLEU-n com penalidade de comprimento  $BP$  é dado por:

$$\text{BLEU} = BP \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right)$$

onde  $p_n$  é a precisão dos n-gramas e  $w_n$  são pesos (tipicamente uniformes).

O fator *BP* (*Brevity Penalty*) tem como objetivo penalizar respostas excessivamente curtas em relação à referência.

- **ROUGE-L** (Lin, 2004): é uma métrica baseada na maior subsequência comum (*Longest Common Subsequence* — LCS) entre duas sentenças. Seja  $L$  o comprimento da LCS,  $m$  o comprimento da resposta gerada e  $n$  o comprimento da resposta de referência. A métrica utiliza dois componentes: precisão (*Precision*), definida como  $\frac{L}{m}$ , que mede a proporção de palavras da resposta gerada que fazem parte da subsequência comum, e revocação (*Recall*), definida como  $\frac{L}{n}$ , que mede a proporção de palavras da referência recuperadas na subsequência comum. Esses valores são combinados em uma média harmônica ponderada, controlada pelo parâmetro  $\beta$ , que ajusta a importância relativa da revocação em relação à precisão. Assim, a métrica é dada por:

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot \text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \beta^2 \cdot \text{Revocação}}$$

Em resumo, o ROUGE-L avalia a similaridade textual considerando a ordem das palavras, sendo menos rígido que métricas baseadas em n-gramas exatos, pois captura a estrutura compartilhada entre os textos.

- **Distância Cosseno**: os vetores de *embeddings* das respostas foram calculados utilizando o modelo `all-mpnet-base-v2`, da biblioteca `sentence-transformers`. A similaridade semântica entre dois vetores  $u$  e  $v$  é avaliada por:

$$\cos(\theta) = \frac{u \cdot v}{\|u\| \cdot \|v\|}$$

sendo a distância definida como  $1 - \cos(\theta)$ .

Essas métricas possibilitaram analisar tanto o grau de variação textual entre as respostas quanto a preservação de similaridade semântica em casos de reformulações. A análise conjunta dos resultados permitiu identificar padrões de adaptação textual e aproximação conceitual entre os perfis, conforme descrito nos experimentos apresentados no Capítulo 6.

#### 5.4 Construção do *Prompt* de Personalização

Esta seção apresenta a construção do *prompt* utilizado pelo módulo de personalização do PersonaReAct, responsável por adaptar a resposta final de acordo com o perfil funcional do usuário. A formulação desse *prompt* seguiu diretrizes baseadas em boas práticas de *prompt*

*engineering* recomendadas pela OpenAI<sup>4</sup>, com o objetivo de garantir maior clareza, precisão e controle sobre a saída do modelo.

O conteúdo completo do *prompt* de personalização encontra-se no Apêndice C desta tese.

As principais boas práticas aplicadas à elaboração do *prompt* foram:

1. **Uso de delimitadores para indicar claramente as partes do input:** o *prompt* é segmentado em seções com marcadores explícitos, como *Etapa 1: Análise do Perfil*, *Etapa 2: Análise da Pergunta*, entre outras. Além disso, as variáveis {profile}, {question} e {history} são destacadas e contextualizadas dentro da estrutura do *prompt*.
2. **Pedido de saída estruturada:** o modelo é orientado a retornar exclusivamente a resposta final, com a instrução explícita “*Retorne apenas a resposta final personalizada.*”. Isso evita a inclusão de explicações intermediárias ou formatações indesejadas.
3. **Inclusão de exemplos de comportamento esperado:** logo na etapa de análise do perfil, são apresentados exemplos conceituais de adaptação de linguagem e estilo. Por exemplo: “*Torcedor do Time → Linguagem informal e empolgada*”.
4. **Divisão da tarefa em subtarefas simples:** a personalização da resposta é dividida em quatro etapas lógicas e independentes: análise do perfil, análise da pergunta, seleção de informações relevantes e geração da resposta final.
5. **Estimulação do raciocínio antes da resposta:** o *prompt* inclui instruções como “*reflita cuidadosamente sobre o perfil informado*” e “*simule mentalmente que você é o usuário descrito*” para induzir o modelo a adotar um comportamento reflexivo.
6. **Imposição de restrições explícitas de formatação:** são indicados casos em que dados quantitativos devem ser arredondados ou evitados, conforme o perfil. Por exemplo, evitar “732 segundos” e preferir “pouco mais de 12 minutos” para certos perfis.
7. **Verificação de suposições e naturalidade:** ao final, o modelo é instruído a revisar a resposta perguntando-se se ela “*soa natural para o perfil informado e se a personalização foi mantida ao longo do texto*”.
8. **Iteração e refinamento progressivo:** o *prompt* final é resultado de ciclos de teste e ajuste, aplicando respostas reais e *feedback* qualitativo como base para melhorias sucessivas na personalização.

Essa estrutura detalhada do *prompt* busca garantir que a resposta final esteja adequada

<sup>4</sup> <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api>

ao perfil informado, soando natural, útil e alinhada às expectativas de quem faz a pergunta.

## 5.5 Desenho Experimental

Esta seção descreve os experimentos planejados para testar as hipóteses apresentadas no Capítulo 1. Cada experimento está associado a uma hipótese específica e foi estruturado para ser executado com base em avaliações predominantemente quantitativas, reduzindo a necessidade de julgamento subjetivo por parte de avaliadores humanos. O Quadro 4 apresenta um breve resumo dos experimentos.

Quadro 4 – Visão geral dos experimentos.

Experimento	Condições e Configuração	Avaliação
<b>E1 — Eficácia e comparação entre abordagens de personalização</b>	<p>- <i>Abordagens</i>:</p> <ol style="list-style-type: none"> <li>1. <b>ReAct</b> (sem personalização),</li> <li>2. <b>ReAct + Prompt Enriquecido</b>,</li> <li>3. <b>PersonaReAct</b> (módulo <i>Tailor</i>).</li> </ol> <p>- <i>Domínios</i>: Esportivo (futebol) e Médico (COVID-19).  - <i>Perfis</i>: 6 por domínio.  - <i>Perguntas</i>: 20 neutras adaptadas por perfil (140 por domínio).  - <i>Modelos</i>: com raciocínio (o3-mini, o4-mini, GPT-5 mini, GPT-5 nano) e sem raciocínio (4o-mini, GPT-4.1 mini, GPT-4.1 nano).</p>	<p>- <i>LLM-as-a-Judge</i> (G-Eval) com o modelo GPT-4o como base e os três seguintes critérios de avaliação:</p> <ol style="list-style-type: none"> <li>1. <b>Linguagem e Estilo Adequados</b>,</li> <li>2. <b>Interpretação e Explicação dos Dados</b>,</li> <li>3. <b>Coerência do Estilo ao Longo do Texto</b>.</li> </ol>
<b>E2 — Variação entre perfis</b>	<p>- Mesmas perguntas (20 neutras) e perfis de E1, para o domínio esportivo apenas.  - Para cada pergunta: 1 resposta neutra (<i>Tailor</i> desativado) + 6 respostas personalizadas (<i>Tailor</i> ativado, uma por perfil).  Conjunto por pergunta: 7 respostas; total agregado por domínio compatível com E1.</p>	<p>- Métricas de similaridade textual entre resposta personalizada e padrão: ROUGE-L, BLEU e Edit Distance; análise qualitativa por perfil.</p>

Fonte: elaborado pelo autor.

### *Experimento 1 – Eficácia e Comparação entre Abordagens de Personalização*

Este experimento foi planejado para avaliar, de forma integrada, três hipóteses centrais desta tese. A primeira (H1) propõe que a personalização das respostas com base no perfil do usuário aumenta sua consistência, alinhamento e adequação comunicacional. A

segunda (H2) defende que a separação explícita do componente de personalização, tal como concebida no PersonaReAct, pode melhorar a qualidade e a organização das respostas em comparação a abordagens que utilizam apenas o ReAct com *prompt* enriquecido. Por fim, a quarta hipótese (H4) sugere que o tipo de modelo de linguagem utilizado (com ou sem capacidades avançadas de raciocínio) influencia diretamente o desempenho relativo das diferentes abordagens de personalização.

Para investigar essas hipóteses, foram comparadas três estratégias de geração de respostas: (i) ReAct Puro, sem qualquer mecanismo de personalização textual; (ii) ReAct usando *Prompt* Enriquecido, que insere o perfil do usuário diretamente no enunciado da pergunta; (iii) PersonaReAct, que trata a personalização como uma etapa explícita do fluxo, realizada por meio do módulo *Tailor*.

O desenho experimental foi aplicado em dois contextos distintos: o domínio esportivo, envolvendo futebol profissional, e o domínio médico, com dados hospitalares relacionados à COVID-19. Essa diversidade de cenários permitiu observar se os padrões de comportamento das abordagens se repetem em áreas com características comunicacionais e informacionais bastante diferentes.

### *Perfis de Usuário e Perguntas*

No estudo do domínio esportivo, os perfis foram definidos com base em personas amplamente reconhecidas no contexto do futebol. Cada perfil representa um modo distinto de consumir e valorizar a informação:

- **Torcedor do Time:** linguagem informal e emocional, com foco no esforço e na entrega dos jogadores;
- **Técnico do Time:** tom técnico e objetivo, voltado à análise tática e física;
- **Narrador de Jogo de Futebol:** tom vibrante, descritivo e visual, buscando dados que embasem a narração ou comentário;
- **Influenciador Digital:** estilo comunicativo e envolvente, com linguagem descontraída e foco em engajamento e curiosidades que gerem conteúdo;
- **Fisioterapeuta do Time:** abordagem técnica e objetiva, com ênfase em carga física, prevenção de lesões e acompanhamento do desempenho atlético;
- **Jornalista Esportivo:** estilo informativo, com foco em performance, narrativa esportiva e possíveis ângulos para matérias.

Já no domínio médico, os perfis foram escolhidos para refletir papéis comuns no ambiente hospitalar e em saúde pública, abrangendo desde a comunicação clínica até a análise epidemiológica:

- **Enfermeiro:** linguagem técnica, mas acessível, com foco no acompanhamento clínico, evolução do quadro e cuidados prestados ao paciente;
- **Familiar de um Paciente:** tom empático e acolhedor, priorizando informações claras sobre estado de saúde e expectativas de evolução, evitando jargões excessivos;
- **Gestor Hospitalar:** estilo objetivo e orientado a indicadores, com ênfase em métricas de ocupação, fluxo de atendimentos e eficiência operacional;
- **Médico:** linguagem técnica e precisa, com foco em interpretação de exames, condutas adotadas e evolução clínica;
- **Paciente:** tom direto e tranquilizador, com informações sobre resultados de exames e orientações de tratamento em linguagem simples;
- **Pesquisador em Saúde Pública:** abordagem analítica e estruturada, com destaque para padrões epidemiológicos e implicações para políticas de saúde.

Para cada perfil, foram inicialmente elaboradas 20 perguntas em tom neutro. Em seguida, essas perguntas foram adaptadas para refletir o estilo, o vocabulário e o foco informacional característico de cada persona. O processo resultou em 140 perguntas por domínio (20 neutras e 120 adaptadas), totalizando 280 perguntas nos dois estudos de caso.

### *Geração de Respostas*

Cada pergunta, associada ao seu perfil, foi processada pelas três abordagens em comparação:

1. **ReAct** (sem personalização): geração de resposta sem considerar qualquer informação sobre o perfil do usuário;
2. **ReAct usando Prompt Enriquecido:** inclusão de uma instrução ao final da pergunta original no formato “ Mude a resposta com base no perfil do usuário que fez a pergunta: ” + perfil;
3. **PersonaReAct:** aplicação da personalização como etapa separada no fluxo, executada pelo módulo *Tailor*.

O Quadro 5 ilustra, com um exemplo do domínio esportivo, como a personalização altera o tom e o conteúdo da resposta. O conjunto completo de perguntas e respostas, abrangendo

ambos os domínios, encontra-se disponível no repositório online do PersonaReAct.

Quadro 5 – Exemplo de pergunta feita para o perfil “Torcedor do Time” com uma resposta padrão e uma resposta personalizada com o PersonaReAct.

<b>Perfil do usuário</b>	<i>Torcedor do Time</i>
<b>Pergunta</b>	<i>Quem foi o cara que mais deu arrancadas durante o jogo contra o Ceará no começo de abril?</i>
<b>Resposta sem Personalização</b>	<i>O jogador que mais deu arrancadas durante o jogo contra o Ceará no começo de abril foi Tomas Pochettino, com uma média de 1.0176470588235293 arrancadas.</i>
<b>Resposta com Personalização</b>	<i>No jogo contra o Ceará no começo de abril, quem se destacou com suas arrancadas foi o nosso guerreiro, Tomas Pochettino! O cara fez bonito em campo, mostrando todo seu ímpeto e garra, com uma média de mais de uma arrancada por minuto. Realmente, ele se entregou demais nesse jogo, não teve pra ninguém! É isso que a gente gosta de ver: jogador dando o sangue pela vitória! Vamos em frente, com ele no time, a esperança é grande!</i>

Fonte: elaborado pelo autor.

### *Modelos Utilizados*

As respostas foram produzidas por sete modelos da OpenAI, organizados em duas categorias:

- **Com capacidades avançadas de raciocínio:** o3-mini, o4-mini, GPT-5 mini, GPT-5 nano;
- **Sem capacidades avançadas de raciocínio:** 4o-mini, GPT-4.1 mini, GPT-4.1 nano.

Cada pergunta foi respondida por todas as combinações de abordagem e modelo. Os modelos foram instanciados com `temperature = 1`, que controla o grau de aleatoriedade na geração de texto (valores mais altos promovem maior diversidade linguística). O parâmetro `top_p = 1.0` (*nucleus sampling*) define que todos os *tokens* com probabilidade acumulada de até 100% são considerados, não limitando a escolha a um subconjunto mais provável. Já `frequency_penalty = 0.0` indica que não há penalização para repetições de termos, permitindo que palavras frequentes possam ser reutilizadas sem restrição.

### *Avaliação Automática com LLM*

A análise da qualidade das respostas foi conduzida automaticamente por meio do *framework* G-Eval, utilizando como julgador o modelo 4o-mini da OpenAI. Três critérios orientaram a avaliação: *Linguagem e Estilo Apropriados*, *Interpretação e Explicação dos Dados* e *Coerência do Estilo ao Longo do Texto*, conforme detalhado na Seção 5.3.

### *Considerações e Limitações*

O delineamento experimental foi pensado para permitir análises comparativas por perfil, por domínio e por tipo de modelo. Assim, além de verificar a eficácia da personalização (H1), também foi possível avaliar o efeito da modularização (H2) e a influência das capacidades de raciocínio do modelo (H4).

Por se tratar de um volume expressivo de respostas e múltiplas combinações, optou-se pela avaliação automatizada com LLM. Embora escalável, essa abordagem pode refletir vieses inerentes ao modelo avaliador, motivo pelo qual as rubricas foram formuladas de forma detalhada para guiar a atribuição das notas e minimizar inconsistências.

### ***Experimento 2 – Variação entre Perfis***

Este experimento tem como objetivo testar a Hipótese 2 (H2), segundo a qual uma mesma pergunta, quando direcionada a diferentes perfis de usuário, deve resultar em respostas significativamente distintas. Parte-se do pressuposto de que o módulo de personalização do PersonaReAct não apenas ajusta o conteúdo, mas também modifica o estilo comunicativo da resposta, de forma coerente com o perfil informado, mesmo quando a informação solicitada é idêntica.

#### *Perfis de Usuário e Perguntas*

Foram utilizados exatamente os mesmos perfis definidos no Experimento 1, abrangendo tanto o domínio esportivo quanto o domínio médico.

O conjunto de perguntas foi construído da seguinte forma: inicialmente, foram elaboradas 20 perguntas em tom neutro, válidas para todos os perfis. Em seguida, cada pergunta foi adaptada para refletir exclusivamente o estilo e o tom característicos de cada perfil, preservando integralmente o conteúdo e a intenção da pergunta original. Dessa maneira, todos os perfis responderam, essencialmente, às mesmas questões, diferenciando-se apenas na forma como elas foram formuladas.

Essa abordagem garante consistência metodológica, permitindo que as diferenças nas respostas sejam atribuídas ao efeito do perfil sobre a personalização, e não a variações no conteúdo da pergunta.

### *Geração de Respostas*

Para cada pergunta neutra, foram produzidos dois tipos de saída:

- **Resposta neutra:** gerada pelo PersonaReAct com o módulo *Tailor* desativado, ignorando qualquer perfil de usuário;
- **Respostas personalizadas:** seis variações, uma para cada perfil, com o módulo *Tailor* ativado, de modo a ajustar linguagem, conteúdo e estilo ao perfil informado.

Assim, cada pergunta resultou em um conjunto de sete respostas: uma neutra e seis personalizadas. Ao todo, foram gerados 140 conjuntos de respostas (20 perguntas × 7 variações). A resposta neutra atua como *baseline* para medir o grau de variação nas versões personalizadas.

Para esse experimento, foram utilizados os modelos o3-mini (modelo com raciocínio) e GPT-4.1 mini (modelo sem raciocínio) da OpenAI por eles terem apresentado um melhor desempenho no Experimento 1. Os modelos foram instanciados com parâmetro *temperature* = 1, a fim de estimular maior variação linguística. Os demais parâmetros mantiveram os valores padrão da API (*top\_p* = 1.0, *frequency\_penalty* = 0.0).

O Quadro 6 apresenta um exemplo de pergunta e suas sete variações. Mais exemplos podem ser encontrados no Apêndice D.

### *Análise de Similaridade Textual*

A primeira etapa de análise consistiu em comparar as respostas personalizadas de cada pergunta com a respectiva resposta neutra, com o objetivo de mensurar a variação textual induzida pela personalização. As métricas de similaridade textual Distância de Edição, BLEU e ROUGE-L, descritas na Seção 5.3 foram utilizadas. Essas métricas quantificam a sobreposição lexical e a proximidade superficial entre as respostas, sem necessariamente capturar sua equivalência semântica.

### *Análise de Similaridade Semântica*

A segunda etapa da análise buscou avaliar a semelhança de significado entre cada resposta personalizada e a respectiva resposta neutra, com o objetivo de verificar se a personalização foi realizada sem comprometer o conteúdo informacional essencial. Para isso, foi utilizada a métrica de Distância Cosseno entre vetores de *embeddings* das sentenças.

Os *embeddings* foram gerados a partir do modelo all-MiniLM-L6-v2, fornecido

Quadro 6 – Respostas personalizadas à pergunta.

<b>Pergunta:</b> “Qual a velocidade média do jogador Bruno no jogo contra o Ceará no dia 1º de abril de 2023?”	
<b>Perfil</b>	<b>Resposta Personalizada</b>
<b>Resposta Padrão</b>	A velocidade média de Bruno Pacheco no jogo contra o Ceará no dia 1º de abril de 2023 foi de 1.3622634444889454.
<b>Torcedor do Time</b>	Olha, meu amigo, sobre o Bruno Pacheco naquele jogo contra o Ceará no dia 1 de abril, a verdade é que ele não se destacou muito na velocidade, viu? A velocidade média dele ficou em torno de 1.36, o que indica que ele até tentou, mas ficou devendo em relação ao que a torcida esperava. A gente sabe que ele pode dar mais, né? Esperamos vê-lo correndo com mais garra na próxima! Vamos torcer!
<b>Técnico do Time</b>	A velocidade média registrada por Bruno Pacheco no jogo contra o Ceará em 1º de abril de 2023 foi de 1,36 m/s. Esse dado é relevante para análises de desempenho, especialmente considerando a intensidade e o ritmo de jogo, que pode impactar não apenas o resultado, mas também a condição física do atleta ao longo da partida.
<b>Narrador de Jogo de Futebol</b>	No jogo contra o Ceará no dia 1 de abril, Bruno Pacheco mostrou-se um verdadeiro corredor de elite em campo! Ele atingiu uma velocidade média de 1.36 metros por segundo, o que, para facilitar a imagem, corresponde a algo em torno de 4.9 km/h. Uma corrida impressionante, sem dúvida! Foi um esforço notável que destacou seu condicionamento físico e dedicação! Ele correu com garra e envolvimento, mantendo a intensidade para ajudar o time. O torcedor pode ficar tranquilo, pois Bruno deu o sangue novamente!
<b>Influenciador Digital</b>	Com certeza! O Bruno Pacheco estava mostrando toda a sua rapidez no jogo contra o Ceará no começo de abril. Ele atingiu uma velocidade média de 1.36 m/s, o que é bem impressionante! Se você está pensando em postar algo legal sobre isso, pode dizer que ele praticamente voou em campo, fazendo jogadas rápidas e ajudando a equipe durante a partida. Aproveita pra trazer um pouco dessa energia e emoção para o seu post, destacando o quanto ele foi um destaque e fez a diferença no jogo!
<b>Fisioterapeuta do Time</b>	A velocidade média do Bruno Pacheco no jogo contra o Ceará, realizado no dia 1º de abril de 2023, foi de aproximadamente 1.36 m/s. Para uma análise da demanda mecânica, esse valor deve ser considerado em conjunto com outros fatores como a carga de trabalho total e a intensidade das corridas durante a partida. Isso nos ajuda a entender melhor o esforço fisiológico ao qual ele foi submetido e a planejar uma recuperação adequada. Se precisar de mais dados sobre a performance dele, ficarei à disposição!
<b>Jornalista Esportivo</b>	A velocidade média do Bruno Pacheco no jogo contra o Ceará, realizado em 1º de abril de 2023, foi de 1.36 m/s. Essa estatística é crucial para entender o desempenho do lateral, que se destacou por sua eficiência e disposição em campo, contribuindo tanto na defesa quanto nas investidas ao ataque. Incluir esse dado em sua matéria pode ressaltar a intensidade com que ele se movimentou durante a partida, destacando seu papel fundamental na equipe.

Fonte: elaborado pelo autor.

pela biblioteca SentenceTransformers (Reimers; Gurevych, 2019), amplamente reconhecida pelo bom desempenho em tarefas de similaridade textual. A escolha desse modelo se deve ao seu equilíbrio entre custo computacional e qualidade semântica, permitindo capturar relações relevantes entre sentenças mesmo em contextos de linguagem natural em português. Embora compacto, o all-MiniLM-L6-v2 já foi avaliado em cenários multilíngues e demonstrou desempenho consistente em português, alcançando resultados competitivos em *benchmarks* de *Semantic Textual Similarity* (STS). Assim, o modelo se mostrou uma opção adequada para este trabalho, em que há grandes volume de dados e a eficiência computacional é um fator relevante.

Ao contrário da análise textual, neste caso espera-se uma alta similaridade semântica

entre as respostas de diferentes perfis, uma vez que todas respondem, em essência, à mesma pergunta (ainda que formulada em tons distintos para cada perfil).

### *Considerações e Limitações*

Este experimento não tem como objetivo identificar qual perfil gerou a melhor resposta, mas sim avaliar o grau de diferenciação das respostas personalizadas em relação à versão neutra. Não foi realizada uma avaliação qualitativa com LLMs, uma vez que o foco está na análise quantitativa da divergência textual e semântica induzida pela personalização.

Cabe destacar que a análise textual é sensível a reescritas e variações estilísticas, podendo capturar diferenças superficiais que não necessariamente refletem mudança de significado. Por outro lado, a análise semântica, ainda que mais robusta, pode não detectar nuances pragmáticas ou estilísticas que são relevantes sob a perspectiva de um usuário humano. Assim, ambas as abordagens são complementares e contribuem para uma compreensão mais abrangente dos efeitos da personalização.

O Quadro 7 apresenta a rastreabilidade entre as hipóteses formuladas, os experimentos propostos para testá-las e as contribuições efetivas do trabalho. Essa rastreabilidade permite evidenciar como cada experimento está alinhado às perguntas de pesquisa e como os resultados esperados contribuem diretamente para os avanços metodológicos e técnicos desta tese.

**Quadro 7 – Rastreabilidade entre Hipóteses, Experimentos e Contribuições**

<b>Hipótese</b>	<b>Experimento</b>	<b>Contribuições Relacionadas</b>
H1 – Personalização melhora utilidade, clareza e interpretabilidade das respostas	E1 – Eficácia e Comparação entre abordagens de personalização	C1 – Definição do PersonaReAct C2 – Implementação do sistema C3 – Módulo Tailor C4 – Estudo no domínio esportivo
H2 – Perfis diferentes geram respostas distintas mesmo para a mesma pergunta	E2 – Análise de variação textual entre perfis	C1 – Definição do PersonaReAct C2 – Implementação do sistema C3 – Módulo Tailor C4 – Estudo no domínio esportivo e médico
H3 – A separação explícita da personalização melhora a organização das respostas	E1 – Eficácia e Comparação entre abordagens de personalização	C4 – Estudo no domínio esportivo e médico C5 - Comparação com o ReAct C7 – Criação de um corpus
H4 – O tipo de modelo utilizado influencia o desempenho das diferentes abordagens de personalização	E1 – Eficácia e Comparação entre abordagens de personalização	C6 – Comparação entre modelos C8 – Criação de critérios de Avaliação

Fonte: elaborado pelo autor.

## 5.6 Limitações e Considerações Metodológicas

Esta seção apresenta as principais limitações da abordagem adotada nesta tese, bem como considerações relevantes que orientam futuras generalizações e melhorias do *framework* PersonaReAct.

### *Idioma*

Todos os *prompts* e avaliações foram realizadas exclusivamente em português. O *prompt* original foi concebido diretamente nesse idioma, sem processo de tradução. Embora isso tenha facilitado a adaptação ao público-alvo e à base de dados do estudo de caso, limita a avaliação da aplicabilidade imediata da abordagem a outros idiomas. Experimentos futuros podem explorar a adaptação do PersonaReAct a contextos multilíngues.

### *Restrição ao uso de modelos mini e nano da OpenAI*

Os experimentos desta tese foram conduzidos exclusivamente com modelos da família *mini* e *nano* da OpenAI. Essa escolha foi motivada principalmente por questões de viabilidade econômica, dado o elevado volume de interações geradas nas avaliações (o que tornaria o uso de modelos mais robustos financeiramente inviável no contexto desta pesquisa).

Embora essa decisão tenha permitido explorar uma ampla variedade de cenários e combinações de forma escalável, ela também introduz limitações. Modelos mais avançados, com maior capacidade de raciocínio e processamento contextual, podem apresentar comportamentos diferentes, tanto na qualidade das respostas quanto na forma de incorporar a personalização. Além disso, o uso restrito a uma única fornecedora impede avaliar possíveis variações de desempenho e compatibilidade em modelos de outros fabricantes, o que pode afetar a generalização dos resultados para ambientes com diferentes infraestruturas ou políticas de acesso.

### *Falta de avaliação humana*

A avaliação das respostas foi realizada exclusivamente por meio do método *LLM-as-a-Judge*, utilizando o *framework* G-Eval. Essa escolha se deveu à indisponibilidade de avaliadores humanos com os perfis definidos e ao tempo reduzido para conduzir avaliações qualitativas manuais. Apesar disso, o uso de critérios bem definidos e a repetição das avaliações por 5 execuções contribuíram para mitigar a variabilidade das notas. Ainda assim, acredita-se que

avaliadores humanos com familiaridade com os perfis utilizados poderiam fornecer julgamentos mais contextualmente informados sobre a utilidade das respostas personalizadas.

### *Outras considerações técnicas*

A escolha do SQLite como banco de dados relacional foi adequada para a escala do projeto, não havendo gargalos de desempenho perceptíveis. Da mesma forma, o uso do LangGraph em sua forma pura (StateGraph) mostrou-se suficiente para estruturar o fluxo do PersonaReAct. No entanto, à medida que o sistema for escalado ou transferido para domínios com estruturas mais complexas, pode ser necessário explorar arquiteturas mais robustas ou bancos de dados com maior capacidade de escalabilidade.

## 6 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados obtidos a partir dos experimentos realizados com o *framework* PersonaReAct, cujo objetivo é personalizar respostas geradas por LLMs de acordo com o perfil do usuário. Os experimentos foram conduzidos com o propósito de validar as hipóteses formuladas no capítulo introdutório desta tese, avaliando aspectos como a eficácia da personalização das respostas, a variação das respostas quando comparados diferentes perfis e as diferenças de personalização com relação a outras abordagens.

Cada seção a seguir corresponde a um dos experimentos descritos na Seção 5.5, divididos nos dois estudos de caso, apresentando análises quantitativas e qualitativas sobre o desempenho da solução. Os resultados são acompanhados de visualizações gráficas, comparações entre diferentes condições experimentais (como a presença ou ausência de personalização) e discussões a respeito dos efeitos observados.

### 6.1 Experimento 1 - Eficácia da Personalização

#### *Estudo de Caso 1: Domínio Esportivo*

Este experimento avaliou o impacto da personalização de respostas no domínio esportivo sobre sua “Linguagem e Estilo Apropriados”, “Interpretação e Explicação dos Dados” e “Coerência do Estilo ao Longo do Texto” ao longo do texto, conforme descrito na Seção 5.5.

As Tabelas 1, 2 e 3 apresentam os resultados agregados das avaliações realizadas por meio do *framework* G-Eval para os três critérios definidos utilizando modelos com capacidade de raciocínio (o3-mini, o4-mini, GPT-5 mini e GPT-5 nano).

Tabela 1 – Médias e desvios padrão por modelo — ReAct (com raciocínio) - Domínio Esportivo.

Critério	o3 mini	o4 mini	GPT 5 mini	GPT 5 nano
Linguagem e Estilo Apropriados	2,42 ± 1,09	<b>3,17 ± 1,30</b>	2,71 ± 0,74	2,58 ± 0,70
Interpretação e Explicação dos Dados	2,90 ± 0,82	3,47 ± 0,96	<b>3,75 ± 1,66</b>	3,20 ± 0,80
Coerência do Estilo ao Longo do Texto	3,10 ± 1,24	4,80 ± 1,95	<b>6,50 ± 2,62</b>	3,79 ± 1,08

Fonte: elaborado pelo autor.

A análise comparativa das três abordagens evidencia um avanço progressivo na qualidade das respostas à medida que se introduzem aprimoramentos no processo de geração. O ReAct, na configuração original, apresentou médias relativamente baixas nos três critérios avaliados, com destaque negativo para “Linguagem e Estilo Apropriados”, cujos valores se

Tabela 2 – Médias e desvios padrão por modelo — ReAct usando *Prompt* Enriquecido (com raciocínio) - Domínio Esportivo.

<b>Critério</b>	o3 mini	o4 mini	GPT 5 mini	GPT 5 nano
Linguagem e Estilo Apropriados	9,07 ± 0,14	<b>9,08 ± 0,09</b>	9,06 ± 0,06	5,94 ± 2,78
Interpretação e Explicação dos Dados	9,02 ± 0,10	9,09 ± 0,10	<b>9,13 ± 0,10</b>	6,09 ± 2,87
Coerência do Estilo ao Longo do Texto	9,14 ± 0,08	<b>9,15 ± 0,07</b>	9,09 ± 0,06	7,98 ± 2,36

Fonte: elaborado pelo autor.

Tabela 3 – Médias e desvios padrão por modelo — PersonaReAct (com raciocínio) - Domínio Esportivo.

<b>Critério</b>	o3 mini	o4 mini	GPT 5 mini	GPT 5 nano
Linguagem e Estilo Apropriados	8,92 ± 0,42	<b>9,11 ± 0,14</b>	9,04 ± 0,08	7,94 ± 1,57
Interpretação e Explicação dos Dados	8,60 ± 0,60	9,05 ± 0,12	<b>9,12 ± 0,10</b>	8,32 ± 1,27
Coerência do Estilo ao Longo do Texto	9,04 ± 0,24	<b>9,14 ± 0,09</b>	9,08 ± 0,06	8,91 ± 0,52

Fonte: elaborado pelo autor.

mantiveram entre 2,42 e 3,17, independentemente do modelo utilizado. A introdução do *Prompt* Enriquecido elevou substancialmente as pontuações para os modelos o3-mini, o4-mini e GPT-5 mini, que passaram a registrar valores superiores a 9,0 em todos os critérios, com desvios-padrão reduzidos, indicando uma consistência mais alta nos resultados. No entanto, o GPT-5 nano mostrou desempenho mais modesto e alta variabilidade, sugerindo maior sensibilidade a limitações de capacidade. Com a adoção do PersonaReAct, as médias se mantiveram elevadas nos modelos maiores e houve melhora expressiva no GPT-5 nano, que passou a apresentar valores acima de 7,9 e menor dispersão, sobretudo em “Coerência do Estilo ao Longo do Texto”. Esses resultados sugerem que, enquanto o *Prompt* Enriquecido já garante ganhos significativos de clareza e consistência, a camada de personalização do PersonaReAct contribui para estabilizar e adaptar a produção textual, especialmente em modelos menos robustos.

As Tabelas 4, 5 e 6 apresentam os resultados agregados das avaliações realizadas por meio do *framework* G-Eval para os três critérios definidos utilizando modelos com capacidade de raciocínio (GPT-4.1 mini, GPT-4.1 nano e GPT-4o mini).

Tabela 4 – Médias e desvios padrão por modelo — ReAct (sem raciocínio) - Domínio Esportivo.

<b>Critério</b>	GPT 4.1 mini	GPT-4o mini	GPT 4.1 nano
Linguagem e Estilo Apropriados	<b>3,47 ± 1,56</b>	3,06 ± 1,38	3,23 ± 1,22
Interpretação e Explicação dos Dados	3,42 ± 0,93	3,31 ± 1,24	<b>3,47 ± 0,79</b>
Coerência do Estilo ao Longo do Texto	6,90 ± 2,53	6,86 ± 2,76	<b>7,01 ± 2,31</b>

Fonte: elaborado pelo autor.

Na configuração sem raciocínio, observa-se que o ReAct, em sua forma original, apresenta desempenho moderado em “Coerência do Estilo ao Longo do Texto”, com médias

Tabela 5 – Médias e desvios padrão por modelo — ReAct usando *Prompt* Enriquecido (sem raciocínio) - Domínio Esportivo.

<b>Critério</b>	GPT 4.1 mini	GPT-4o mini	GPT 4.1 nano
Linguagem e Estilo Apropriados	<b>8,91 ± 0,30</b>	6,20 ± 2,48	6,16 ± 2,46
Interpretação e Explicação dos Dados	<b>8,88 ± 0,31</b>	5,82 ± 2,44	5,52 ± 2,56
Coerência do Estilo ao Longo do Texto	<b>9,08 ± 0,11</b>	8,55 ± 1,41	8,38 ± 1,64

Fonte: elaborado pelo autor.

Tabela 6 – Médias e desvios padrão por modelo — PersonaReAct (sem raciocínio) - Domínio Esportivo.

<b>Critério</b>	GPT 4.1 mini	GPT-4o mini	GPT 4.1 nano
Linguagem e Estilo Apropriados	<b>9,06 ± 0,15</b>	8,94 ± 0,29	8,14 ± 1,47
Interpretação e Explicação dos Dados	<b>9,02 ± 0,09</b>	8,88 ± 0,24	7,84 ± 1,45
Coerência do Estilo ao Longo do Texto	<b>9,10 ± 0,08</b>	9,04 ± 0,13	8,90 ± 0,62

Fonte: elaborado pelo autor.

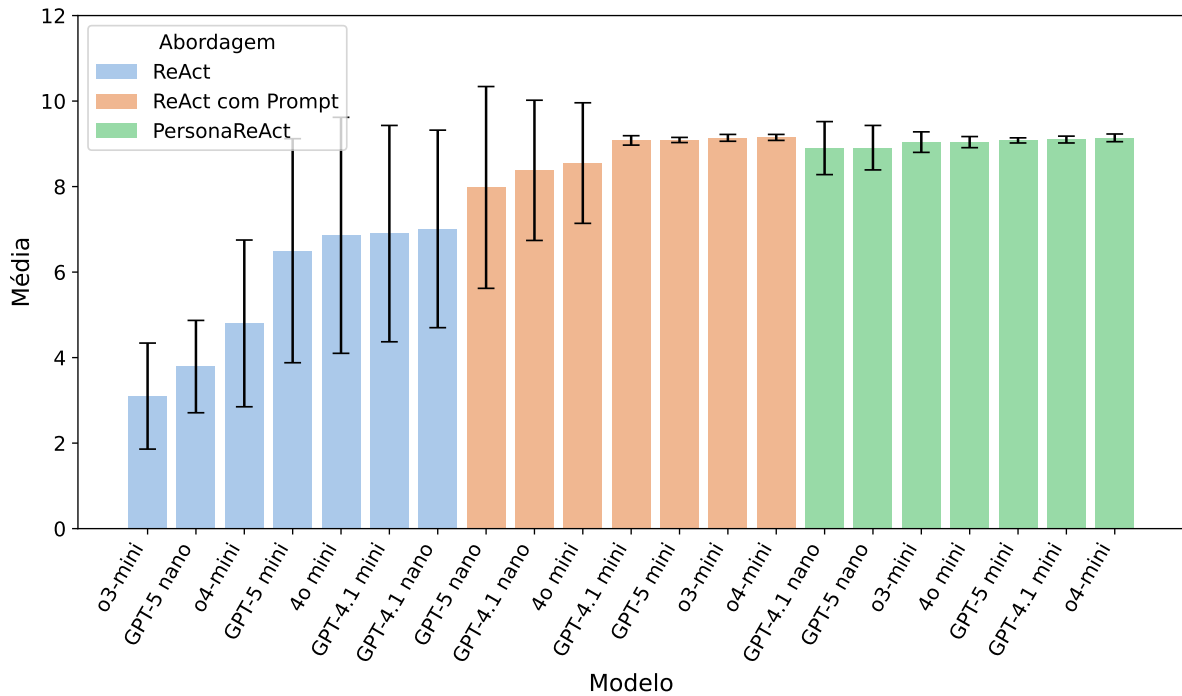
próximas a 7,0, mas resultados consideravelmente mais baixos para “Interpretação e Explicação dos Dados” e “Linguagem e Estilo Apropriados”, situando-se entre 3,06 e 3,47. A introdução do *Prompt* Enriquecido eleva substancialmente as médias do GPT-4.1 mini, que atinge valores acima de 8,8 em todos os critérios, mantendo desvios-padrão baixos e indicando respostas mais consistentes. Entretanto, os modelos GPT-4o mini e GPT-4.1 nano apresentam ganhos significativos apenas em “Coerência do Estilo”, mas mantêm desempenho irregular nos demais critérios, com variações acentuadas. Com a aplicação do PersonaReAct, os três modelos alcançam médias mais homogêneas e próximas entre si, com destaque para a melhoria expressiva do GPT-4.1 nano, que ultrapassa a marca de 7,8 em todos os critérios e reduz a dispersão dos resultados. Esses achados indicam que, na ausência de raciocínio estruturado, o PersonaReAct consegue não apenas manter alta qualidade textual, mas também atenuar diferenças de desempenho entre modelos de diferentes capacidades.

As Figuras 6, 7 e 8 apresentam os gráficos de barras de todos os modelos e abordagens, de forma conjunta, permitindo que as diferenças entre os cenários sejam percebidas de forma mais imediata, facilitando a identificação de padrões e contrastes relevantes.

No caso do critério de “Coerência do Estilo ao Longo do Texto” (Figura 6), o gráfico evidencia o ganho expressivo proporcionado pelas abordagens com maior grau de personalização, sobretudo o PersonaReAct, que mantém médias elevadas e estáveis mesmo na ausência de raciocínio explícito. Esse comportamento contrasta com o desempenho mais irregular do ReAct usando *Prompt* Enriquecido, que apresenta maior dispersão nos resultados.

Para o critério de “Interpretação e Explicação dos Dados” (Figura 7), a visualização

Figura 6 – Avaliação do critério “Coerência do Estilo ao Longo do Texto” nos modelos aplicados ao domínio esportivo.



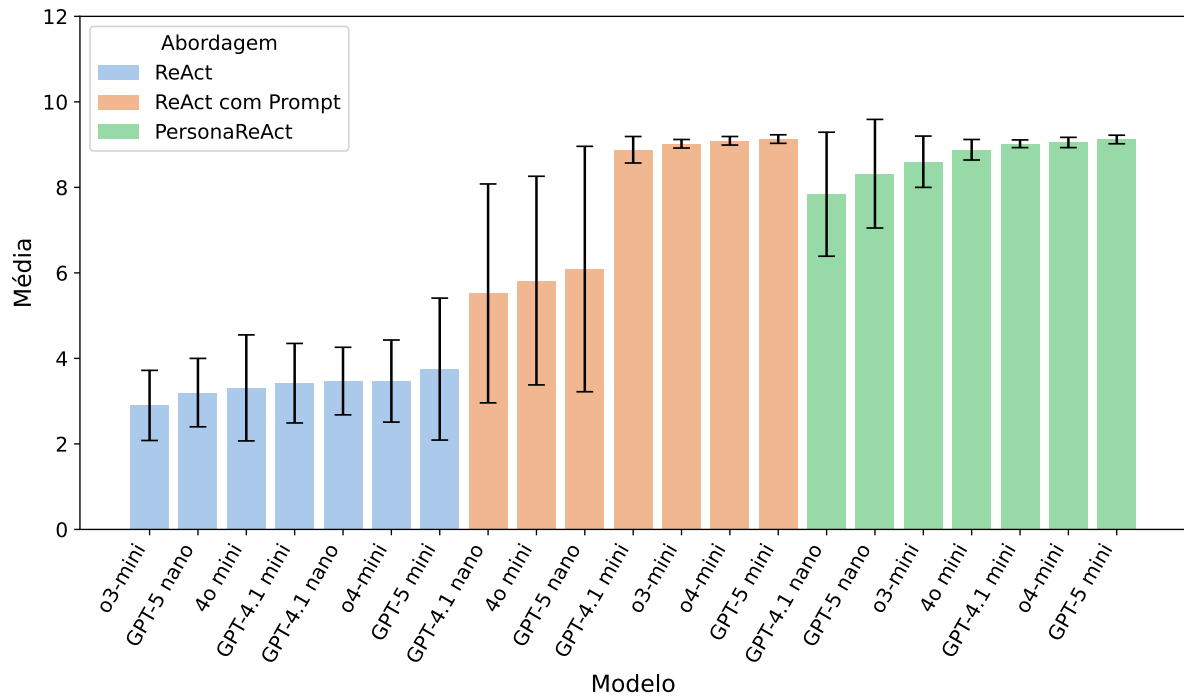
Fonte: elaborado pelo autor.

confirma que o uso de *prompts* enriquecidos e da camada de personalização do PersonaReAct contribui para respostas mais claras e informativas, especialmente nos modelos GPT-4.1 mini, GPT-4o mini e GPT-5 mini. Observa-se também que, nos modelos com menor capacidade, as diferenças entre abordagens são mais pronunciadas, indicando maior impacto das melhorias no processo de geração.

Já no critério de “Linguagem e Estilo Adequados” (Figura 7), o gráfico revela um padrão semelhante: as abordagens que incorporam estratégias adicionais de contextualização e personalização obtêm resultados mais consistentes e próximos do valor máximo. A comparação visual facilita a percepção de como essas melhorias reduzem a variabilidade dos desempenhos entre modelos e asseguram um padrão de qualidade mais uniforme.

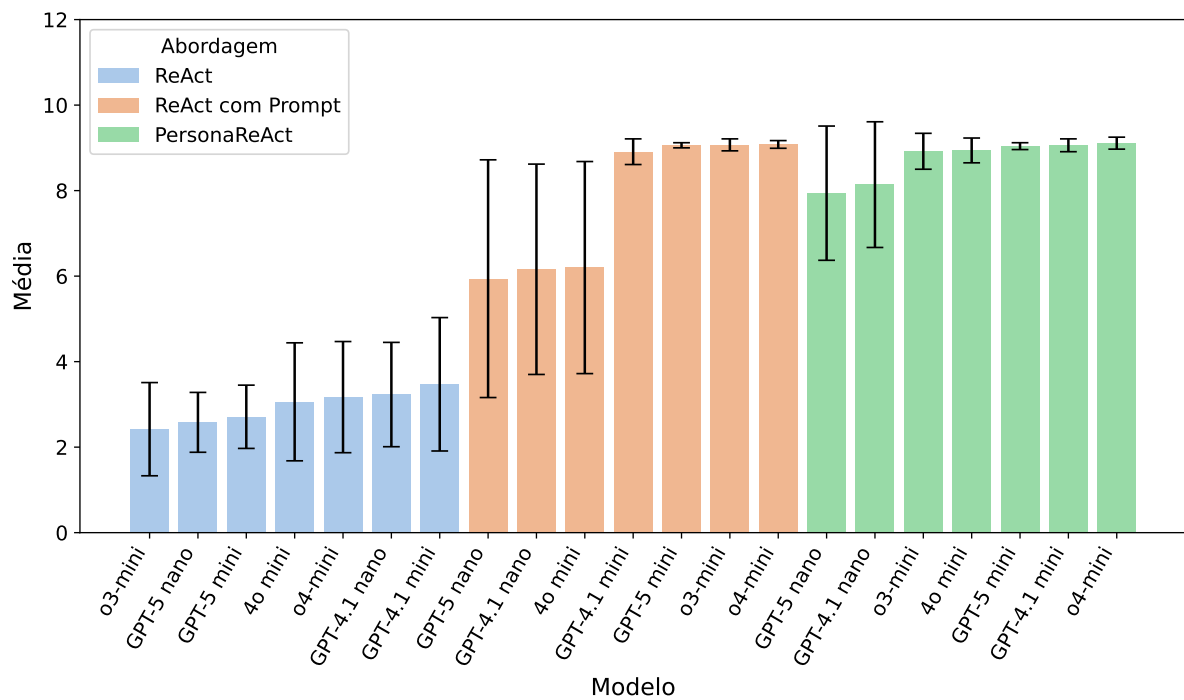
Os Quadros 8, 9 e 10 apresentam as notas e respectivas justificativas atribuídas pelo *framework* G-Eval para a seguinte pergunta: “Qual dos nossos jogadores teve a maior diferença entre o tempo que passou correndo e o tempo que ficou parado no jogo contra o Ceará?”. As respostas exibidas correspondem ao perfil Torcedor do Time e foram geradas pelo modelo o4 mini da OpenAI. Este é apenas um exemplo entre as diversas combinações de perfis e modelos avaliados ao longo do experimento. Além de ilustrar o desempenho, esses quadros permitem

Figura 7 – Avaliação do critério “Interpretação e Explicação dos Dados” nos modelos aplicados ao domínio esportivo.



Fonte: elaborado pelo autor.

Figura 8 – Avaliação do critério “Linguagem e Estilo Apropriados” nos modelos aplicados ao domínio esportivo.



Fonte: elaborado pelo autor.

compreender os motivos que levaram o G-Eval a atribuir as notas, oferecendo uma visão mais clara dos critérios qualitativos utilizados na avaliação.

Quadro 8 – Notas e justificativas do *framework* G-Eval para o domínio esportivo. Abordagem: ReAct.

<b>Resposta:</b> “O jogador com a maior diferença entre tempo parado e tempo correndo no jogo contra o Ceará foi Emanuel Brites, com diferença de 1.7437425506555424.”		
<b>Critério</b>	<b>Justificativa (G-Eval)</b>	<b>Nota</b>
Linguagem e Estilo Apropriado	A resposta fornece as informações solicitadas, mas carece do tom informal e apaixonado típico de um perfil Torcedor do Time. O uso de dados numéricos precisos parece deslocado e não condiz com o estilo informal esperado.	2,97
Interpretação e Explicação dos Dados	A saída identifica o jogador com a maior diferença no tempo de corrida e de permanência, mas a explicação carece de profundidade e relevância para o perfil do torcedor, pois não fornece contexto ou significância dessa estatística.	3,28
Coerência do Estilo ao Longo do Texto	A resposta usa valores preciso, que não são típicos de um perfil "Torcedor do Time", fazendo com que pareça forçado e não alinhado com a linguagem esperada de um fã casual.	3,12

Fonte: elaborado pelo autor.

Quadro 9 – Notas e justificativas do *framework* G-Eval para o domínio esportivo. Abordagem: ReAct usando *Prompt* Enriquecido.

<b>Resposta:</b> “Que orgulho, torcedor! Emanuel Brites foi quem mais se destacou contra o Ceará, apresentando a maior diferença entre o tempo correndo e parado, com aproximadamente 1,74. É dedicação em cada lance!”		
<b>Critério</b>	<b>Justificativa (G-Eval)</b>	<b>Nota</b>
Linguagem e Estilo Apropriado	A resposta usa expressões comuns e um tom que combina bem com o entusiasmo do torcedor, como “Que orgulho, torcedor!” e “É atendido em cada lance!”. O estilo é consistente e natural do começo ao fim, combinando perfeitamente com o perfil do torcedor.	9,08
Interpretação e Explicação dos Dados	O resultado identifica com precisão Emanuel Brites como o jogador com a maior diferença entre tempo de corrida e tempo parado, em linha com a pergunta inserida. Ele fornece uma explicação relevante e entusiasmada, adaptada à perspectiva do torcedor, aumentando o engajamento.	8,84
Coerência do Estilo ao Longo do Texto	O conteúdo mantém um tom entusiasmado e consistente, além de um vocabulário adequado para um fã, alinhado ao perfil do usuário. O estilo é coerente e apropriado em todos os aspectos, sem mudanças abruptas.	9,12

Fonte: elaborado pelo autor.

A análise qualitativa das respostas mostra uma melhoria na adaptação do conteúdo ao perfil “Torcedor do Time” à medida que as abordagens avançam. O ReAct apresenta uma resposta correta, porém excessivamente técnica e distante do tom esperado, evidenciando baixa adequação linguística e explicativa. Com o ReAct usando *Prompt* Enriquecido, observa-se um salto na personalização: a linguagem torna-se mais envolvente, com expressões emocionais e simplificação dos dados numéricos, aproximando-se do discurso típico do torcedor. Por fim, o PersonaReAct consolida essa evolução, entregando uma resposta que combina precisão factual com espontaneidade, utilizando expressões naturais e um tom entusiasmado consistente do início

Quadro 10 – Notas e justificativas do *framework* G-Eval para o domínio esportivo. Abordagem: PersonaReAct.

<b>Resposta:</b> “Emanuel Brites foi quem mais correu no jogo contra o Ceará – o cara praticamente não parou, passou quase o dobro do tempo em movimento do que parado. Deu um show de disposição!”		
<b>Critério</b>	<b>Justificativa (G-Eval)</b>	<b>Nota</b>
Linguagem e Estilo Apropriado	A resposta usa linguagem e expressões comuns, típicas de um torcedor, com tom entusiasmado e estilo informal, consistente com o perfil Torcedor do Time.	9,32
Interpretação e Explicação dos Dados	O resultado identifica com precisão Emanuel Brites como o jogador com a maior diferença no tempo de movimento, em linha com a pergunta inserida. Ele fornece uma explicação relevante e entusiasmada, adequada para um torcedor, que corresponde ao contexto do perfil do usuário.	8,55
Coerência do Estilo ao Longo do Texto	A resposta mantém um tom informal e entusiasmado consistente, adequado para um fã, com vocabulário e estrutura que se alinham bem ao perfil do usuário. Não há mudanças abruptas de estilo ou tom.	9.18

Fonte: elaborado pelo autor.

ao fim. Embora as notas do G-Eval para as duas últimas abordagens sejam próximas, a percepção qualitativa indica que o PersonaReAct alcança um equilíbrio entre clareza, estilo e relevância para o público-alvo, tornando a interação mais autêntica e alinhada às expectativas do usuário.

A análise dos resultados separados por perfil, apresentada nas Figuras 9, 10 e 11, evidencia o desempenho comparativo das abordagens ReAct, ReAct usando *Prompt* Enriquecido e PersonaReAct para os seis perfis avaliados, considerando os três critérios já mencionados (Linguagem, Coerência e Interpretação) em modelos específicos (GPT-5 mini, GPT-5 nano e GPT-4o mini). Esses modelos foram escolhidos por representarem de forma consistente o comportamento observado nos demais, sendo inviável, por restrições de espaço, apresentar todos os gráficos.

Nos modelos com capacidade de raciocínio, como o GPT-5 mini e o GPT-5 nano, há um padrão de desempenho mais uniforme entre os três critérios, enquanto no GPT-4o mini — sem raciocínio explícito — aparecem quedas mais pronunciadas, especialmente no critério Linguagem, para perfis que demandam maior adaptação contextual, como “Narrador de Jogo de Futebol” e “Influenciador Digital”. Esse comportamento sugere que a presença de mecanismos internos de raciocínio auxilia na preservação do estilo e na adequação semântica ao perfil, reforçando a hipótese de que o raciocínio deliberado favorece a personalização com consistência.

Ao comparar os dois modelos com raciocínio, o GPT-5 mini apresenta maior estabilidade entre perfis e melhores resultados, sobretudo em Consistência e Personalização, enquanto o GPT-5 nano exhibe oscilações mais perceptíveis, principalmente em perfis que exigem mudanças de registro mais radicais. Isso indica que modelos menores, ainda que com raciocínio, podem ter mais dificuldade para manter coerência estilística em personalizações profundas.

No contraste entre abordagens, o GPT-5 mini e o GPT-4o mini mostram vantagem consistente do PersonaReAct, principalmente nos critérios de Personalização e Consistência, enquanto no GPT-5 nano as diferenças entre PersonaReAct e ReAct usando *Prompt* Enriquecido são menos expressivas, sugerindo que, nesse caso, o enriquecimento do *prompt* é quase suficiente para atingir níveis altos de desempenho. Ainda assim, em modelos menores ou sem raciocínio, como o GPT-4o mini, a arquitetura dedicada do PersonaReAct se mostra mais benéfica.

Em síntese, destacam-se dois achados principais: (i) modelos com raciocínio mantêm melhor equilíbrio entre critérios e perfis, e (ii) o PersonaReAct entrega ganhos mais relevantes em cenários de menor capacidade computacional ou ausência de raciocínio explícito, consolidando-se como uma solução vantajosa para personalizações que demandam maior adaptação estilística e semântica.

### ***Estudo de Caso 2: Domínio Médico***

Este experimento avaliou o impacto da personalização de respostas no domínio médico, assim como no Estudo de Caso 1, sobre sua “Linguagem e Estilo Adequados”, “Interpretação e Explicação dos Dados” e “Coerência do Estilo ao Longo do Texto” ao longo do texto, conforme descrito na Seção 5.5.

As Tabelas 7, 8 e 9 apresentam os resultados agregados das avaliações realizadas por meio do *framework* G-Eval para os três critérios definidos utilizando modelos com capacidade de raciocínio (o3-mini, o4-mini, GPT-5 mini e GPT-5 nano).

Tabela 7 – Médias e desvios padrão por modelo — ReAct (com raciocínio) - Domínio Médico.

<b>Critério</b>	o3 mini	o4 mini	GPT 5 mini	GPT 5 nano
Linguagem e Estilo Adequados	4,67 ± 2,43	<b>6,08 ± 2,27</b>	6,08 ± 2,35	5,46 ± 2,17
Interpretação e Explicação dos Dados	3,49 ± 1,39	4,73 ± 1,45	<b>5,11 ± 2,12</b>	4,17 ± 1,23
Coerência do Estilo ao Longo do Texto	6,90 ± 2,40	<b>8,37 ± 1,23</b>	8,21 ± 1,48	8,05 ± 1,34

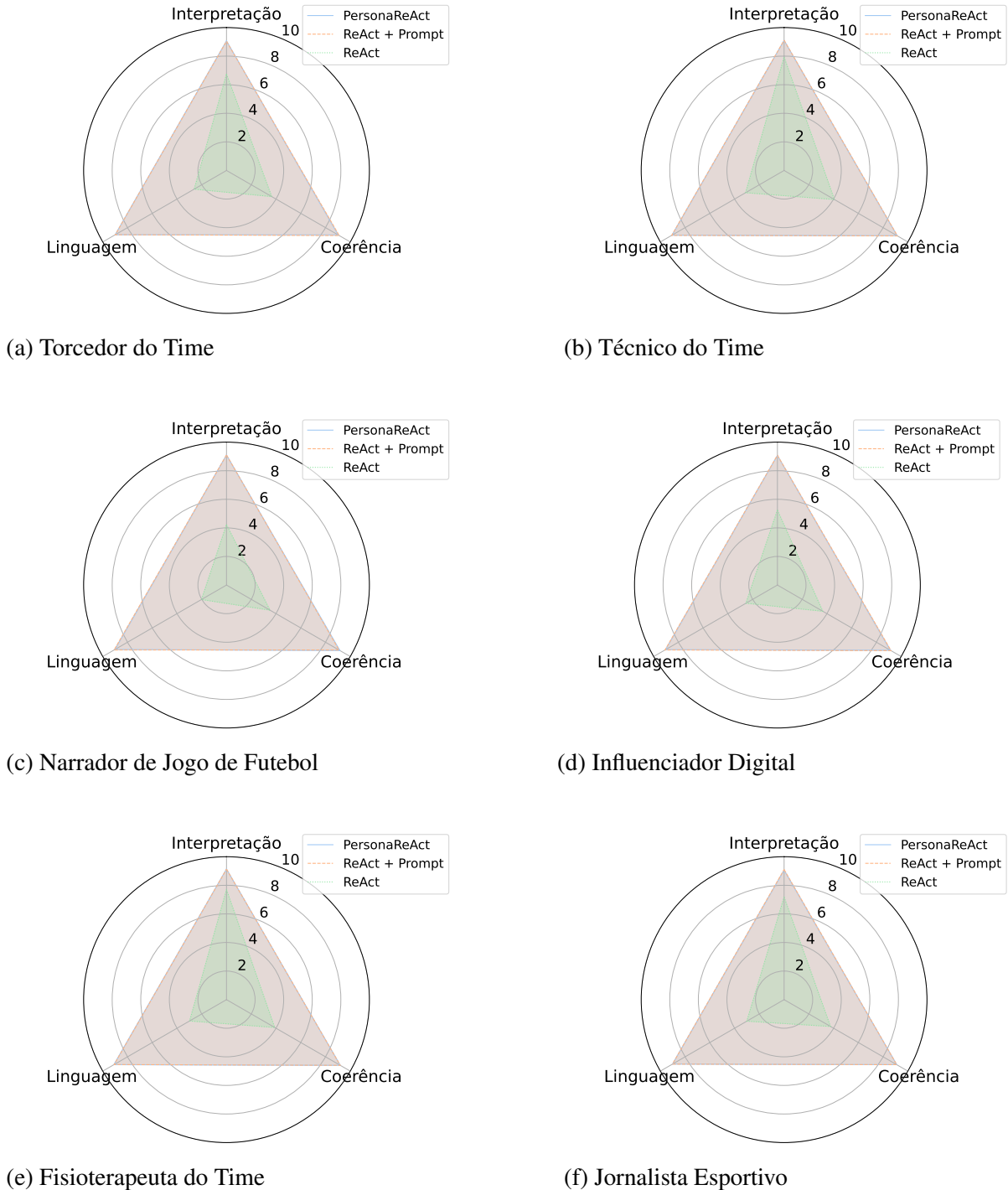
Fonte: elaborado pelo autor.

Tabela 8 – Médias e desvios padrão por modelo — ReAct usando *Prompt* Enriquecido (com raciocínio) - Domínio Médico.

<b>Critério</b>	o3 mini	o4 mini	GPT 5 mini	GPT 5 nano
Linguagem e Estilo Adequados	9,03 ± 0,06	<b>9,05 ± 0,06</b>	9,05 ± 0,06	8,05 ± 1,74
Interpretação e Explicação dos Dados	9,02 ± 0,10	9,08 ± 0,09	<b>9,11 ± 0,08</b>	5,82 ± 2,64
Coerência do Estilo ao Longo do Texto	9,13 ± 0,08	<b>9,16 ± 0,07</b>	9,09 ± 0,06	9,14 ± 0,27

Fonte: elaborado pelo autor.

Figura 9 – Comparação das médias por critério entre ReAct, ReAct usando *Prompt* Enriquecido e PersonaReAct para o modelo GPT-5 mini no domínio esportivo.



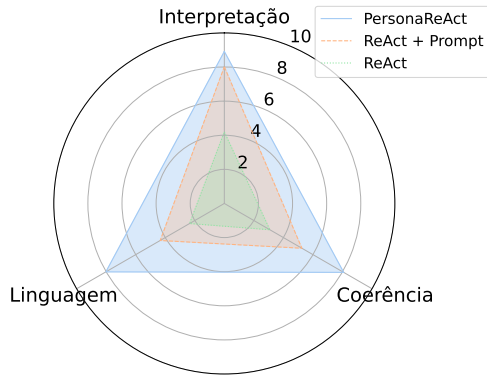
Fonte: elaborado pelo autor.

Tabela 9 – Médias e desvios padrão por modelo — PersonaReAct (com raciocínio) - Domínio Médico.

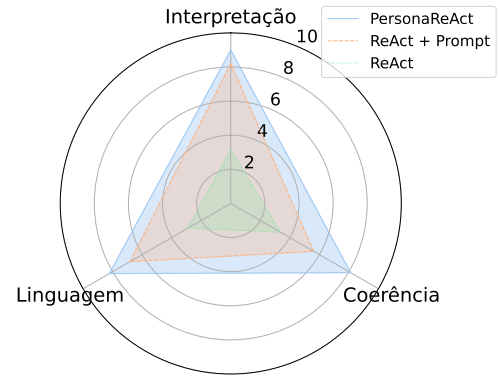
Critério	o3 mini	o4 mini	GPT 5 mini	GPT 5 nano
Linguagem e Estilo Adequados	8,95 ± 0,22	9,05 ± 0,05	<b>9,07 ± 0,05</b>	8,97 ± 0,14
Interpretação e Explicação dos Dados	8,81 ± 0,35	9,08 ± 0,11	<b>9,10 ± 0,08</b>	8,96 ± 0,17
Coerência do Estilo ao Longo do Texto	9,14 ± 0,09	<b>9,16 ± 0,07</b>	9,09 ± 0,05	9,14 ± 0,12

Fonte: elaborado pelo autor.

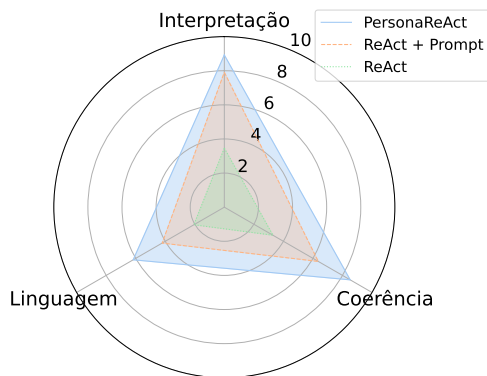
Figura 10 – Comparação das médias por critério entre ReAct, ReAct usando *Prompt* Enriquecido e PersonaReAct para o modelo GPT-5 nano no domínio esportivo.



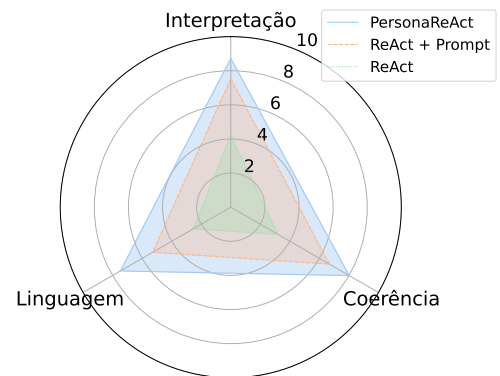
(a) Torcedor do Time



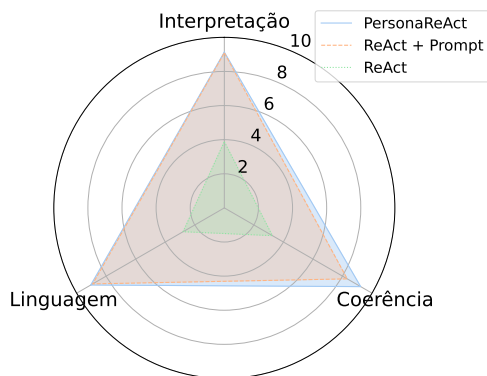
(b) Técnico do Time



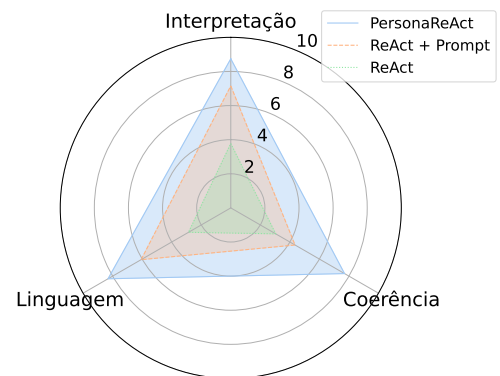
(c) Narrador de Jogo de Futebol



(d) Influenciador Digital



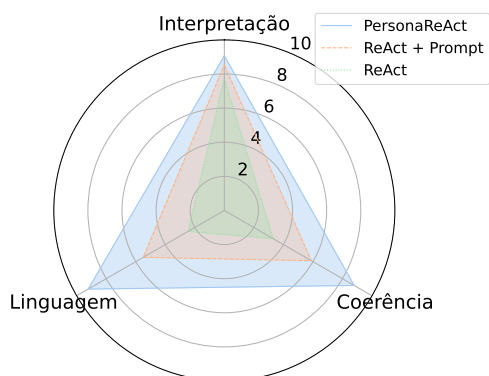
(e) Fisioterapeuta do Time



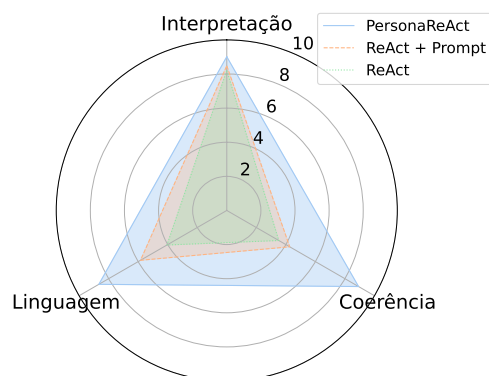
(f) Jornalista Esportivo

Fonte: elaborado pelo autor.

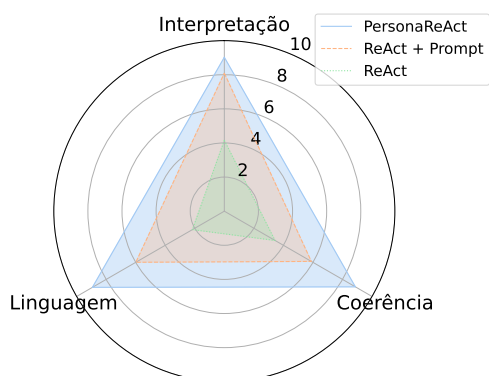
Figura 11 – Comparação das médias por critério entre ReAct, ReAct usando *Prompt* Enriquecido e PersonaReAct para o modelo GPT-4o mini no domínio esportivo.



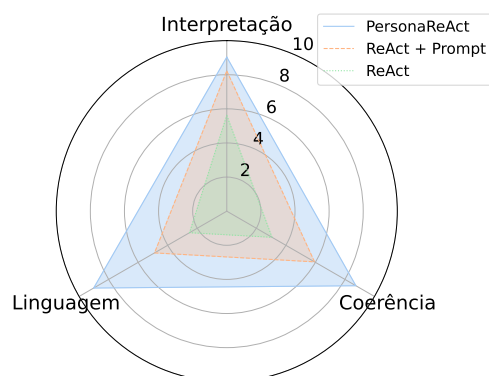
(a) Torcedor do Time



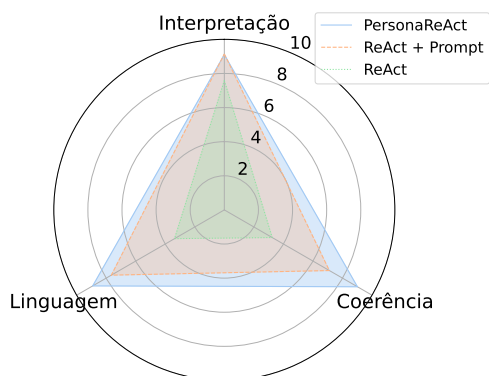
(b) Técnico do Time



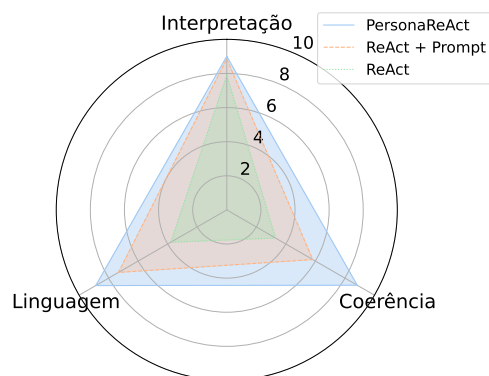
(c) Narrador de Jogo de Futebol



(d) Influenciador Digital



(e) Fisioterapeuta do Time



(f) Jornalista Esportivo

Fonte: elaborado pelo autor.

A análise comparativa das três abordagens revela diferenças claras no desempenho entre as configurações testadas. O ReAct, em sua forma original, apresentou médias baixas nos três critérios avaliados, com “Linguagem e Estilo Apropriados” sendo o ponto mais fraco, registrando valores entre 2,42 e 3,17, independentemente do modelo. Esse comportamento indica que, sem ajustes adicionais no *prompt*, os modelos mantêm limitações significativas na consistência e adequação da comunicação.

A introdução do *Prompt* Enriquecido produziu um salto expressivo no desempenho para os modelos o3-mini, o4-mini e GPT-5 mini, todos superando 9,0 em todos os critérios, com desvios-padrão muito baixos, sinalizando alta estabilidade nas respostas. No entanto, o GPT-5 nano manteve médias mais baixas (entre 5,94 e 7,98) e elevada variabilidade, reforçando sua maior sensibilidade a limitações de capacidade e à complexidade das instruções.

Com o PersonaReAct, as médias se mantiveram elevadas para os modelos maiores, praticamente no mesmo patamar do *Prompt* Enriquecido, mas o destaque esteve na melhora substancial do GPT-5 nano, que passou a registrar valores entre 7,94 e 8,91, além de apresentar redução consistente na dispersão, sobretudo em “Coerência do Estilo ao Longo do Texto”.

As Tabelas 10, 11 e 12 apresentam os resultados agregados das avaliações realizadas por meio do *framework* G-Eval para os três critérios definidos utilizando modelos com capacidade de raciocínio (GPT-4.1 mini, GPT-4.1 nano e GPT-4o mini).

Tabela 10 – Médias e desvios padrão por modelo — ReAct (sem raciocínio) - Domínio Médico.

<b>Critério</b>	GPT 4.1 mini	GPT-4o mini	GPT 4.1 nano
Linguagem e Estilo Apropriados	<b>5,68 ± 2,29</b>	5,36 ± 2,21	5,22 ± 2,40
Interpretação e Explicação dos Dados	<b>4,64 ± 1,72</b>	3,83 ± 1,48	3,54 ± 1,41
Coerência do Estilo ao Longo do Texto	<b>8,20 ± 1,35</b>	7,83 ± 1,69	7,42 ± 2,29

Fonte: elaborado pelo autor.

Tabela 11 – Médias e desvios padrão por modelo — ReAct usando *Prompt* Enriquecido (sem raciocínio) - Domínio Médico.

<b>Critério</b>	GPT 4.1 mini	GPT-4o mini	GPT 4.1 nano
Linguagem e Estilo Apropriados	<b>8,97 ± 0,17</b>	7,48 ± 2,35	7,92 ± 1,76
Interpretação e Explicação dos Dados	<b>8,80 ± 0,42</b>	4,79 ± 2,51	4,84 ± 2,39
Coerência do Estilo ao Longo do Texto	<b>9,14 ± 0,19</b>	8,56 ± 2,08	8,89 ± 1,15

Fonte: elaborado pelo autor.

Na configuração sem raciocínio, o ReAct em sua forma original apresenta desempenho razoável em “Coerência do Estilo ao Longo do Texto”, com médias variando entre 7,42 e 8,20, mas resultados significativamente inferiores para “Interpretação e Explicação dos Dados”,

Tabela 12 – Médias e desvios padrão por modelo — PersonaReAct (sem raciocínio) - Domínio Médico.

<b>Critério</b>	<b>GPT 4.1 mini</b>	<b>GPT-4o mini</b>	<b>GPT 4.1 nano</b>
Linguagem e Estilo Apropriados	<b>9,02 ± 0,07</b>	8,97 ± 0,11	8,84 ± 0,47
Interpretação e Explicação dos Dados	<b>9,04 ± 0,07</b>	8,94 ± 0,13	8,47 ± 0,67
Coerência do Estilo ao Longo do Texto	<b>9,11 ± 0,07</b>	9,07 ± 0,10	9,10 ± 0,11

Fonte: elaborado pelo autor.

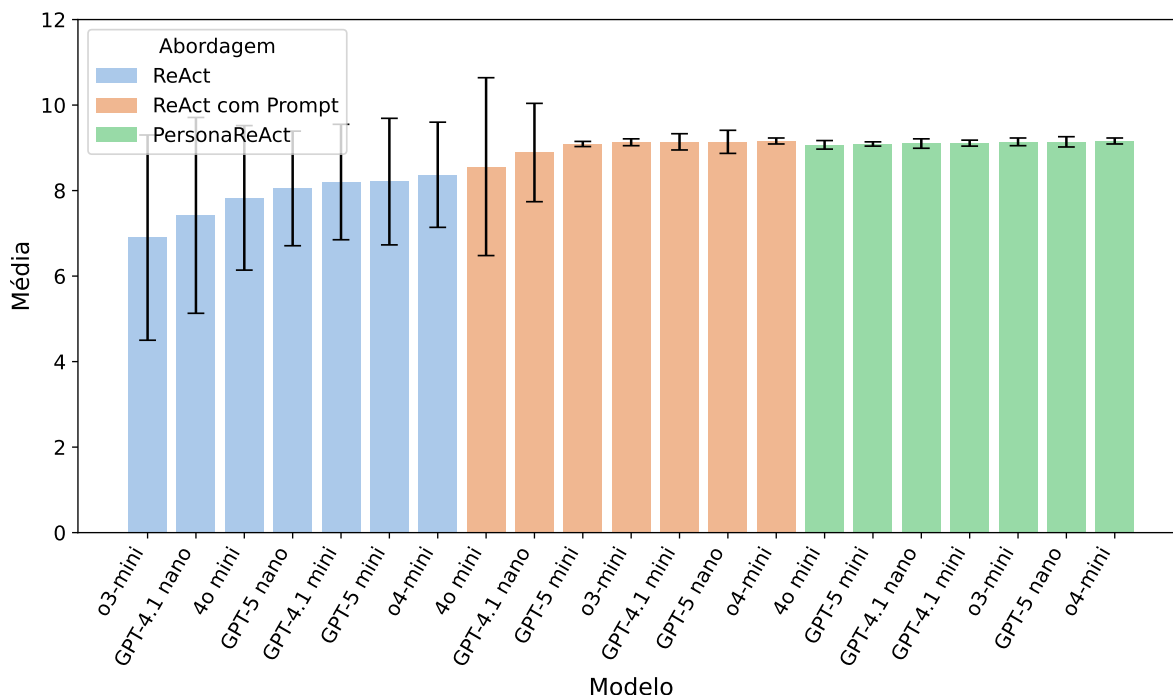
situando-se entre 3,54 e 4,64, e para “Linguagem e Estilo Apropriados”, com valores entre 5,22 e 5,68. A introdução do *Prompt* Enriquecido eleva substancialmente as médias do GPT-4.1 mini, que supera 8,8 em todos os critérios, mantendo desvios-padrão baixos, sinalizando maior consistência nas respostas. Contudo, os modelos GPT-4o mini e GPT-4.1 nano mostram ganhos claros em “Coerência do Estilo”, mas mantêm resultados mais modestos e variáveis em “Interpretação e Explicação dos Dados” e “Linguagem e Estilo Apropriados”, evidenciando sensibilidade à complexidade do conteúdo. Com a aplicação do PersonaReAct, os três modelos atingem médias elevadas e próximas entre si, acima de 8,4 em todos os critérios, com destaque para a expressiva evolução do GPT-4.1 nano, que alcança 9,10 em “Coerência do Estilo” e reduz consideravelmente a dispersão dos resultados. Esses achados indicam que, na ausência de raciocínio estruturado, o PersonaReAct não apenas mantém alta qualidade textual, mas também reduz disparidades de desempenho entre modelos de diferentes capacidades, garantindo maior homogeneidade e estabilidade nas respostas.

As Figuras 12, 13 e 14 reúnem, em um mesmo conjunto, os gráficos de barras referentes a todos os modelos e abordagens avaliados. Essa apresentação integrada favorece uma visualização imediata das variações entre os cenários, tornando mais simples reconhecer tendências, diferenças marcantes e relações de desempenho.

No critério de “Coerência do Estilo ao Longo do Texto”, a visualização evidencia que as abordagens com maior grau de refinamento (especialmente o PersonaReAct) alcançam resultados bastante estáveis, com médias elevadas em todos os modelos. O ganho é mais perceptível quando comparado ao ReAct original, que apresenta desempenho inferior e mais disperso, sobretudo nos modelos de menor capacidade. Já o ReAct usando *Prompt* Enriquecido mostra melhora considerável, mas ainda com alguma variação, indicando que a camada de personalização do PersonaReAct contribui para consolidar o padrão de qualidade.

Para o critério de “Interpretação e Explicação dos Dados”, o gráfico reforça que o uso de estratégias adicionais, como *prompts* enriquecidos e personalização, aumenta a clareza e a completude das respostas. Modelos como o o3-mini, o4-mini e GPT-5 mini atingem

Figura 12 – Avaliação do critério “Coerência do Estilo ao Longo do Texto” nos modelos aplicados ao domínio médico.



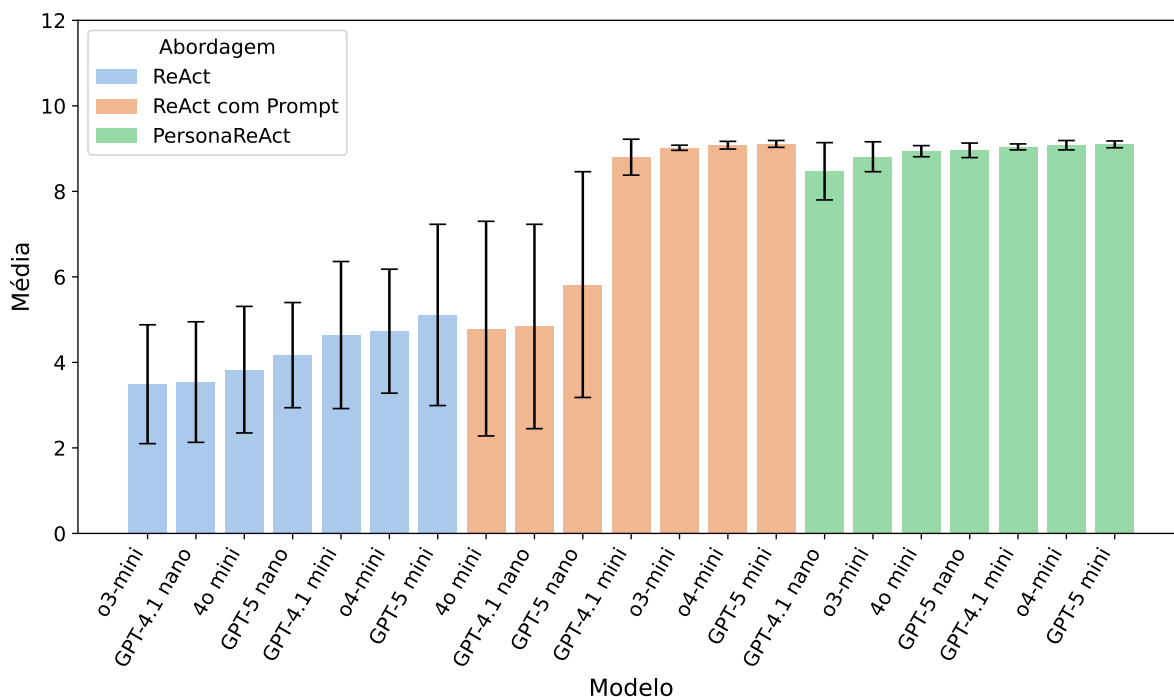
Fonte: elaborado pelo autor.

valores próximos do máximo com o PersonaReAct, enquanto no ReAct original os resultados são visivelmente mais modestos. Nos modelos de menor capacidade, como o GPT-5 nano, a diferença entre abordagens é acentuada, revelando maior sensibilidade às melhorias no processo de geração.

No critério de “Linguagem e Estilo Apropriados”, o padrão observado é semelhante: a incorporação de técnicas de contextualização e personalização eleva substancialmente as médias, ao mesmo tempo que reduz a dispersão entre modelos. Enquanto o ReAct original mantém valores mais baixos e irregulares, as versões com *Prompt* Enriquecido e PersonaReAct convergem para um desempenho mais uniforme, garantindo que a qualidade textual se mantenha alta independentemente da capacidade do modelo.

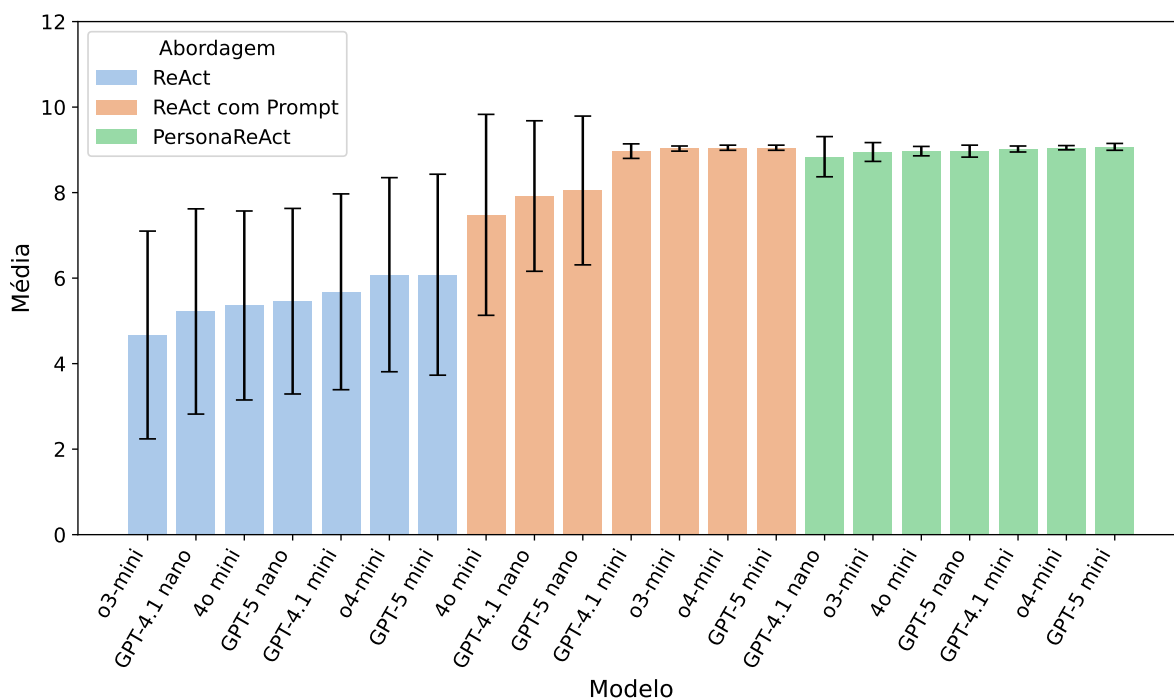
Os Quadros 11, 12 e 13 apresentam as notas e as respectivas justificativas geradas pelo *framework* G-Eval para a seguinte pergunta: “Qual exame apresentou a maior variedade de resultados distintos no período de 01/01/2020 a 01/09/2020?”. As respostas apresentadas nos quadros correspondem ao perfil Médico e foram geradas com o modelo o4 mini da OpenAI. Trata-se de uma entre várias combinações de perfis e modelos analisadas neste experimento. Mais do que exemplificar o desempenho obtido, esses quadros auxiliam na compreensão das

Figura 13 – Avaliação do critério “Interpretação e Explicação dos Dados” nos modelos aplicados ao domínio médico.



Fonte: elaborado pelo autor.

Figura 14 – Avaliação do critério “Linguagem e Estilo Apropriados” nos modelos aplicados ao domínio médico.



Fonte: elaborado pelo autor.

razões que levaram o G-Eval a atribuir as pontuações, oferecendo uma perspectiva detalhada sobre os critérios qualitativos empregados na avaliação.

Quadro 11 – Notas e justificativas do *framework* G-Eval para o domínio médico. Abordagem: ReAct.

<b>Resposta:</b> “O exame Hemograma apresentou a maior diversidade de resultados distintos (6272) no período de 1 de janeiro de 2020 a 1 de setembro de 2020.”		
<b>Critério</b>	<b>Justificativa (G-Eval)</b>	<b>Nota</b>
Linguagem e Estilo Apropriado	A resposta utiliza terminologia médica precisa e um tom factual, que se alinha bem com o estilo de comunicação esperado de um profissional médico. A linguagem é clara e direta, adequada ao perfil sem complexidade desnecessária.	8,85
Interpretação e Explicação dos Dados	O resultado identifica o Hemograma como tendo os resultados mais diversos, mas carece de profundidade na interpretação e relevância para as necessidades de um profissional médico.	5,13
Coerência do Estilo ao Longo do Texto	O resultado mantém um tom consistente e profissional, adequado ao contexto médico, condizente com o perfil do usuário como médico. O vocabulário e a estrutura são coerentes com a pergunta inserida, e não há mudanças abruptas de estilo ou tom.	8,88

Fonte: elaborado pelo autor.

Quadro 12 – Notas e justificativas do *framework* G-Eval para o domínio médico. Abordagem: ReAct usando *Prompt* Enriquecido.

<b>Resposta:</b> “No período de 01/01/2020 a 01/09/2020, o exame que apresentou a maior variedade de resultados distintos foi o Hemograma, com 6.272 resultados diferentes. Isso reflete a complexidade e abrangência dos parâmetros hematológicos avaliados nesse painel, informação relevante para sua análise clínica.”		
<b>Critério</b>	<b>Justificativa (G-Eval)</b>	<b>Nota</b>
Linguagem e Estilo Apropriado	A linguagem e as expressões utilizadas são condizentes com o estilo de um profissional médico, utilizando termos como "hemograma" e "parâmetros hematológicos". O tom é informativo e analítico, condizente com a postura de um médico. O estilo de comunicação é formal e preciso, alinhado com a forma como um perfil médico normalmente se comunica.	9,04
Interpretação e Explicação dos Dados	A saída identifica com precisão o hemograma como o que apresenta os resultados mais distintos, alinhando-se com os dados de entrada. Ela fornece uma explicação relevante sobre a complexidade dos parâmetros hematológicos, o que é útil para a análise clínica de um profissional médico.	9,06
Coerência do Estilo ao Longo do Texto	O resultado mantém um tom profissional consistente e um vocabulário adequado para um profissional da área médica, alinhando-se bem ao perfil do usuário. A transição entre as frases é suave e as informações são apresentadas de forma clara e coerente, atendendo ao contexto e aos requisitos de entrada.	9,27

Fonte: elaborado pelo autor.

A análise qualitativa evidencia uma melhoria progressiva na adequação das respostas ao perfil Médico à medida que as abordagens evoluem. No ReAct original, observa-se uma comunicação tecnicamente correta, com uso de terminologia médica e tom formal, mas limitada em profundidade interpretativa (especialmente no critério Interpretação e Explicação dos Dados, onde a nota é sensivelmente inferior às demais). Isso indica que, embora a resposta seja precisa,

Quadro 13 – Notas e justificativas do *framework* G-Eval para o domínio médico. Abordagem: PersonaReAct.

<p><b>Resposta:</b> “Durante o período de 01/01/2020 a 01/09/2020, o hemograma foi o exame que apresentou a maior variedade de resultados distintos, totalizando 6 272 valores únicos. Esse elevado número de padrões decorre da multiplicidade de parâmetros avaliados no hemograma — como contagem de eritrócitos, leucócitos (e seus subtipos), plaquetas e índices eritrocitários — além de variações analíticas e pré-analíticas inerentes ao método. Do ponto de vista clínico, essa diversidade reforça a sensibilidade do hemograma para detectar alterações hematológicas sutis em diferentes contextos (infecção, anemia, inflamação, distúrbios de coagulação etc.). Recomendo atenção especial à padronização de coleta e calibração dos analisadores para minimizar vieses e assegurar a comparabilidade dos resultados no acompanhamento longitudinal dos pacientes. Caso seu objetivo seja reduzir a complexidade de interpretação ou priorizar exames com menor variabilidade intrínseca, você pode complementar essa análise comparando o coeficiente de variação de outros painéis laboratoriais (como bioquímica sérica ou painel de coagulação) no mesmo intervalo. Isso ajudará a identificar quais protocolos oferecem maior reprodutibilidade e robustez diagnóstica em sua rotina clínica.”</p>		
Critério	Justificativa (G-Eval)	Nota
Linguagem e Estilo Apropriado	A resposta utiliza terminologia médica e explicações detalhadas típicas de um profissional da área médica, alinhando-se bem ao perfil esperado. O tom é formal e informativo, adequado ao estilo de comunicação de um médico. O uso de termos específicos como "hemograma", "eritrócitos" e "coeficiente de variação" é consistente com a forma como um médico se comunicaria.	9,09
Interpretação e Explicação dos Dados	O resultado identifica com precisão o hemograma como o que apresenta os resultados mais distintos, fornecendo uma explicação detalhada de seus parâmetros e relevância clínica. Oferece conselhos estratégicos sobre padronização e sugere a comparação da variabilidade com outros exames, alinhando-se bem ao perfil médico.	9,08
Coerência do Estilo ao Longo do Texto	O conteúdo mantém um tom profissional consistente e um vocabulário adequado para um profissional da área médica, alinhando-se perfeitamente ao perfil do usuário. A explicação é coerente, com transições suaves e sem mudanças abruptas de estilo ou tom, abordando efetivamente a dúvida sobre a variabilidade dos resultados dos exames.	9,12

Fonte: elaborado pelo autor.

carece de contextualização clínica e de elementos que a tornem mais útil para o profissional de saúde.

Com o ReAct usando *Prompt* Enriquecido, há um salto expressivo na qualidade: a resposta mantém o rigor terminológico, mas acrescenta explicações sobre a complexidade dos parâmetros hematológicos, aproximando-se mais das necessidades práticas de um médico. O tom segue formal e consistente, e a coerência textual se mantém elevada. As notas do G-Eval refletem essa melhoria, com ganhos claros em todos os critérios, especialmente em Interpretação e Explicação dos Dados.

O PersonaReAct consolida essa evolução ao integrar precisão factual, detalhamento técnico e recomendações práticas, como estratégias de padronização e sugestões de análises complementares. Essa abordagem entrega não apenas a resposta à pergunta, mas também um raciocínio clínico mais abrangente e aplicável, alinhando-se perfeitamente ao perfil médico. As notas do G-Eval mostram estabilidade e alto desempenho nos três critérios, confirmando que o

PersonaReAct atinge um equilíbrio entre clareza, profundidade e relevância, elevando a utilidade da interação para o público-alvo.

A análise dos resultados separados por perfil é apresentada nas Figuras 15, 16 e 17, que mostram gráficos de radar para cada um dos seis perfis avaliados, comparando as três abordagens: ReAct, ReAct usando *Prompt* Enriquecido e PersonaReAct. Cada gráfico representa as médias obtidas nos três critérios (Interpretação, Coerência e Linguagem) para um modelo específico. Esses modelos (GPT-5 mini, GPT-5 nano e GPT-4o mini) foram escolhidos por representarem de forma consistente os padrões observados nos demais experimentos, sendo inviável, por restrições de espaço, apresentar os resultados de todos os modelos avaliados.

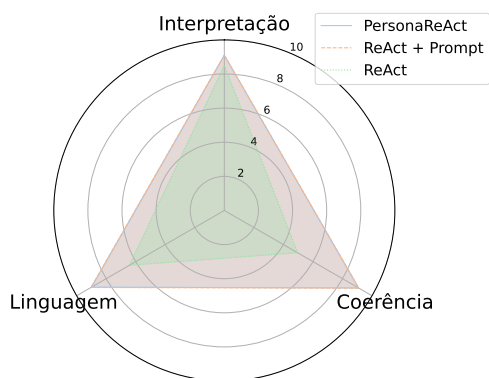
No conjunto analisado, observa-se que o modelo GPT-4o mini, sem capacidade explícita de raciocínio, apresenta maior irregularidade entre os critérios, especialmente no de Linguagem, onde o desempenho cai de forma mais acentuada em perfis que demandam maior adaptação contextual, como “Familiar de um Paciente” e “Pesquisador em Saúde Pública”. Isso sugere que a ausência de mecanismos de raciocínio interno pode limitar a capacidade de adequação a variações linguísticas mais sofisticadas, afetando a coerência estilística e a personalização.

Entre os modelos com raciocínio, o GPT-5 mini mantém resultados mais equilibrados e altos nos três critérios quando comparado ao GPT-5 nano. Este último, embora apresente desempenho consistente na maior parte dos perfis, mostra oscilações mais marcantes em contextos que exigem mudanças profundas no registro textual, indicando que modelos menores podem enfrentar mais dificuldades para preservar consistência estilística diante de personalizações mais complexas.

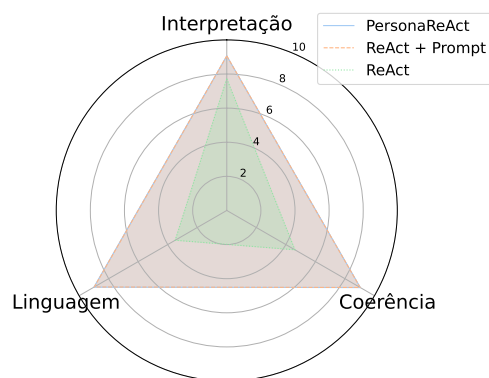
Quanto às abordagens, o PersonaReAct demonstra desempenho superior ao ReAct e, em boa parte dos casos, ao ReAct usando *Prompt* Enriquecido, sobretudo no GPT-4o mini e no GPT-5 nano, com ganhos claros em Coerência e Linguagem. No GPT-5 mini, entretanto, a diferença entre PersonaReAct e ReAct usando *Prompt* Enriquecido é menor, sugerindo que modelos mais robustos já conseguem, apenas com enriquecimento de *prompt*, atingir níveis elevados nos três critérios.

Assim, os resultados reforçam três conclusões principais: (i) modelos com raciocínio mantêm maior estabilidade e qualidade mesmo em perfis que exigem personalização intensa; (ii) modelos menores tendem a apresentar maior variação entre perfis, especialmente na consistência e na adaptação linguística; e (iii) a arquitetura especializada do PersonaReAct agrega mais

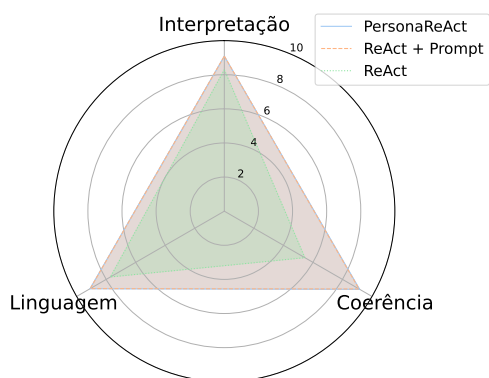
Figura 15 – Comparação das médias por critério entre ReAct, ReAct usando *Prompt* Enriquecido e PersonaReAct para o modelo GPT-5 mini no domínio médico.



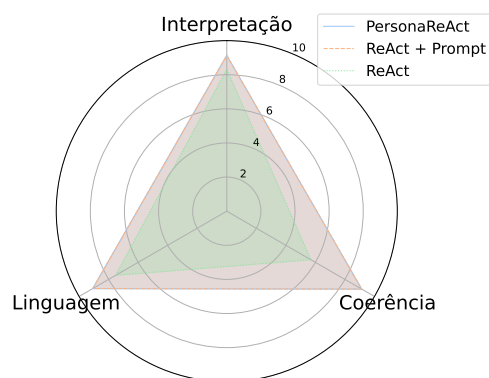
(a) Enfermeiro



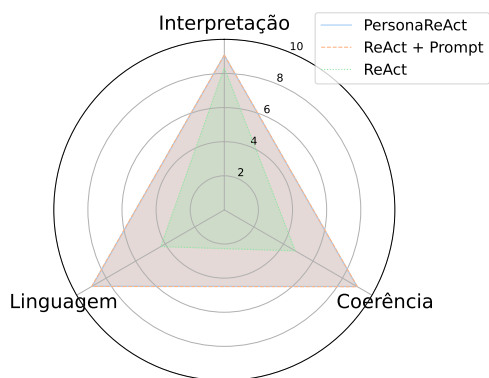
(b) Familiar de um Paciente



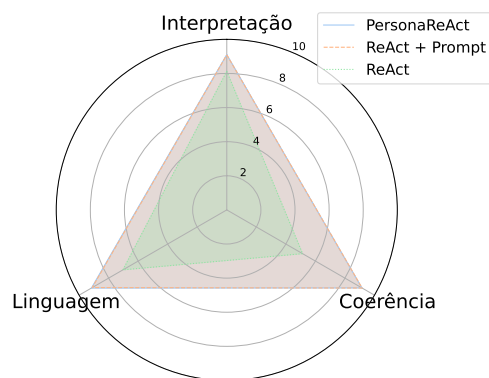
(c) Gestor Hospitalar



(d) Médico



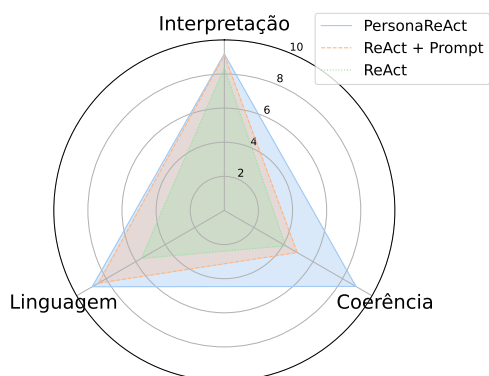
(e) Paciente



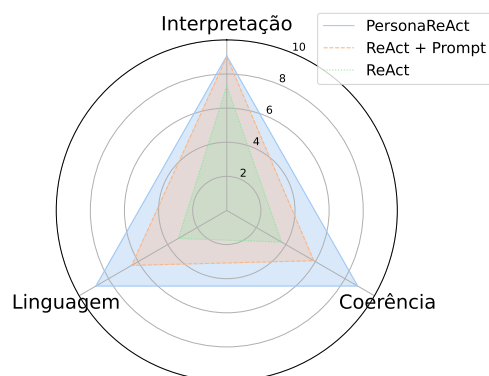
(f) Pesquisador em Saúde Pública

Fonte: elaborado pelo autor.

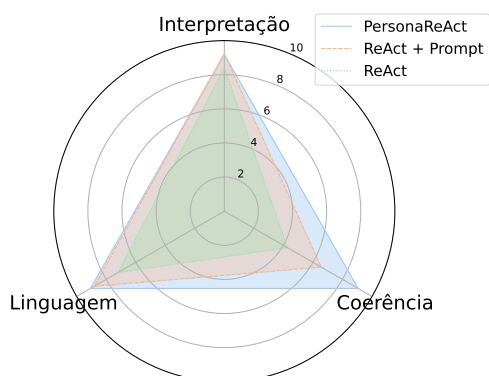
Figura 16 – Comparação das médias por critério entre ReAct, ReAct usando *Prompt* Enriquecido e PersonaReAct para o modelo GPT-5 nano no domínio médico.



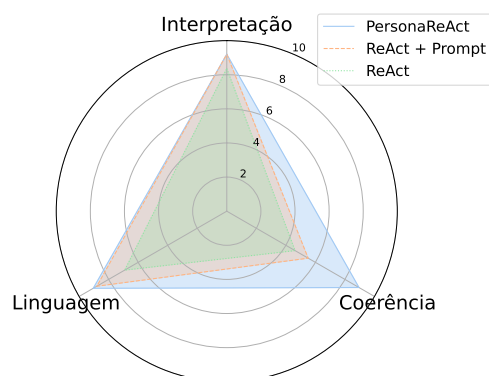
(a) Enfermeiro



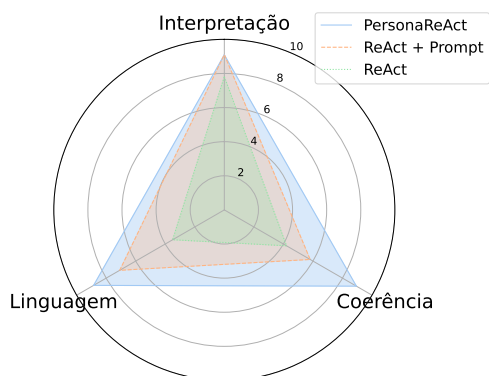
(b) Familiar de um Paciente



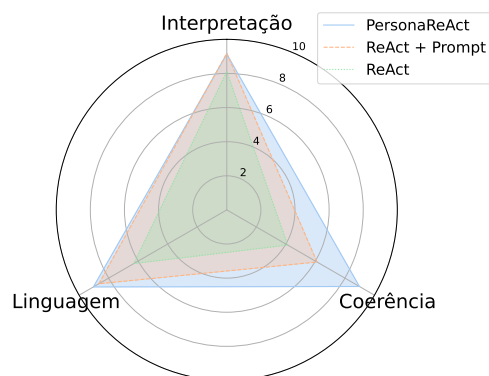
(c) Gestor Hospitalar



(d) Médico



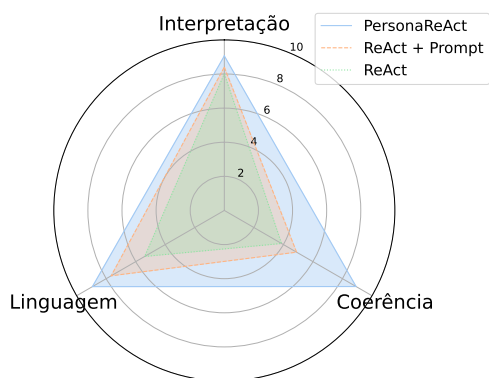
(e) Paciente



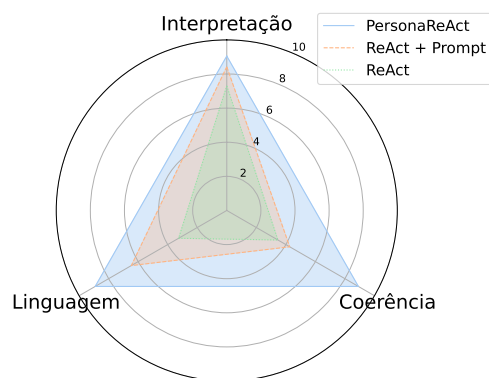
(f) Pesquisador em Saúde Pública

Fonte: elaborado pelo autor.

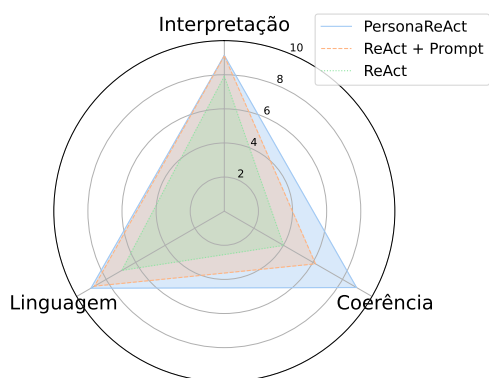
Figura 17 – Comparação das médias por critério entre ReAct, ReAct usando *Prompt* Enriquecido e PersonaReAct para o modelo GPT-4o mini no domínio médico.



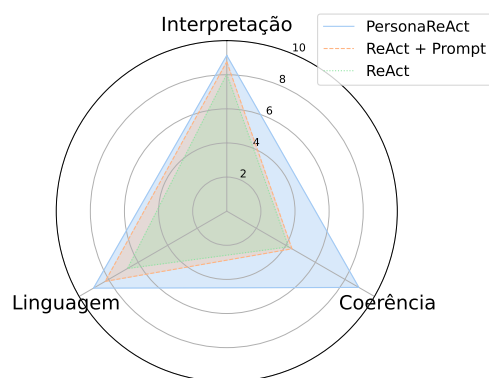
(a) Enfermeiro



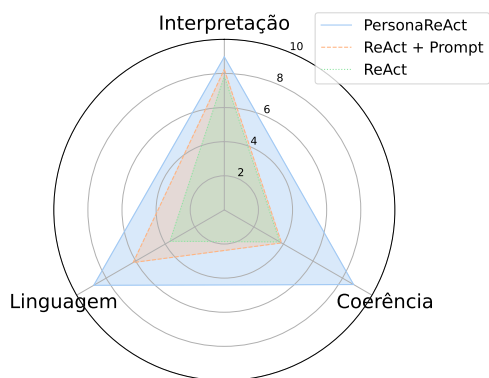
(b) Familiar de um Paciente



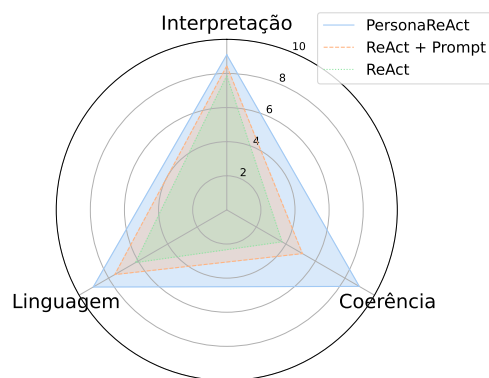
(c) Gestor Hospitalar



(d) Médico



(e) Paciente



(f) Pesquisador em Saúde Pública

Fonte: elaborado pelo autor.

valor em modelos menores ou sem raciocínio, sendo particularmente vantajosa em cenários com restrições computacionais e alta demanda por personalização.

### ***Discussão Sobre os Resultados do Experimento 1***

A comparação do desempenho do PersonaReAct com o ReAct usando *Prompt* Enriquecido revelou dois padrões distintos, dependendo se o modelo de linguagem utilizado possui ou não capacidades explícitas de raciocínio.

Nos modelos com raciocínio (o3-mini, o4-mini, GPT-5 mini e GPT-5 nano), a diferença entre as abordagens foi pequena. Isso ocorre porque, no *Prompt* Enriquecido, a instrução de personalização é incorporada antes do início do ciclo *Reason + Act*, o que significa que o perfil do usuário é internalizado desde o primeiro passo do raciocínio. Com isso, tanto as ações (*Acts*) quanto as observações (*Observations*) e o texto final já são naturalmente adaptados ao perfil. Como o raciocínio desses modelos é mais estruturado e segue múltiplos ciclos de reflexão, o perfil é mantido de forma consistente ao longo de toda a resposta. Assim, quando o módulo *Tailor* do PersonaReAct atua, o texto-base já apresenta um nível elevado de personalização, reduzindo o impacto marginal do pós-processamento.

Além disso, a literatura recente indica que modelos avançados com raciocínio tendem a apresentar maior capacidade de seguir instruções complexas. Fu *et al.* (2025), por meio do *benchmark* MathIF, mostram que *Large Reasoning Models* (LRMs) mantêm elevado desempenho em tarefas complexas justamente por integrar a interpretação das instruções ao fluxo de raciocínio (*chain-of-thought*). Qin *et al.* (2025) reforçam que esse processamento estruturado permite lidar melhor com demandas multifacetadas, enquanto He *et al.* (2024) observam que, diante de instruções com múltiplas restrições, modelos treinados para raciocínio cumprem essas condições com mais eficácia. No entanto, esses mesmos estudos apontam um *trade-off* importante: quanto mais profunda a capacidade de raciocínio, maior a probabilidade de o modelo priorizar a solução do problema em detrimento da obediência literal a instruções simples. É justamente nesse ponto que o desacoplamento proposto pelo PersonaReAct se torna relevante, pois preserva a cadeia lógica intacta e garante que a personalização final seja aplicada de forma controlada e consistente, independentemente da complexidade do raciocínio subjacente.

O cenário é diferente nos modelos sem raciocínio (GPT-4.1 mini, GPT-4.1 nano e GPT-4o mini), onde o PersonaReAct apresentou vantagem clara em praticamente todos os critérios avaliados nos dois estudos de caso. Sem um ciclo formal de raciocínio, esses modelos

tendem a tratar a instrução de personalização no *Prompt* Enriquecido como um ajuste superficial, muitas vezes aplicado apenas no início da resposta ou de forma inconsistente ao longo do texto. Isso compromete a Coerência do Estilo ao Longo do Texto, já que partes da resposta podem manter um tom neutro, enquanto outras refletem o perfil do usuário.

O PersonaReAct supera essa limitação por desacoplar o raciocínio da personalização. Após gerar a resposta-base pelo ReAct tradicional, o módulo *Tailor* processa o texto por completo com foco exclusivo no perfil, reforçando vocabulário, tom e nível de detalhamento. Esse pós-processamento garante personalização uniforme do início ao fim, elevando as notas não apenas em Coerência do Estilo e Linguagem e Estilo Adequados, mas também em Interpretação e Explicação dos Dados. Isso ocorre porque o *Tailor* pode inserir exemplos, analogias e interpretações contextualizadas, mesmo que a resposta-base seja genérica.

Os resultados quantitativos e gráficos apresentados nos Estudos de Caso 1 e 2 reforçam esses achados. Nos cenários sem raciocínio explícito, a aplicação do PersonaReAct resultou em médias mais altas e menor dispersão, enquanto nos modelos com raciocínio as diferenças foram mais sutis, mas ainda indicaram ganhos de consistência. Essa observação é consistente com a premissa de que a arquitetura proposta reduz as disparidades entre modelos de diferentes capacidades, garantindo um padrão de qualidade mais uniforme.

Com base nessa análise, podemos avaliar as Hipóteses 1, 3 e 4 de forma mais direta:

- Hipótese 1 (a personalização melhora a utilidade, clareza e interpretabilidade das respostas): em ambos os estudos de caso, o PersonaReAct apresentou ganhos claros e consistentes nos três critérios avaliados, com destaque para a melhoria na clareza e adequação da linguagem, independentemente da presença de raciocínio.
- Hipótese 3 (separação explícita da personalização melhora a organização das respostas): o desacoplamento permitiu um controle mais preciso sobre a aplicação da personalização, assegurando coerência estilística e consistência estrutural, sobretudo nos modelos sem raciocínio.
- Hipótese 4 (o tipo de modelo utilizado influencia o desempenho das diferentes abordagens de personalização): os modelos com raciocínio apresentaram diferenças menores entre abordagens, enquanto nos modelos sem raciocínio os ganhos com o PersonaReAct foram significativamente mais expressivos, evidenciando que a capacidade de raciocínio é um fator determinante para o impacto da personalização.

De forma resumida, o Experimento 1 demonstra que, embora modelos mais avança-

dos consigam internalizar e manter a personalização diretamente no fluxo lógico de raciocínio, a separação explícita proposta pelo PersonaReAct oferece melhores resultados em outros cenários. Esses resultados se mostram particularmente relevantes em cenários com modelos de menor capacidade ou ausência de raciocínio estruturado, nos quais o ganho de clareza, consistência e adaptação ao perfil do usuário é maior.

## 6.2 Experimento 2 - Variação entre Perfis

Nesta seção, são apresentados os resultados dos experimentos que investigam se uma mesma pergunta gera respostas significativamente diferentes quando o perfil do usuário é variado, conforme a metodologia descrita na Seção 5.5.

### *Análise de Similaridade Textual*

Para avaliar em que medida as respostas personalizadas diferem das respostas padrão em termos textuais, foram aplicadas três métricas amplamente utilizadas: ROUGE-L, BLEU e Distância de Edição. Essas métricas foram calculadas individualmente para cada par de respostas (*resposta personalizada vs. resposta padrão*), considerando a mesma pergunta como referência.

Neste experimento, foram selecionados os modelos o3 mini e GPT-4.1 mini por representarem, respectivamente, um modelo com capacidade avançada de raciocínio e outro sem essa habilidade, além de ambos terem apresentado bom desempenho no Experimento 1. Os resultados obtidos com os demais modelos foram omitidos devido a limitações de espaço. Essa análise refere-se especificamente ao estudo de caso no domínio esportivo. Já os gráficos correspondentes ao estudo de caso no domínio médico encontram-se no Apêndice E.

Considerando diferentes granularidades de comparação, a avaliação da similaridade textual foi conduzida em três níveis, apresentados a seguir.

### **Similaridade Média por Pergunta**

Neste nível, para cada pergunta, foram selecionadas todas as respostas personalizadas geradas para os diferentes perfis. Em seguida, cada uma dessas respostas foi comparada com a resposta padrão correspondente utilizando as três métricas. A média dos valores obtidos para cada métrica foi então computada. Essas médias permitem identificar quais perguntas geraram maior ou menor variação textual nas personalizações, independentemente do perfil. A Tabela

13 apresenta as médias e desvios padrão das métricas de similaridade textual entre as respostas personalizadas e suas respectivas versões neutras, organizadas por pergunta, para o modelo o3 mini. A Tabela 14 apresenta as mesmas características, mas para o modelo GPT-4.1 mini.

Tabela 13 – Médias e desvios padrão das métricas de similaridade textual por pergunta para o modelo o3 mini no contexto do conjunto de dados do domínio de futebol.

Pergunta	ROUGE-L (Média ± DP)	BLEU (Média ± DP)	Distância de Edição (Média ± DP)
P1	0,168 ± 0,053	0,002 ± 0,001	0,183 ± 0,024
P2	0,271 ± 0,069	0,026 ± 0,029	0,208 ± 0,037
P3	0,157 ± 0,040	0,008 ± 0,005	0,223 ± 0,016
P4	0,037 ± 0,043	0,000 ± 0,000	0,036 ± 0,014
P5	0,154 ± 0,037	0,001 ± 0,001	0,147 ± 0,027
P6	0,252 ± 0,106	0,020 ± 0,015	0,233 ± 0,050
P7	0,191 ± 0,044	0,020 ± 0,010	0,267 ± 0,036
P8	0,050 ± 0,033	0,000 ± 0,000	0,042 ± 0,010
P9	0,229 ± 0,080	0,013 ± 0,013	0,249 ± 0,063
P10	0,079 ± 0,025	0,014 ± 0,010	0,212 ± 0,019
P11	0,285 ± 0,144	0,035 ± 0,037	0,221 ± 0,085
P12	0,016 ± 0,040	0,000 ± 0,000	0,031 ± 0,013
P13	0,199 ± 0,075	0,007 ± 0,009	0,189 ± 0,048
P14	0,023 ± 0,026	0,000 ± 0,000	0,011 ± 0,004
P15	0,155 ± 0,049	0,002 ± 0,001	0,157 ± 0,031
P16	0,264 ± 0,083	0,021 ± 0,032	0,235 ± 0,067
P17	0,049 ± 0,041	0,000 ± 0,000	0,042 ± 0,014
P18	0,029 ± 0,034	0,000 ± 0,000	0,040 ± 0,010
P19	0,115 ± 0,032	0,003 ± 0,004	0,199 ± 0,040
P20	0,153 ± 0,058	0,002 ± 0,002	0,162 ± 0,042

Fonte: elaborado pelo autor.

A Tabela 13 apresenta os valores médios e desvios padrão das métricas de similaridade textual para o modelo o3 mini, enquanto a Tabela 14 exhibe os mesmos indicadores para o modelo GPT-4.1 mini. Cada métrica possui uma escala e sensibilidade próprias, de modo que não é adequado compará-las diretamente em termos absolutos. Assim, a análise deve ser conduzida observando separadamente o comportamento de cada métrica.

No caso do o3-mini, observam-se perguntas como P2 (“*Quem foi o jogador que mais deu sprints na partida contra o Ceará em 1º de abril de 2023?*”) que apresentam valores mais altos nas três métricas, sugerindo que, mesmo com a personalização, o conteúdo manteve proximidade com a resposta de referência, possivelmente pela objetividade da informação solicitada. Por outro lado, questões como P4 (“*Quem foi que ficou mais tempo parado naquele jogo contra o Ceará?*”), P8 (“*Quem do nosso time teve a velocidade mais instável durante o jogo contra o Ceará, variando mais ao longo da partida?*”), P14 (“*Quem foi o jogador mais estável na corrida, com a velocidade mais constante, no jogo contra o Ceará?*”) e P17 (“*Quem foi o jogador que mais acumulou eventos intensos, somando acelerações e desacelerações fortes*”

Tabela 14 – Médias e desvios padrão das métricas de similaridade textual por pergunta para o modelo gpt-4.1-mini no contexto do conjunto de dados do domínio de futebol.

Pergunta	ROUGE-L (Média ± DP)	BLEU (Média ± DP)	Distância de Edição (Média ± DP)
P1	0,251 ± 0,021	0,047 ± 0,015	0,226 ± 0,025
P2	0,253 ± 0,026	0,024 ± 0,020	0,167 ± 0,022
P3	0,137 ± 0,060	0,004 ± 0,006	0,187 ± 0,028
P4	0,178 ± 0,044	0,005 ± 0,007	0,143 ± 0,044
P5	0,199 ± 0,080	0,014 ± 0,014	0,159 ± 0,022
P6	0,209 ± 0,061	0,012 ± 0,015	0,178 ± 0,044
P7	0,207 ± 0,049	0,026 ± 0,022	0,287 ± 0,023
P8	0,168 ± 0,046	0,009 ± 0,009	0,204 ± 0,025
P9	0,194 ± 0,063	0,010 ± 0,009	0,197 ± 0,038
P10	0,141 ± 0,054	0,014 ± 0,006	0,263 ± 0,048
P11	0,200 ± 0,036	0,003 ± 0,002	0,168 ± 0,031
P12	0,160 ± 0,045	0,017 ± 0,010	0,238 ± 0,028
P13	0,201 ± 0,057	0,002 ± 0,002	0,155 ± 0,033
P14	0,170 ± 0,066	0,006 ± 0,005	0,172 ± 0,036
P15	0,126 ± 0,040	0,002 ± 0,001	0,145 ± 0,017
P16	0,232 ± 0,043	0,015 ± 0,012	0,246 ± 0,032
P17	0,174 ± 0,051	0,013 ± 0,007	0,248 ± 0,025
P18	0,239 ± 0,062	0,025 ± 0,020	0,246 ± 0,033
P19	0,131 ± 0,049	0,001 ± 0,001	0,149 ± 0,020
P20	0,146 ± 0,066	0,002 ± 0,004	0,156 ± 0,025

Fonte: elaborado pelo autor.

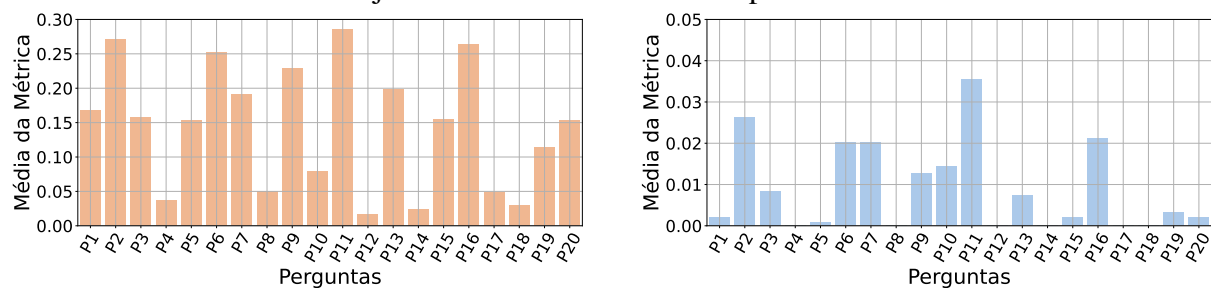
*no jogo contra o Ceará?*”) mostram valores baixos em todas as métricas, indicando grande variação lexical e estrutural entre perfis.

Para o gpt-4.1-mini, nota-se que algumas perguntas, como P1 (“*Qual foi a velocidade média que o Bruno Pacheco alcançou naquele clássico contra o Ceará no dia 1º de abril de 2023?*”), P4 e P18 (“*Qual jogador teve a maior diferença entre sua velocidade média e a média de todo o time naquele jogo contra o Ceará?*”), apresentam valores relativamente mais altos em BLEU e ROUGE-L do que no o3 mini, indicando maior aderência à forma textual da referência. Perguntas como P7 (“*Como foi a variação nas curvas que o Bruno Pacheco fez durante o jogo contra o Ceará? Teve muito desvio?*”), P10 (“*Na média, quanto tempo cada jogador passou correndo durante aquele jogo contra o Ceará?*”) e P16 (“*Qual foi a diferença entre o tempo que o Bruno Pacheco passou sprintando e o tempo que ele ficou parado durante o jogo contra o Ceará?*”) apresentam variação estrutural mais moderada, evidenciada pela Distância de Edição.

Para uma melhor visualização das tendências gerais, as Figuras 18 e 19 mostram as médias dessas métricas em formato gráfico.

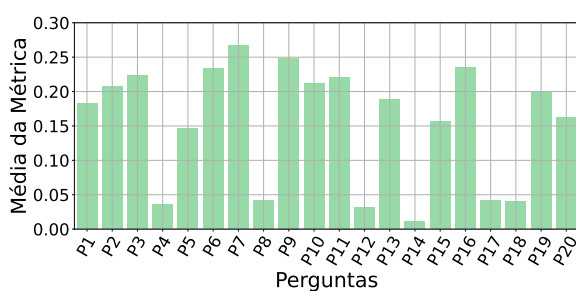
Em geral, os resultados indicam que o gpt-4.1-mini tende a gerar respostas mais próximas lexicalmente da referência em várias perguntas, enquanto o o3 mini apresenta maior dispersão e, em alguns casos, maior variação estilística entre perfis. Essas diferenças sugerem

Figura 18 – Gráfico de barras das métricas de similaridade textual por pergunta para o modelo o3 mini no contexto do conjunto de dados do domínio esportivo.



(a) ROUGE-L

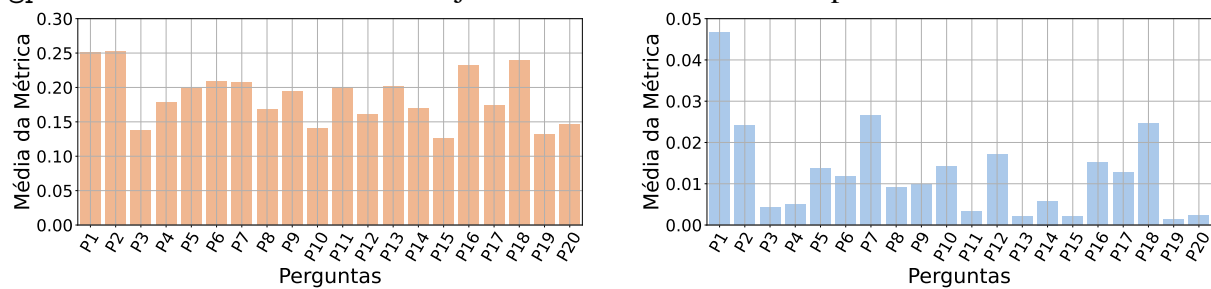
(b) BLEU



(c) Distância de Edição

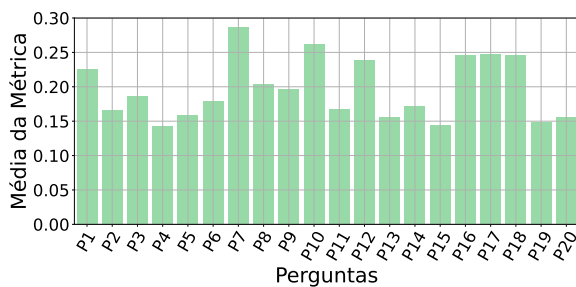
Fonte: elaborado pelo autor.

Figura 19 – Gráfico de barras das métricas de similaridade textual por pergunta para o modelo gpt-4.1-mini no contexto do conjunto de dados do domínio esportivo.



(a) ROUGE-L

(b) BLEU



(c) Distância de Edição

Fonte: elaborado pelo autor.

que o primeiro modelo tende a preservar mais a forma textual original, ao passo que o segundo se mostra mais propenso a reformulações criativas ou estruturais.

### Similaridade Média por Perfil

Neste experimento, a agregação é feita por perfil de usuário. Para cada perfil, todas as respostas personalizadas geradas foram comparadas com suas respectivas versões padrão (mantendo a mesma pergunta de referência) e, para cada métrica, foi calculada a média. Essa análise revela quais perfis geraram personalizações mais ou menos distantes das versões originais.

A Tabela 15 apresenta as médias e desvios padrão das métricas de similaridade textual entre as respostas personalizadas e suas respectivas versões neutras, organizadas por perfil, para o modelo o3 mini. A Tabela 16 apresenta as mesmas características, mas para o modelo GPT-4.1 mini.

Tabela 15 – Médias e desvios padrão das métricas de similaridade textual por perfil para o modelo o3 mini no contexto do conjunto de dados do domínio de futebol.

Perfil	ROUGE-L (Média ± DP)	BLEU (Média ± DP)	Edit Distance (Média ± DP)
Fisioterapeuta	0,110 ± 0,086	0,004 ± 0,007	0,127 ± 0,084
Influenciador	0,138 ± 0,101	0,006 ± 0,010	0,157 ± 0,091
Jornalista	0,166 ± 0,119	0,010 ± 0,018	0,155 ± 0,089
Narrador	0,104 ± 0,067	0,003 ± 0,004	0,141 ± 0,079
Técnico	0,178 ± 0,131	0,019 ± 0,029	0,178 ± 0,109
Torcedor	0,167 ± 0,105	0,010 ± 0,015	0,168 ± 0,099

Fonte: elaborado pelo autor.

Tabela 16 – Médias e desvios padrão das métricas de similaridade textual por perfil para o modelo GPT-4.1 mini no contexto do conjunto de dados do domínio de futebol.

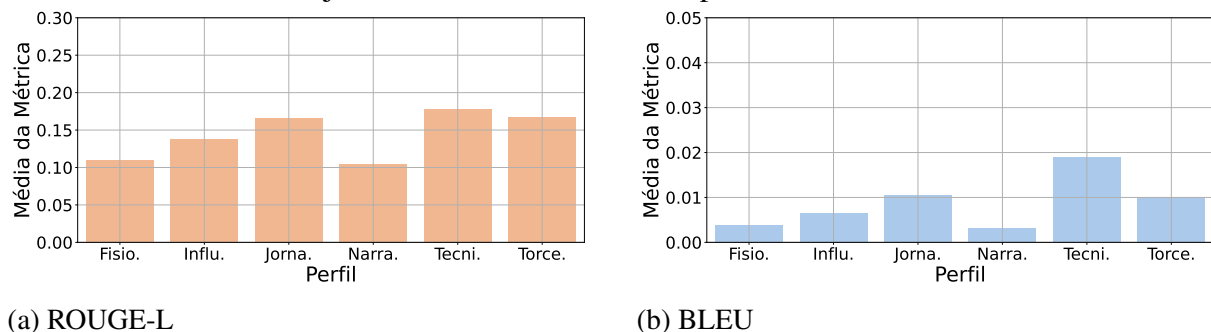
Perfil	ROUGE-L (Média ± DP)	BLEU (Média ± DP)	Edit Distance (Média ± DP)
Fisioterapeuta	0,165 ± 0,052	0,007 ± 0,008	0,174 ± 0,059
Influenciador	0,161 ± 0,054	0,009 ± 0,014	0,196 ± 0,059
Jornalista	0,209 ± 0,062	0,016 ± 0,017	0,198 ± 0,045
Narrador	0,168 ± 0,056	0,010 ± 0,017	0,194 ± 0,050
Técnico	0,213 ± 0,073	0,019 ± 0,018	0,207 ± 0,052
Torcedor	0,199 ± 0,057	0,015 ± 0,013	0,211 ± 0,045

Fonte: elaborado pelo autor.

As Figuras 20 e 21 apresentam a média das métricas ROUGE-L, BLEU e Edit Distance entre as respostas personalizadas e suas respectivas versões padrão, agora agrupadas de acordo com o perfil do usuário para os modelos o3 mini e GPT-4.1 mini, respectivamente. O objetivo dessa análise é identificar quais perfis geram respostas mais próximas, em termos

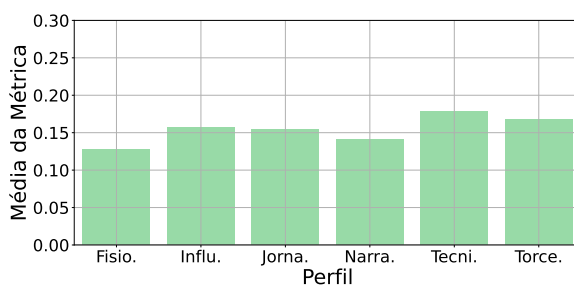
textuais, das versões padrão (ou seja, menos afetadas pelas modificações de personalização).

Figura 20 – Gráfico de barras das métricas de similaridade textual por perfil para o modelo o3 mini no contexto do conjunto de dados do domínio esportivo.



(a) ROUGE-L

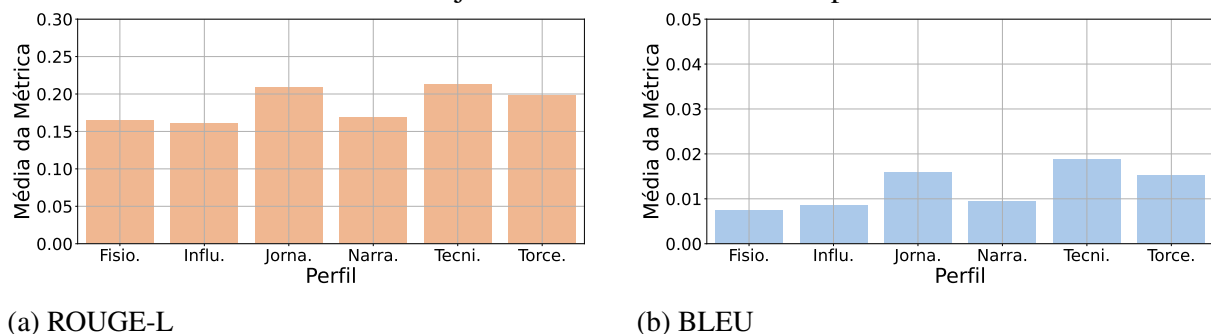
(b) BLEU



(c) Distância de Edição

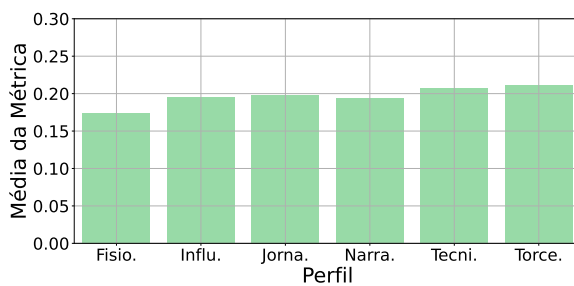
Fonte: elaborado pelo autor.

Figura 21 – Gráfico de barras das métricas de similaridade textual por perfil para o modelo GPT-4.1 mini no contexto do conjunto de dados do domínio esportivo.



(a) ROUGE-L

(b) BLEU



(c) Distância de Edição

Fonte: elaborado pelo autor.

A análise conjunta das Tabelas 15 e 16 e das Figuras 20 e 21, considerando que aqui desejamos afastamento em relação à resposta neutra (ou seja, ROUGE-L e BLEU baixos e distância de edição alta indicam maior diferenciação), revela um padrão coerente por perfil. Os perfis “Narrador de Jogo de Futebol” e “Influenciador Digital” tendem a apresentar menor sobreposição com a resposta neutra (ROUGE-L e BLEU mais baixos), refletindo estilos mais livres e expressivos; já os perfis “Técnico do Time” e, em seguida, “Jornalista Esportivo” permanecem mais próximos do texto base (valores relativamente mais altos de ROUGE-L/BLEU), o que é compatível com registros mais objetivos e informativos. No o3 mini, por exemplo, ROUGE-L cai para 0,104 no perfil “Narrador de Jogo de Futebol” e 0,138 no “Influenciador Digital”, enquanto sobe para 0,178 no perfil “Técnico do Time”; cenário análogo aparece no GPT-4.1 mini (0,168/0,161 vs. 0,213).

Comparando os modelos, o GPT-4.1 mini exhibe médias mais altas de ROUGE-L e BLEU e menor dispersão em todos os perfis, indicando maior proximidade e homogeneidade em relação à resposta neutra. Já o o3 mini apresenta ROUGE-L e BLEU sistematicamente menores (por exemplo, no perfil “Técnico do Time”, 0,178 vs. 0,213; no perfil “Narrador de Jogo de Futebol”, 0,104 vs. 0,168), sinalizando maior grau de reescrita e, portanto, diferenciação textual mais intensa (exatamente o comportamento desejado quando se busca personalização que se afaste do texto base). Em resumo, as tabelas e figuras mostram que: (i) perfis mais expressivos (“Narrador de Jogo de Futebol” e “Influenciador Digital”) se distanciam mais da resposta neutra, enquanto perfis técnicos se mantêm mais próximos; e (ii) entre os modelos comparados, o o3 mini favorece maior divergência estilística, ao passo que o GPT-4.1 mini preserva uma proximidade maior com o tom neutro. Essa diferença de comportamento é útil para orientar a escolha do modelo conforme a ênfase desejada entre estabilidade e grau de personalização.

### **Distribuição Geral das Métricas**

Por fim, foi feita uma análise agregada da distribuição de todas as comparações realizadas, sem distinção por pergunta ou perfil. Cada par (resposta personalizada, resposta padrão) foi submetido ao cálculo das três métricas, e os valores foram representados em um *boxplot*, permitindo observar a variação, tendência central e presença de *outliers*. Essa abordagem ajuda a compreender o comportamento geral das transformações textuais promovidas pelo sistema. As Tabelas 17 e 18 apresentam os valores obtidos para os modelos o3 mini e GPT-4.1 mini, respectivamente.

Tabela 17 – Distribuição geral das métricas de similaridade textual entre respostas personalizadas e padrão para o modelo o3 mini no domínio esportivo.

Métrica	Média	DP	Mínimo	Q1 (25%)	Mediana (50%)	Q3 (75%)	Máximo
ROUGE-L	0,1439	0,1052	0,0000	0,0693	0,1316	0,2073	0,4545
BLEU	0,0088	0,0166	0,0000	0,0000	0,0026	0,0114	0,0859
Edit Distance	0,1544	0,0918	0,0040	0,0480	0,1747	0,2216	0,3731

Fonte: elaborado pelo autor.

Tabela 18 – Distribuição geral das métricas de similaridade textual entre respostas personalizadas e padrão (modelo GPT-4.1 mini) no domínio esportivo.

Métrica	Média	DP	Mínimo	Q1 (25%)	Mediana (50%)	Q3 (75%)	Máximo
ROUGE-L	0,1858	0,0619	0,0750	0,1331	0,1806	0,2324	0,3210
BLEU	0,0126	0,0152	0,0001	0,0017	0,0059	0,0190	0,0700
Edit Distance	0,1966	0,0522	0,0862	0,1529	0,1874	0,2373	0,3262

Fonte: elaborado pelo autor.

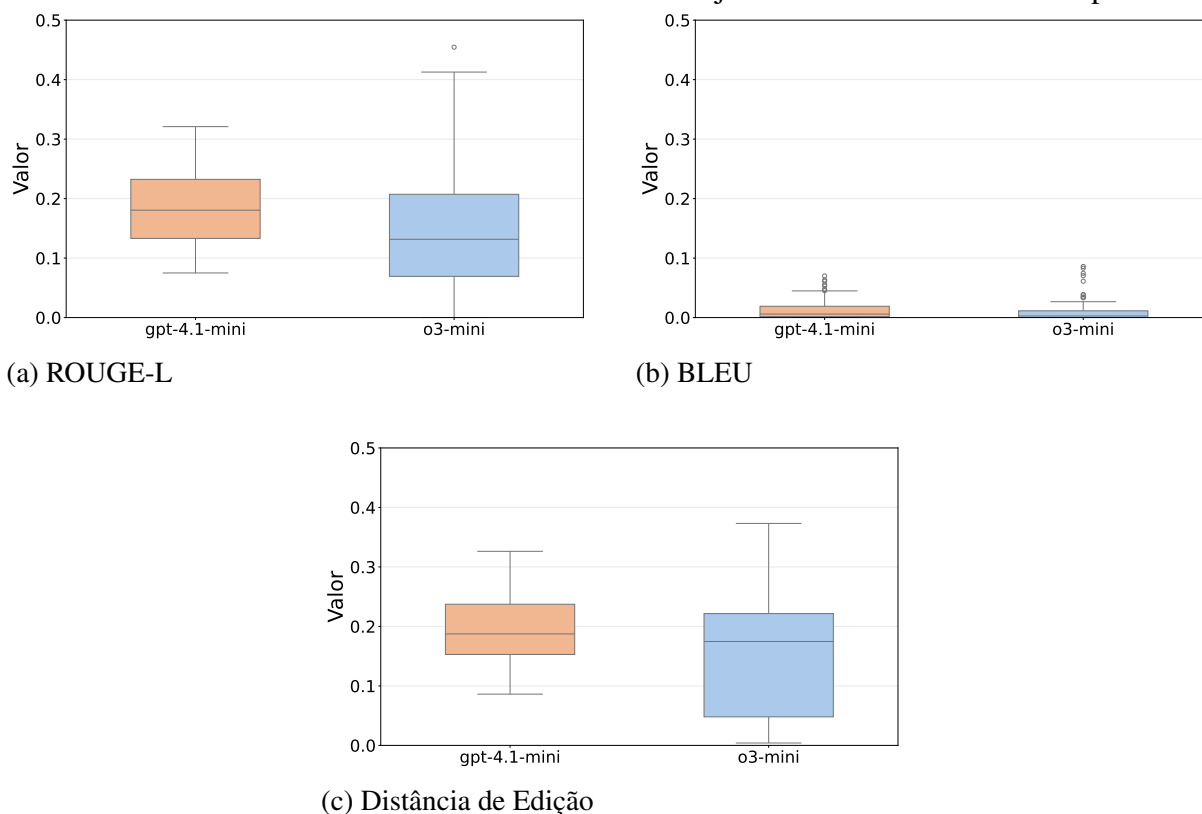
A Figura 22 apresenta as distribuições gerais das métricas ROUGE-L, BLEU e Distância de Edição entre todas as respostas personalizadas e suas respectivas versões padrão, independentemente da pergunta ou perfil, para os modelos o3 mini e GPT-4.1 mini. A visualização em *boxplot* permite identificar padrões de centralidade, dispersão e presença de valores atípicos (*outliers*).

As Tabelas 17 e 18, em conjunto com os *boxplots* da Figura 22, sintetizam a distribuição global das três métricas de similaridade (ROUGE-L, BLEU e Distância de Edição) para todas as comparações entre resposta personalizada e resposta padrão. Considerando que, neste experimento, buscamos afastamento da resposta neutra, interpretamos ROUGE-L/BLEU baixos como maior diferenciação e distância de edição alta como maior divergência.

Considerando o ROUGE-L, observa-se que o GPT-4.1 mini mantém maior proximidade lexical com a resposta base: mediana de 0,1806 (IQR 0,1331–0,2324) contra 0,1316 (0,0693–0,2073) do o3 mini. Em BLEU (cujos valores absolutos permanecem baixos, como esperado quando há reescritas de estilo) a mediana também é um pouco mais alta no GPT-4.1 mini (0,0059 vs. 0,0026 no o3 mini), reforçando uma tendência a respostas mais literais e homogêneas. Na Distância de Edição, as medianas são próximas (0,1874 no GPT-4.1 mini e 0,1747 no o3 mini), mas os perfis de dispersão diferem: o o3 mini apresenta Q1 muito baixo (0,0480), indicando casos de edição mínima ao lado de reescritas mais profundas, enquanto o GPT-4.1 mini concentra-se em patamares moderados de alteração (Q1=0,1529), com menos ocorrências próximas de zero.

Em conjunto, as tabelas e figuras mostram uma direção útil para a escolha do

Figura 22 – Distribuição Geral das Métricas (ROUGE-L, BLEU, Distância de Edição) para os modelos o3 mini e GPT-4.1 mini no contexto do conjunto de dados do domínio esportivo.



Fonte: elaborado pelo autor.

modelo: o3 mini tende a produzir personalizações mais distantes do texto neutro (ROUGE-L/BLEU menores e maior variabilidade), enquanto o GPT-4.1 mini entrega respostas mais próximas e consistentes da base (ROUGE-L/BLEU mais altos e dispersão menor). Assim, quando a prioridade é maximizar a diferenciação estilística entre perfis, o o3 mini mostra-se mais alinhado; quando a meta é preservar fidelidade lexical e estabilidade, o GPT-4.1 mini se revela a opção mais adequada.

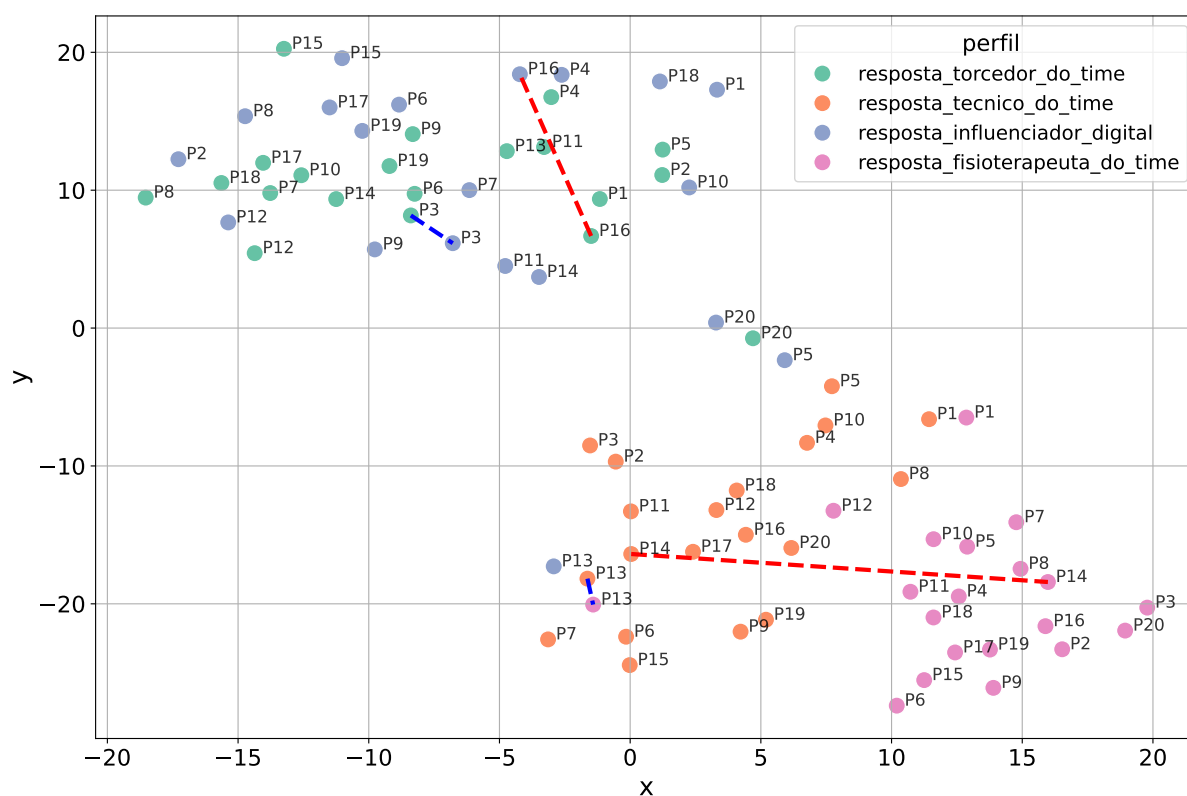
### ***Análise de Similaridade Semântica***

#### *Estudo de Caso 1: Domínio Esportivo*

Com o objetivo de avaliar a consistência semântica entre as respostas geradas para diferentes perfis, foi realizada uma análise de similaridade textual utilizando *embeddings* derivados do modelo `all-mpnet-base-v2` da biblioteca Sentence Transformers. Esse modelo transforma cada resposta textual em um vetor denso de alta dimensionalidade que captura relações semânticas latentes.

Para facilitar a visualização dos dados, foi aplicada uma redução de dimensionalidade com t-SNE (*t-distributed Stochastic Neighbor Embedding*), que projeta os vetores em duas dimensões, preservando relações de vizinhança local. A Figura 23 apresenta o resultado dessa projeção para os quatro perfis selecionados: *Torcedor do Time*, *Influenciador Digital*, *Técnico do Time* e *Fisioterapeuta do Time* para o modelo GTP-4.1 mini, enquanto a Figura 24 apresenta o resultado dessa projeção para os mesmos perfis, mas para o modelo o3 mini. Cada ponto corresponde à resposta de um perfil para uma pergunta específica, identificada por um código (por exemplo, P1, P2, etc.).

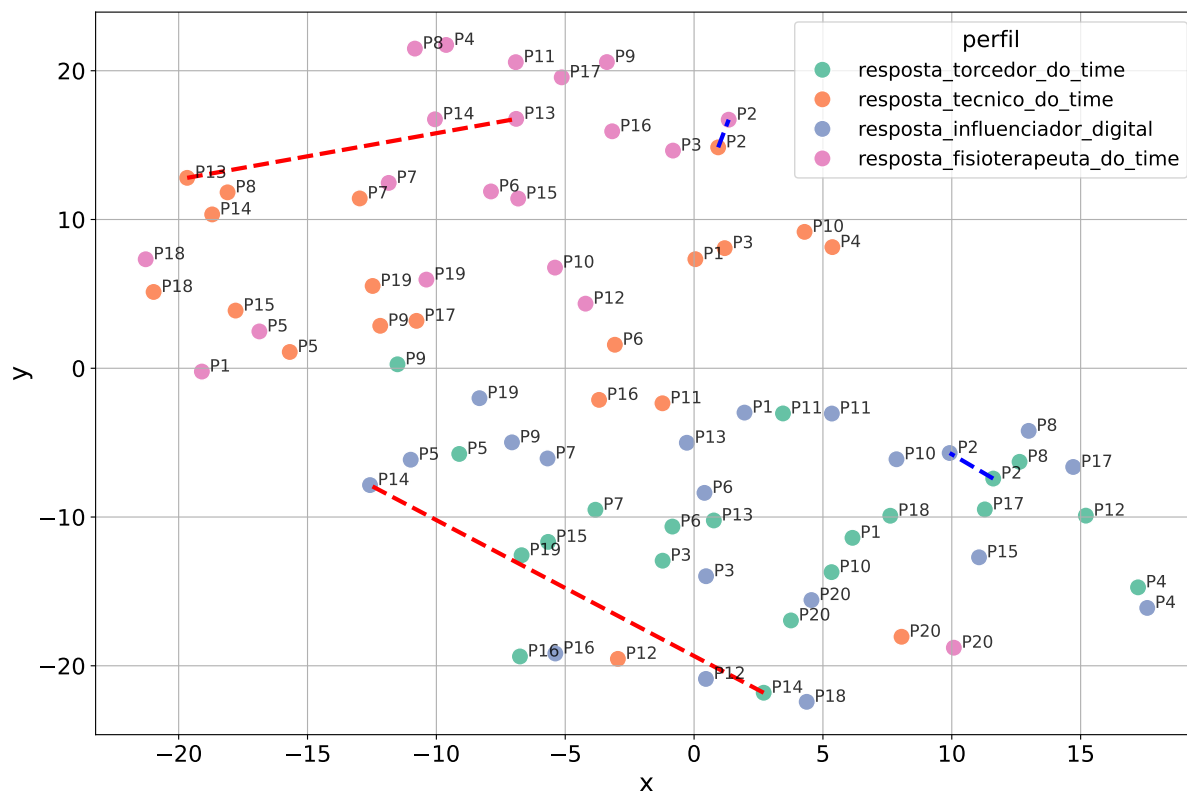
Figura 23 – Clusterização das respostas com destaque para pares de respostas mais próximas (azul) e mais distantes (vermelho) em cada *cluster* (modelo GTP-4.1 mini) no domínio esportivo.



Fonte: elaborado pelo autor.

Verifica-se, por inspeção visual, a formação de dois agrupamentos principais no espaço vetorial de ambas as figuras: um envolvendo os perfis “Torcedor do Time” e “Influenciador Digital”, e outro reunindo “Técnico do Time” e “Fisioterapeuta do Time”. Cada um desses *clusters* apresenta uma dispersão interna relativamente homogênea, indicando que as respostas de perfis mais semelhantes entre si mantiveram proximidade semântica. Além disso, observa-se

Figura 24 – Clusterização das respostas com destaque para pares de respostas mais próximas (azul) e mais distantes (vermelho) em cada *cluster* (modelo o3 mini) no domínio esportivo.



Fonte: elaborado pelo autor.

uma separação clara entre os dois agrupamentos, o que reforça a distinção semântica entre os estilos de comunicação dos perfis mais leigos (“Torcedor do Time” e “Influenciador Digital”) e os mais técnicos (“Técnico do Time” e “Fisioterapeuta do Time”).

Para quantificar essas observações, foram calculadas as distâncias de cosseno entre os *embeddings* das respostas de perfis diferentes para uma mesma pergunta.

A partir dessas distâncias, também foi possível identificar os pares de respostas mais próximas e mais distantes dentro de cada agrupamento. Nas Figuras 23 e 24, as ligações azuis destacam os pares de respostas semanticamente mais próximas, enquanto as vermelhas indicam os pares mais distantes.

### Análise Qualitativa dos Extremos

Para a análise qualitativa dos extremos, vamos considerar as respostas geradas com o modelo GPT-4.1 mini.

**Cluster Técnico–Fisioterapeuta:** a maior distância semântica foi observada na

resposta à pergunta “Qual atleta apresentou a menor variabilidade de velocidade (maior constância) durante a partida contra o Ceará?”:

- **Técnico do Time:** *Após análise detalhada, identificamos que, durante a partida contra o Ceará, o atleta Titi apresentou o menor coeficiente de variação na velocidade, com um índice médio de aproximadamente 0,326. Isso indica que sua velocidade foi a mais constante entre os jogadores avaliados nessa partida, evidenciando estabilidade no desempenho dinâmico. Essa constância pode ser relevante para a manutenção da estrutura tática e eficiência defensiva da equipe durante o jogo.*
- **Fisioterapeuta do Time:** *O jogador que apresentou a maior estabilidade em sua velocidade ao longo da partida foi João Ricardo, com uma velocidade média de 0.536 m/s. Essa característica sugere um bom controle físico e uma melhor capacidade de evitar sobrecargas, o que é crucial para a prevenção de lesões e para a manutenção de um desempenho consistente durante os 90 minutos. Fique atento a esses dados, pois eles podem ser significativos para a nossa análise na próxima avaliação física do time.*

No cluster “Técnico–Fisioterapeuta”, a maior distância semântica pode ser resultado de três pontos principais. A resposta do perfil “Técnico do Time” operacionaliza “menor variabilidade” em termos estritamente estatísticos, identificando Titi como o atleta com menor coeficiente de variação da velocidade ( $CV \approx 0,326$ ) e conectando o achado a implicações táticas, como manutenção da estrutura e eficiência defensiva. Já a resposta do perfil “Fisioterapeuta do Time” interpreta “estabilidade” sob um olhar fisiológico, reportando a velocidade média de João Ricardo ( $\approx 0,536$  m/s) como evidência de controle de carga e prevenção de lesões ao longo dos 90 minutos, além de sugerir acompanhamento em avaliações futuras. Assim, mudam simultaneamente a métrica e o conceito (CV vs. média), a entidade central (Titi vs. João Ricardo) e a finalidade discursiva (tática coletiva vs. saúde e recuperação). Esse deslocamento de método, jogador e objetivo analítico afasta o núcleo informativo de cada texto e explica a maior distância observada entre as respostas.

A menor distância semântica, por sua vez, ocorreu na resposta à pergunta “O coeficiente de variação da velocidade do Bruno Pacheco foi maior do que o do restante da equipe nesse jogo?”:

- **Técnico do Time:** *Analisando os dados do jogo, o coeficiente de variação da velocidade do Bruno Pacheco foi de aproximadamente 0,412, valor que indica a dispersão relativa da velocidade durante a partida. No entanto, este índice não foi superior ao coeficiente*

*médio registrado pelo restante da equipe. Portanto, do ponto de vista estatístico, sua variabilidade na velocidade foi menor ou igual à média do time, o que sugere uma performance mais consistente em termos de dinâmica de movimentação durante o jogo.*

- **Fisioterapeuta do Time:** *O coeficiente de variação da velocidade do Bruno Pacheco, que é de aproximadamente 0.39, é, na verdade, inferior aos dos outros jogadores analisados. Por exemplo, o jogador com o maior coeficiente, Habraão, apresentou um índice de 0.44, seguido por Junior Santos com 0.42. Essa comparação sugere que Bruno não está apresentando um padrão locomotor instável em relação aos demais atletas do time. Na verdade, sua variabilidade de velocidade é menor, o que pode ser um indicativo de maior consistência em suas performances. Vale considerar que essa estabilidade pode ser vantajosa para a sua participação em campo.*

No *cluster* “Técnico–Fisioterapeuta”, a menor distância semântica aparece porque as duas respostas convergem quase ponto a ponto: ambas operam com a mesma métrica (coeficiente de variação da velocidade), avaliam o mesmo jogador (Bruno Pacheco) e respondem à mesma comparação (frente ao restante da equipe) com a mesma conclusão: o CV de Bruno não é superior ao do time, indicando maior consistência de movimentação. As pequenas diferenças residem no enquadramento: o perfil “Técnico do Time” descreve o resultado de forma estatística e tática (consistência para a dinâmica de jogo), enquanto o perfil “Fisioterapeuta do Time” adota um olhar fisiológico e adiciona uma validação numérica com exemplos de colegas (Habraão e Júnior Santos). Esses acréscimos, contudo, não alteram o núcleo informativo; ao contrário, reforçam a mesma tese com léxico e estrutura muito próximos (“coeficiente de variação”, “velocidade”, “média da equipe”, “consistência”). Por isso, apesar de nuances de ênfase, o conteúdo, a entidade central e a inferência final são praticamente idênticos, o que explica a baixa distância entre os vetores das duas respostas.

**Cluster Torcedor–Influenciador Digital:** o maior afastamento semântico foi identificado na pergunta “Qual foi a diferença entre o tempo que o Bruno Pacheco passou sprintando e o tempo que ele ficou parado durante o jogo contra o Ceará?”.

- **Torcedor do Time:** *O Bruno Pacheco deu aquele gás na velocidade só de vez em quando, mas ficou mais tempo na marca, esperando a jogada desenrolar. Ou seja, apesar de ter mostrado pique nas arrancadas, ele passou bem mais tempo parado do que acelerando naquela partida contra o Ceará. Isso mostra que ele sabia dosar o esforço, economizando energia pra quando o time realmente precisava dele em alta velocidade.*

- **Influenciador Digital:** *E aí, tudo bem? Olha só, analisando o jogo do Bruno Pacheco, ele ficou voando em campo por um tempo bem maior do que parado. Especificamente, o tempo médio que ele passou se movimentando de forma mais ativa foi cerca de 9.9 unidades de tempo a mais do que o que ele ficou parado. Isso mostra a intensidade e o dinamismo da participação dele no jogo, sabe? É o tipo de dado que dá pra usar tranquilamente pros seus conteúdos, destacando como ele se mantém ligado e agilíssimo durante a partida, o que é ótimo para engajar seu público com insights legais e números que fazem sentido de verdade.*

No *cluster* “Torcedor–Influenciador Digital”, a maior distância semântica pode ser resultado de três pontos principais. Primeiro, há conclusões opostas sobre o comportamento de Bruno Pacheco: o perfil “Torcedor do Time” afirma que ele passou “bem mais tempo parado do que acelerando”, enquanto o perfil “Influenciador Digital” sustenta que ele ficou mais tempo em movimento, chegando a quantificar uma diferença de 9,9 “unidades de tempo” em favor do deslocamento ativo. Segundo, verifica-se um desalinhamento de métrica/alvo: o perfil “Torcedor do Time” responde exatamente à diferença entre tempo sprintando e tempo parado, ao passo que o perfil “Influenciador Digital” troca o foco para “movimento mais ativo” versus parado (uma categoria mais ampla e ambígua do que “sprintar”) além de recorrer a uma unidade temporal não especificada, o que reforça a divergência conceitual. Terceiro, os textos operam com propósitos discursivos e registros muito distintos: o perfil “Torcedor do Time” narra em tom afetivo e interno ao jogo (“deu aquele gás”, “soube dosar o esforço”), enquanto o perfil “Influenciador Digital” fala diretamente ao criador de conteúdo (“E aí, tudo bem?”), enfatizando engajamento, “insights” e utilidade para o público, com linguagem promocional e metacomunicação. A combinação de polos semânticos contraditórios (parado > sprintando vs. ativo > parado), mudança de variável (sprint  $\neq$  “movimento ativo”) e diferença de estilo/pragmática (torcida vs. marketing de conteúdo) desloca fortemente o núcleo informativo de cada resposta, explicando a distância vetorial elevada observada para esse par de perfis.

A menor distância foi verificada na pergunta “Durante aquele jogo contra o Ceará, quanto tempo o Bruno Pacheco passou correndo no gás, em sprint?”.

- **Torcedor do Time:** *Cara, naquele jogo contra o Ceará, o Bruno Pacheco realmente deu um show de disposição! Ele ficou numa pegada frenética, sprintando por uns bons momentos — tipo uns 20 segundos em alta velocidade, o que no futebol é uma baita arrancada. Dá pra dizer que o cara correu no gás, mostrando todo o sangue e raça que a*

*gente espera de um guerreiro em campo. Foi desses lances que fazem a diferença, sabe? Bruno não ficou parado um segundo, foi pura correria e entrega!*

- **Influenciador Digital:** *E aí, tudo certo? Sobre o Bruno Pacheco na partida contra o Ceará, ele ficou no modo turbo por cerca de 14 segutinhos sprintando no campo – é tipo aquele pique que acelera tudo, sabe? Não é um tempo longo, mas mostrar esse estalo de velocidade pode fazer toda a diferença no jogo, principalmente para levantar a galera e abrir espaço. Se quiser, posso te ajudar a montar uns posts legais explorando esses dados!*

No *cluster* “Torcedor–Influenciador Digital”, a menor distância semântica decorre da forte convergência de conteúdo e de objetivo informacional. As duas respostas tratam exatamente da mesma variável (o tempo de sprint do Bruno Pacheco no jogo contra o Ceará), usam a mesma unidade temporal (segundos) e chegam a estimativas de mesma ordem de grandeza ( $\approx 20s$  vs.  $\approx 14s$ ). Além disso, ambas constroem a mesma leitura de impacto contextual: sprint como “modo turbo/pegada frenética” que “faz diferença” ao imprimir intensidade e abrir espaço. Ou seja, entidade central, evento, métrica e inferência coincidem. As diferenças ficam quase todo no registro: o Torcedor adota um tom emocional, com hipérboles (“não ficou parado um segundo”, “sangue e raça”), enquanto o Influenciador fala com apelo de conteúdo e engajamento (diminutivos, convite a “montar posts”). Como esse desvio é predominantemente estilístico (sem troca de variável, sem mudança de comparação e sem introdução de novos atores) o núcleo semântico permanece praticamente idêntico. Por isso, mesmo com nuances de linguagem e uma pequena discrepância numérica, o vetor de significado das respostas se mantém muito próximo, explicando a baixa distância observada para esse par.

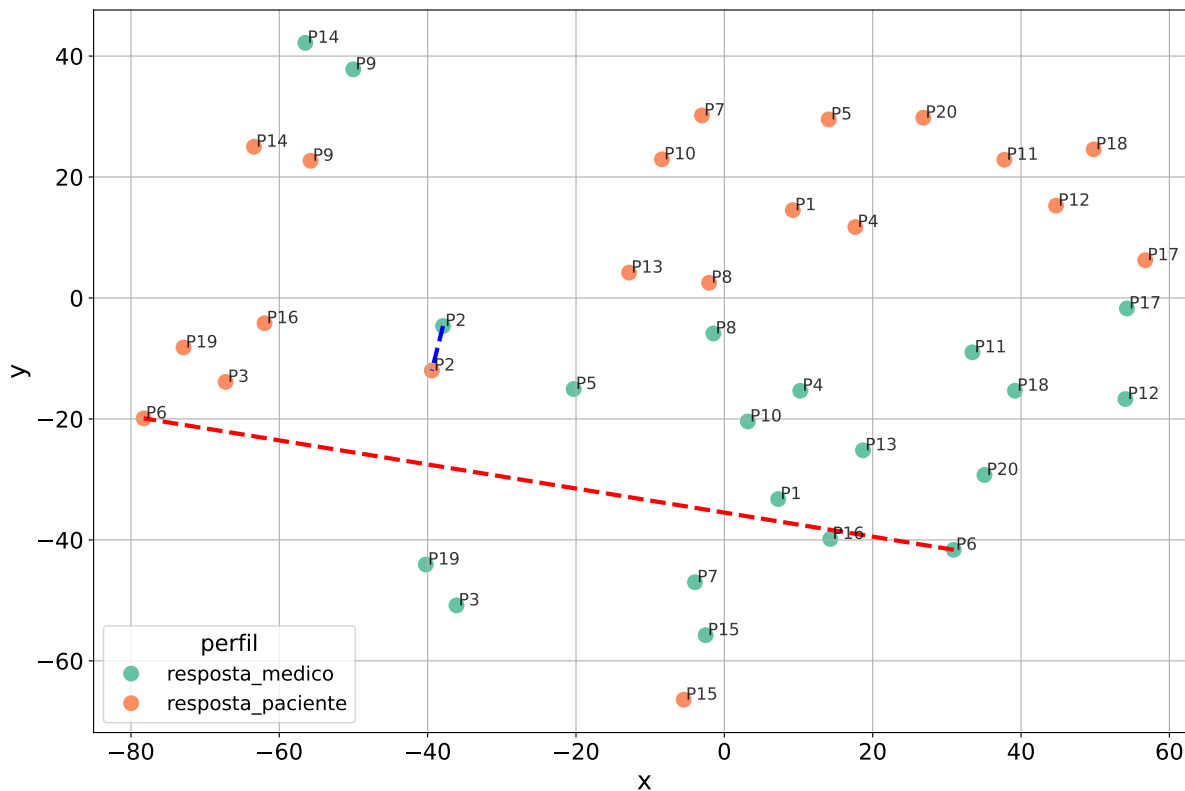
A análise dos *embeddings* mostrou que os pares de perfis analisados (Técnico com Fisioterapeuta e Torcedor com Influenciador) tendem a gerar respostas mais parecidas entre si. Isso acontece porque, apesar das diferenças de tom, os perfis de cada par compartilham objetivos comunicativos semelhantes. Por exemplo, tanto o Técnico quanto o Fisioterapeuta se preocupam com dados técnicos e desempenho físico, enquanto o Torcedor e o Influenciador valorizam o impacto emocional e a construção de uma narrativa envolvente. Esses pontos em comum ajudam a explicar as menores distâncias observadas nos pares destacados em azul na Figura 24. Assim, mesmo com estilos diferentes, os perfis de cada par acabam se aproximando semanticamente, o que reforça a ideia de que a intenção da mensagem influencia mais do que apenas a forma como é escrita.

### Estudo de Caso 2: Domínio Médico

Para examinar a coerência semântica das respostas fornecidas pelos diferentes perfis, conduziu-se uma análise de similaridade textual a partir de *embeddings* gerados com o modelo *all-mpnet-base-v2*, disponibilizado pela biblioteca *Sentence Transformers*. Esse modelo projeta cada resposta em um vetor denso de alta dimensionalidade, capaz de representar relações semânticas subjacentes e permitir comparações mais precisas entre os textos.

Para facilitar a visualização dos dados, foi aplicada uma redução de dimensionalidade com t-SNE (*t-distributed Stochastic Neighbor Embedding*), que projeta os vetores em duas dimensões, preservando relações de vizinhança local. A Figura 25 apresenta o resultado dessa projeção para os dois perfis selecionados: *Médico* e *Paciente* para o modelo GTP-4.1 mini, enquanto a Figura 26 apresenta o resultado dessa projeção para os mesmos perfis, mas para o modelo o3 mini. Cada ponto corresponde à resposta de um perfil para uma pergunta específica, identificada por um código (por exemplo, P1, P2, etc.).

Figura 25 – Projeção das respostas com destaque para pares de respostas mais próximas (azul) e mais distantes (vermelho) (modelo GTP-4.1 mini) no domínio médico.



Fonte: elaborado pelo autor.



os pares mais distantes.

### **Análise Qualitativa dos Extremos**

Para a análise qualitativa dos extremos, vamos considerar as respostas geradas com o modelo GPT-4.1 mini.

**Par Médico–Paciente:** a maior distância semântica foi observada na resposta à pergunta “*Qual a quantidade de pacientes que realizaram o exame “Magnésio”?*”:

- **Médico:** *Conforme verificado nos registros disponíveis, 3.612 pacientes foram submetidos ao exame de magnésio. Esse dado é fundamental para análises epidemiológicas e planejamento clínico relacionados ao monitoramento desse eletrólito, considerando sua importância em diversos processos fisiológicos e patologias associadas.*
- **Paciente:** *Oi! Sim, muitas pessoas já fizeram o exame de magnésio aqui. Até o momento, 3.612 pacientes passaram por esse teste. Se você quiser, posso te ajudar a entender melhor para que serve esse exame ou como ele pode ser útil para você.*

No par de perfis “Médico–Paciente”, a maior distância semântica observada está relacionada não ao conteúdo factual, que permanece idêntico em ambos os casos, com o número de pacientes ou exames corretamente reportado, mas à forma de apresentação e ao enquadramento comunicativo. O perfil “Médico” adota um discurso técnico, com foco em interpretação clínica e epidemiológica, destacando a relevância dos dados para análises e planejamento em saúde. O perfil “Paciente”, por outro lado, utiliza um tom conversacional, próximo e acolhedor, incluindo expressões de interação direta e explicações adicionais voltadas ao entendimento leigo. Esse contraste entre um estilo objetivo, centrado na utilidade profissional das informações, e outro mais empático e acessível, orientado ao engajamento, explica a maior distância encontrada entre os textos, ainda que ambos partam da mesma informação numérica.

A menor distância semântica, por sua vez, ocorreu na resposta à pergunta “*Quantos exames foram coletados no local de origem “Recepção do Centro Diagnóstico” no mês de Agosto de 2020?*”:

- **Médico:** *No mês de agosto de 2020, foram coletados exatamente 21.792 exames na origem identificada como “Recepção do Centro Diagnóstico”. Este dado pode ser útil para avaliar a demanda e o fluxo de pacientes nesse ponto específico de atendimento, auxiliando no planejamento operacional e na alocação de recursos para otimizar a qualidade e a eficiência do serviço.*

- **Paciente:** *No mês de agosto de 2020, foram realizados 21.792 exames logo na recepção do Centro Diagnóstico. Essa informação mostra a quantidade significativa de atendimentos que o centro realizou nesse período, o que pode representar um bom ritmo de atividade na unidade. Se precisar de mais detalhes ou informações sobre esses exames, fique à vontade para perguntar!*

No par de perfis “Médico–Paciente”, a menor distância semântica ocorre justamente porque ambas as respostas mantêm praticamente o mesmo núcleo informativo: o número de exames realizados (21.792) no mesmo local e período. As variações ficam restritas ao estilo de apresentação. O perfil “Médico” adota um tom técnico, sugerindo implicações do dado para planejamento e eficiência operacional, enquanto o perfil “Paciente” recorre a uma formulação mais coloquial, destacando o volume de atendimentos de forma acessível e acrescentando uma abertura para diálogo. Essas diferenças de estilo, contudo, não alteram o conteúdo essencial, que permanece idêntico. A proximidade semântica entre os dois textos se explica, portanto, pela coincidência total nos fatos relatados, acompanhada de pequenas diferenças discursivas.

A análise dos *embeddings* mostrou que o par de perfis analisado (“Médico–Paciente”) tende a produzir respostas semanticamente mais próximas entre si. Essa proximidade ocorre porque, mesmo com diferenças de tom e formalidade, os perfis do par compartilham objetivos comunicativos semelhantes. O perfil “Médico” e o perfil “Paciente”, por exemplo, convergem em torno da apresentação de informações clínicas claras e diretamente relacionadas aos exames ou atendimentos, ainda que o primeiro adote uma linguagem mais técnica e o segundo opte por uma formulação mais acessível.

Esses pontos em comum explicam as menores distâncias observadas entre os pares destacados em azul nas figuras 25 e 26. Assim, ainda que os estilos de linguagem variem (do tom técnico-científico ao coloquial e explicativo), a intenção comunicativa do par atua como fator de alinhamento, tornando suas respostas semanticamente próximas no espaço vetorial.

### ***Discussão Sobre os Resultados do Experimento 2***

Os resultados do Experimento 2 demonstram de forma clara que o perfil do usuário influencia significativamente a forma como as respostas são formuladas, tanto em nível textual quanto semântico, para ambos os modelos analisados. A análise de similaridade textual, com base nas métricas ROUGE-L, BLEU e Distância de Edição, indicou que perfis como o de “Técnico do Time” tendem a manter maior proximidade com a resposta padrão, enquanto perfis como

“Torcedor” e “Influenciador Digital” apresentam maior variação linguística, refletindo diferentes estilos narrativos e níveis de informalidade. Já a análise semântica, baseada em *embeddings* e projeção via t-SNE, mostrou que pares de perfis com estilos e propósitos comunicativos semelhantes (como “Técnico” e “Fisioterapeuta”, de um lado, e “Torcedor” e “Influenciador”, de outro) formaram agrupamentos bem definidos no espaço vetorial, indicando alinhamento conceitual entre suas respostas. Essa proximidade sugere que, mesmo com variações na forma textual, os conteúdos transmitidos tendem a ocupar regiões próximas no espaço semântico.

No estudo de caso do domínio médico, entretanto, os agrupamentos se mostraram menos separados, o que pode ser explicado pela natureza altamente factual e objetiva das perguntas, diretamente vinculadas a dados numéricos da base. Nesse cenário, ainda que os perfis tenham adotado tons distintos (como a formalidade técnica do “Médico” e a abordagem mais acessível e coloquial do “Paciente”), o núcleo semântico permaneceu praticamente inalterado, resultando em maior sobreposição entre os vetores. Pares como “Médico–Paciente” tenderam a se aproximar justamente pela convergência de seus objetivos comunicativos, seja na apresentação clara de dados clínicos ou na contextualização prática das informações.

Análises qualitativas reforçam esse padrão ao evidenciar que, mesmo diante de perguntas factuais, as respostas variam em tom, foco e densidade argumentativa. Esses resultados confirmam a Hipótese 2 (H2), validando a premissa de que a adaptação por perfil não apenas altera o estilo da resposta, mas também sua estrutura conceitual, o que destaca a importância de arquiteturas como o PersonaReAct na produção de respostas personalizadas e sensíveis ao perfil do usuário.

## 7 CONCLUSÃO

Esta tese apresentou o PersonaReAct, uma especialização do paradigma ReAct, com o objetivo de incorporar mecanismos explícitos de personalização orientados ao perfil funcional do usuário. A proposta parte do reconhecimento de que modelos baseados em linguagem natural, apesar de sua capacidade de geração coerente e contextual, ainda carecem de mecanismos robustos para adaptar suas respostas a diferentes perfis de usuários com necessidades e expectativas distintas.

A arquitetura proposta preserva o ciclo original de *Reason* e *Act* do ReAct, mas introduz uma etapa final denominada *Tailor*, responsável por transformar as observações obtidas ao longo do processo em uma resposta final personalizada. Essa separação modular permite isolar a lógica de raciocínio e ação da camada de adaptação, viabilizando ajustes personalizados sem comprometer a base de informação construída anteriormente.

Ao longo da tese, conduzimos experimentos em quatro frentes complementares:

1. **Análise qualitativa por perfil:** evidenciou que o PersonaReAct é capaz de adaptar não apenas o tom e vocabulário, mas também o foco informacional das respostas conforme o perfil do usuário, resultando em interações mais alinhadas às expectativas comunicativas de cada grupo funcional.
2. **Análise de similaridade textual:** demonstrou que as respostas geradas para diferentes perfis apresentam alterações substanciais em termos de estrutura, vocabulário e ênfases semânticas, mesmo quando a entrada (pergunta) é idêntica, comprovando a influência efetiva da variável de perfil.
3. **Avaliação quantitativa com abordagens de LLM-as-a-Judge:** comparando o PersonaReAct com o ReAct original, observou-se superioridade do modelo proposto nos critérios de alinhamento, consistência e perceptibilidade. As respostas do PersonaReAct foram avaliadas como mais claras, úteis e adequadas ao perfil do usuário final.
4. **Testes com diferentes modelos de LLMs:** replicamos os experimentos com sete modelos heterogêneos (o3-mini, o4-mini, GPT-5 mini, GPT-5 nano, 4o-mini, GPT-4.1 mini e GPT-4.1 nano), abrangendo variantes com e sem capacidades explícitas de raciocínio e com diferentes portes (mini/nano). Essa frente permitiu avaliar a sensibilidade das abordagens (ReAct, ReAct com *prompt* enriquecido e PersonaReAct) ao tipo de modelo, comparando desempenho médio, consistência (dispersão), e custo computacional, além de discutir *trade-offs* entre fidelidade lexical e divergência estilística sob diferentes

arquiteturas.

Esses resultados permitiram confirmar três das quatro hipóteses formuladas:

1. **Hipótese 1 — confirmada.** A personalização baseada no perfil do usuário aumentou a utilidade, a clareza e a interpretabilidade das respostas. Nas avaliações com *LLM-as-a-Judge*, observamos ganhos consistentes em *Linguagem e Estilo Adequados*, *Interpretação e Explicação dos Dados* e *Coerência do Estilo*, tanto no domínio esportivo quanto no médico.
2. **Hipótese 2 — confirmada.** A análise de similaridade textual (ROUGE-L e BLEU mais baixos, maior distância de edição), somada aos resultados qualitativos por perfil, mostrou que a personalização por perfis gera modificações substanciais no conteúdo, na linguagem e no foco das respostas, mesmo quando a pergunta neutra é idêntica.
3. **Hipótese 3 — confirmada.** A separação explícita do componente de personalização no PersonaReAct (módulo *Tailor*) melhorou a organização e a qualidade das respostas quando comparada ao ReAct com *prompt* enriquecido, especialmente em modelos sem capacidades avançadas de raciocínio, nos quais a personalização se manteve mais estável e consistente ao longo do texto. Em modelos com raciocínio, a diferença foi menor (por o perfil já ser internalizado no ciclo *Reason+Act*), mas o PersonaReAct preservou vantagens de controle e padronização do estilo.
4. **Hipótese 4 — confirmada.** O tipo de modelo de linguagem influenciou o desempenho relativo das abordagens de personalização: modelos com capacidades de raciocínio (p.,ex., o3-mini, o4-mini, GPT-5 mini) tenderam a aproximar os resultados do ReAct com *prompt* enriquecido e do PersonaReAct; já em modelos sem raciocínio e/ou de menor porte (p.,ex., GPT-4.1 mini, GPT-4.1 nano, 4o-mini, GPT-5 nano) o PersonaReAct apresentou vantagens mais claras em alinhamento e consistência, evidenciando a sensibilidade das abordagens ao tipo e ao tamanho do modelo.

## Trabalhos Futuros

A partir dos resultados obtidos e das limitações identificadas, apontam-se os seguintes caminhos para continuidade desta tese:

- **Personalização iterativa:** transformar o módulo *Tailor* em um ciclo de refinamento, permitindo que o agente gere versões sucessivas da resposta final com base em múltiplas etapas de ajuste, possivelmente incorporando feedback do usuário ou heurísticas de

controle por perfil.

- **Construção de um chatbot funcional:** utilizar o PersonaReAct como base para um agente conversacional completo, com capacidade de adaptação ao perfil e interações multi-turno. O domínio esportivo é o primeiro alvo de aplicação, mas a arquitetura pode ser estendida a outros contextos, como educação e finanças.
- **Detecção automática de perfil:** explorar formas de inferência dinâmica do perfil do usuário a partir da linguagem utilizada, histórico de interações ou sinais contextuais, dispensando a necessidade de configuração explícita no início do diálogo.
- **Novas fontes de recuperação de dados:** estender o sistema para integrar outras formas de recuperação de conhecimento, como documentos semi-estruturados, bases de dados temporais ou mecanismos híbridos envolvendo vetores semânticos, por exemplo.
- **Validação com usuários reais:** realizar estudos de usabilidade e testes em campo com participantes reais de diferentes perfis, a fim de avaliar a percepção subjetiva da personalização, medir o impacto na experiência de uso e refinar os critérios de adaptação do sistema.

De forma geral, o PersonaReAct apresentou resultados positivos ao incorporar personalização ao fluxo de raciocínio e ação dos agentes baseados em linguagem. A proposta demonstrou ser eficaz para adaptar as respostas ao perfil do usuário sem comprometer a lógica construída ao longo do processo. Com base nos experimentos realizados, a arquitetura mostrou-se promissora para aplicações práticas que exigem respostas mais adequadas ao contexto do usuário, servindo como base para o desenvolvimento de sistemas mais flexíveis e responsivos em diferentes domínios.

## REFERÊNCIAS

AHER, G. V.; ARRIAGA, R. I.; KALAI, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In: KRAUSE, A.; BRUNSKILL, E.; CHO, K.; ENGELHARDT, B.; SABATO, S.; SCARLETT, J. (Ed.). **Proceedings of the 40th International Conference on Machine Learning**. PMLR, 2023. (Proceedings of Machine Learning Research, v. 202), p. 337–371. Disponível em: <https://proceedings.mlr.press/v202/aher23a.html>.

AI, .; ; YOUNG, A.; CHEN, B.; LI, C.; HUANG, C.; ZHANG, G.; ZHANG, G.; WANG, G.; LI, H.; ZHU, J.; CHEN, J.; CHANG, J.; YU, K.; LIU, P.; LIU, Q.; YUE, S.; YANG, S.; YANG, S.; XIE, W.; HUANG, W.; HU, X.; REN, X.; NIU, X.; NIE, P.; LI, Y.; XU, Y.; LIU, Y.; WANG, Y.; CAI, Y.; GU, Z.; LIU, Z.; DAI, Z. **Yi: Open Foundation Models by 01.AI**. 2025. Disponível em: <https://arxiv.org/abs/2403.04652>.

ARGYLE, L. P.; BUSBY, E. C.; FULDA, N.; GUBLER, J. R.; RYTTING, C.; WINGATE, D. Out of one, many: Using language models to simulate human samples. **Political Analysis**, v. 31, n. 3, p. 337–351, 2023.

BAI, Y.; KADAVATH, S.; KUNDU, S.; ASKELL, A.; KERNION, J.; JONES, A.; CHEN, A.; GOLDIE, A.; MIRHOSEINI, A.; MCKINNON, C.; CHEN, C.; OLSSON, C.; OLAH, C.; HERNANDEZ, D.; DRAIN, D.; GANGULI, D.; LI, D.; TRAN-JOHNSON, E.; PEREZ, E.; KERR, J.; MUELLER, J.; LADISH, J.; LANDAU, J.; NDOUSSE, K.; LUKOSUITE, K.; LOVITT, L.; SELLITTO, M.; ELHAGE, N.; SCHIEFER, N.; MERCADO, N.; DASSARMA, N.; LASENBY, R.; LARSON, R.; RINGER, S.; JOHNSTON, S.; KRAVEC, S.; SHOWK, S. E.; FORT, S.; LANHAM, T.; TELLEEN-LAWTON, T.; CONERLY, T.; HENIGHAN, T.; HUME, T.; BOWMAN, S. R.; HATFIELD-DODDS, Z.; MANN, B.; AMODEI, D.; JOSEPH, N.; MCCANDLISH, S.; BROWN, T.; KAPLAN, J. **Constitutional AI: Harmlessness from AI Feedback**. 2022. Disponível em: <https://arxiv.org/abs/2212.08073>.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language models are few-shot learners. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2020. v. 33, p. 1877–1901. Disponível em: [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf).

CHARPENTIER, L. G. G.; SAMUEL, D. GPT or BERT: why not both? In: HU, M. Y.; MUELLER, A.; ROSS, C.; WILLIAMS, A.; LINZEN, T.; ZHUANG, C.; CHOSHEN, L.; COTTERELL, R.; WARSTADT, A.; WILCOX, E. G. (Ed.). **The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning**. Miami, FL, USA: Association for Computational Linguistics, 2024. p. 262–283. Disponível em: <https://aclanthology.org/2024.conll-babylm.24/>.

CHEN, J.; WANG, X.; XU, R.; YUAN, S.; ZHANG, Y.; SHI, W.; XIE, J.; LI, S.; YANG, R.; ZHU, T.; CHEN, A.; LI, N.; CHEN, L.; HU, C.; WU, S.; REN, S.; FU, Z.; XIAO, Y. **From Persona to Personalization: A Survey on Role-Playing Language Agents**. 2024. Disponível em: <https://arxiv.org/abs/2404.18231>.

CHEN, M.; TWOREK, J.; JUN, H.; YUAN, Q.; PINTO, H. P. de O.; KAPLAN, J.; EDWARDS, H.; BURDA, Y.; JOSEPH, N.; BROCKMAN, G.; RAY, A.; PURI, R.; KRUEGER, G.; PETROV, M.; KHLAAF, H.; SASTRY, G.; MISHKIN, P.; CHAN, B.; GRAY, S.; RYDER, N.; PAVLOV, M.; POWER, A.; KAISER, L.; BAVARIAN, M.; WINTER, C.; TILLET, P.; SUCH, F. P.; CUMMINGS, D.; PLAPPERT, M.; CHANTZIS, F.; BARNES, E.; HERBERT-VOSS, A.; GUSS, W. H.; NICHOL, A.; PAINO, A.; TEZAK, N.; TANG, J.; BABUSCHKIN, I.; BALAJI, S.; JAIN, S.; SAUNDERS, W.; HESSE, C.; CARR, A. N.; LEIKE, J.; ACHIAM, J.; MISRA, V.; MORIKAWA, E.; RADFORD, A.; KNIGHT, M.; BRUNDAGE, M.; MURATI, M.; MAYER, K.; WELINDER, P.; MCGREW, B.; AMODEI, D.; MCCANDLISH, S.; SUTSKEVER, I.; ZAREMBA, W. **Evaluating Large Language Models Trained on Code**. 2021. Disponível em: <https://arxiv.org/abs/2107.03374>.

CHEN, W.; KOENIG, S.; DILKINA, B. **RePrompt: Planning by Automatic Prompt Engineering for Large Language Models Agents**. 2025. Disponível em: <https://arxiv.org/abs/2406.11132>.

CHOWDHERY, A.; NARANG, S.; DEVLIN, J.; BOSMA, M.; MISHRA, G.; ROBERTS, A.; BARHAM, P.; CHUNG, H. W.; SUTTON, C.; GEHRMANN, S.; SCHUH, P.; SHI, K.; TSVYASHCHENKO, S.; MAYNEZ, J.; RAO, A.; BARNES, P.; TAY, Y.; SHAZEER, N.; PRABHAKARAN, V.; REIF, E.; DU, N.; HUTCHINSON, B.; POPE, R.; BRADBURY, J.; AUSTIN, J.; ISARD, M.; GUR-ARI, G.; YIN, P.; DUKE, T.; LEVSKAYA, A.; GHEMAWAT, S.; DEV, S.; MICHALEWSKI, H.; GARCIA, X.; MISRA, V.; ROBINSON, K.; FEDUS, L.; ZHOU, D.; IPPOLITO, D.; LUAN, D.; LIM, H.; ZOPH, B.; SPIRIDONOV, A.; SEPASSI, R.; DOHAN, D.; AGRAWAL, S.; OMERNICK, M.; DAI, A. M.; PILLAI, T. S.; PELLAT, M.; LEWKOWYCZ, A.; MOREIRA, E.; CHILD, R.; POLOZOV, O.; LEE, K.; ZHOU, Z.; WANG, X.; SAETA, B.; DIAZ, M.; FIRAT, O.; CATASTA, M.; WEI, J.; MEIER-HELLSTERN, K.; ECK, D.; DEAN, J.; PETROV, S.; FIEDEL, N. Palm: Scaling language modeling with pathways. **Journal of Machine Learning Research**, v. 24, n. 240, p. 1–113, 2023. Disponível em: <http://jmlr.org/papers/v24/22-1144.html>.

CHRISTAKOPOULOU, K.; LALAMA, A.; ADAMS, C.; QU, I.; AMIR, Y.; CHUCRI, S.; VOLLUCCI, P.; SOLDI, F.; BSEISO, D.; SCODEL, S.; DIXON, L.; CHI, E. H.; CHEN, M. **Large Language Models for User Interest Journeys**. 2023. Disponível em: <https://arxiv.org/abs/2305.15498>.

DESHPANDE, A.; MURAHARI, V.; RAJPUROHIT, T.; KALYAN, A.; NARASIMHAN, K. **Toxicity in ChatGPT: Analyzing Persona-assigned Language Models**. 2023. Disponível em: <https://arxiv.org/abs/2304.05335>.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423/>.

DUMITRU, H.; GIBIEC, M.; HARIRI, N.; CLELAND-HUANG, J.; MOBASHER, B.; CASTRO-HERRERA, C.; MIRAKHORLI, M. On-demand feature recommendations derived from mining public product descriptions. In: **Proceedings of the 33rd International Conference on Software Engineering**. New York, NY, USA: Association for Computing

Machinery, 2011. (ICSE '11), p. 181–190. ISBN 9781450304450. Disponível em: <https://doi.org/10.1145/1985793.1985819>.

DURMUS, E.; NGUYEN, K.; LIAO, T. I.; SCHIEFER, N.; ASKELL, A.; BAKHTIN, A.; CHEN, C.; HATFIELD-DODDS, Z.; HERNANDEZ, D.; JOSEPH, N.; LOVITT, L.; MCCANDLISH, S.; SIKDER, O.; TAMKIN, A.; THAMKUL, J.; KAPLAN, J.; CLARK, J.; GANGULI, D. **Towards Measuring the Representation of Subjective Global Opinions in Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2306.16388>.

FU, T.; GU, J.; LI, Y.; QU, X.; CHENG, Y. **Scaling Reasoning, Losing Control: Evaluating Instruction Following in Large Reasoning Models**. 2025. Disponível em: <https://arxiv.org/abs/2505.14810>.

GU, J.; JIANG, X.; SHI, Z.; TAN, H.; ZHAI, X.; XU, C.; LI, W.; SHEN, Y.; MA, S.; LIU, H.; WANG, S.; ZHANG, K.; WANG, Y.; GAO, W.; NI, L.; GUO, J. **A Survey on LLM-as-a-Judge**. 2025. Disponível em: <https://arxiv.org/abs/2411.15594>.

GUPTA, S.; SHRIVASTAVA, V.; DESHPANDE, A.; KALYAN, A.; CLARK, P.; SABHARWAL, A.; KHOT, T. **Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs**. 2024. Disponível em: <https://arxiv.org/abs/2311.04892>.

HE, Q.; ZENG, J.; HE, Q.; LIANG, J.; XIAO, Y. **From Complex to Simple: Enhancing Multi-Constraint Complex Instruction Following Ability of Large Language Models**. 2024.

IP, J.; VONGTHONGSRI, K. **deepeval**. 2025. Disponível em: <https://github.com/confident-ai/deepeval>.

JIANG, A. Q.; SABLAYROLLES, A.; ROUX, A.; MENSCH, A.; SAVARY, B.; BAMFORD, C.; CHAPLOT, D. S.; CASAS, D. de las; HANNA, E. B.; BRESSAND, F.; LENGYEL, G.; BOUR, G.; LAMPLE, G.; LAVAUD, L. R.; SAULNIER, L.; LACHAUX, M.-A.; STOCK, P.; SUBRAMANIAN, S.; YANG, S.; ANTONIAK, S.; SCAO, T. L.; GERVET, T.; LAVRIL, T.; WANG, T.; LACROIX, T.; SAYED, W. E. **Mixtral of Experts**. 2024. Disponível em: <https://arxiv.org/abs/2401.04088>.

JIANG, G.; XU, M.; ZHU, S.-C.; HAN, W.; ZHANG, C.; ZHU, Y. Evaluating and inducing personality in pre-trained language models. In: OH, A.; NAUMANN, T.; GLOBERSON, A.; SAENKO, K.; HARDT, M.; LEVINE, S. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2023. v. 36, p. 10622–10643. Disponível em: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/21f7b745f73ce0d1f9bcea7f40b1388e-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/21f7b745f73ce0d1f9bcea7f40b1388e-Paper-Conference.pdf).

KANG, W.-C.; NI, J.; MEHTA, N.; SATHIAMOORTHY, M.; HONG, L.; CHI, E.; CHENG, D. Z. **Do LLMs Understand User Preferences? Evaluating LLMs On User Rating Prediction**. 2023. Disponível em: <https://arxiv.org/abs/2305.06474>.

KAPLAN, J.; MCCANDLISH, S.; HENIGHAN, T.; BROWN, T. B.; CHESSE, B.; CHILD, R.; GRAY, S.; RADFORD, A.; WU, J.; AMODEI, D. **Scaling Laws for Neural Language Models**. 2020. Disponível em: <https://arxiv.org/abs/2001.08361>.

KHATTAB, O.; SINGHVI, A.; MAHESHWARI, P.; ZHANG, Z.; SANTHANAM, K.; VARDHAMANAN, S.; HAQ, S.; SHARMA, A.; JOSHI, T. T.; MOAZAM, H.; MILLER, H.; ZAHARIA, M.; POTTS, C. **DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines**. 2023. Disponível em: <https://arxiv.org/abs/2310.03714>.

KIM, J.; YANG, Y. **Few-shot Personalization of LLMs with Mis-aligned Responses**. 2025. Disponível em: <https://arxiv.org/abs/2406.18678>.

LANGCHAIN. **Introducing LangGraph**. 2024. Accessed: 2025-06-27. Disponível em: <https://blog.langchain.com/langgraph/>.

LangChain Team. **LangGraph: A stateful orchestration framework for agentic applications**. 2024. <https://github.com/langchain-ai/langgraph>. Acessado em: 22/06/2025.

LEVENSHTEIN, V. I. *et al.* Binary codes capable of correcting deletions, insertions, and reversals. In: SOVIET UNION. **Soviet physics doklady**. [S. l.], 1966. v. 10, n. 8, p. 707–710.

LI, C.; ZHANG, M.; MEI, Q.; KONG, W.; BENDERSKY, M. Learning to rewrite prompts for personalized text generation. In: **Proceedings of the ACM Web Conference 2024**. New York, NY, USA: Association for Computing Machinery, 2024. (WWW '24), p. 3367–3378. ISBN 9798400701719. Disponível em: <https://doi.org/10.1145/3589334.3645408>.

LI, G.; HAMMOUD, H. A. A. K.; ITANI, H.; KHIZBULLIN, D.; GHANEM, B. Camel: communicative agents for "mind" exploration of large language model society. In: **Proceedings of the 37th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2023. (NIPS '23).

LI, X.; ZHANG, Y.; MALTHOUSE, E. C. **A Preliminary Study of ChatGPT on News Recommendation: Personalization, Provider Fairness, Fake News**. 2023. Disponível em: <https://arxiv.org/abs/2306.10702>.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013/>.

LIU, J.; LIU, C.; ZHOU, P.; LV, R.; ZHOU, K.; ZHANG, Y. **Is ChatGPT a Good Recommender? A Preliminary Study**. 2023. Disponível em: <https://arxiv.org/abs/2304.10149>.

LIU, N. F.; LIN, K.; HEWITT, J.; PARANJAPE, A.; BEVILACQUA, M.; PETRONI, F.; LIANG, P. **Lost in the Middle: How Language Models Use Long Contexts**. 2023. Disponível em: <https://arxiv.org/abs/2307.03172>.

LIU, N. F.; LIN, K.; HEWITT, J.; PARANJAPE, A.; BEVILACQUA, M.; PETRONI, F.; LIANG, P. Lost in the middle: How language models use long contexts. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 12, p. 157–173, 2024. Disponível em: <https://aclanthology.org/2024.tacl-1.9/>.

LIU, Q.; CHEN, N.; SAKAI, T.; WU, X.-M. Once: Boosting content-based recommendation with both open- and closed-source large language models. In: **Proceedings of the 17th ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2024. (WSDM '24), p. 452–461. ISBN 9798400703713. Disponível em: <https://doi.org/10.1145/3616855.3635845>.

LIU, Y.; ITER, D.; XU, Y.; WANG, S.; XU, R.; ZHU, C. **G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment**. 2023. Disponível em: <https://arxiv.org/abs/2303.16634>.

LU, Y.; BARTOLO, M.; MOORE, A.; RIEDEL, S.; STENETORP, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In: MURESAN, S.; NAKOV, P.; VILLAVICENCIO, A. (Ed.). **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022**. Association for Computational Linguistics, 2022. p. 8086–8098. Disponível em: <https://doi.org/10.18653/v1/2022.acl-long.556>.

LYU, H.; JIANG, S.; ZENG, H.; XIA, Y.; WANG, Q.; ZHANG, S.; CHEN, R.; LEUNG, C.; TANG, J.; LUO, J. **LLM-Rec: Personalized Recommendation via Prompting Large Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2307.15780>.

MATARAZZO, A.; TORLONE, R. **A Survey on Large Language Models with some Insights on their Capabilities and Limitations**. 2025. Disponível em: <https://arxiv.org/abs/2501.04040>.

MELLO, L. E.; SUMAN, A.; MEDEIROS, C. B.; PRADO, C. A.; RIZZATTI, E. G.; NUNES, F. L. S.; BARNABÉ, G. F.; FERREIRA, J. E.; Sá, J.; REIS, L. F. L.; RIZZO, L. V.; SARNO, L.; LAMONICA, R. de; MACIEL, R. M. d. B.; CESAR-JR, R. M.; CARVALHO, R. **Opening Brazilian COVID-19 patient data to support world research on pandemics**. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3966427>.

MINAEE, S.; MIKOLOV, T.; NIKZAD, N.; CHENAGHLU, M.; SOCHER, R.; AMATRIAIN, X.; GAO, J. **Large Language Models: A Survey**. 2025. Disponível em: <https://arxiv.org/abs/2402.06196>.

MOK, J.; KIM, I. hwan; PARK, S.; YOON, S. **Exploring the Potential of LLMs as Personalized Assistants: Dataset, Evaluation, and Analysis**. 2025. Disponível em: <https://arxiv.org/abs/2506.01262>.

NARAYANAN, D.; SHOEYBI, M.; CASPER, J.; LEGRESLEY, P.; PATWARY, M.; KORTHIKANTI, V.; VAINBRAND, D.; KASHINKUNTI, P.; BERNAUER, J.; CATANZARO, B.; PHANISHAYEE, A.; ZAHARIA, M. Efficient large-scale language model training on gpu clusters using megatron-lm. In: **Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis**. New York, NY, USA: Association for Computing Machinery, 2021. (SC '21). ISBN 9781450384421. Disponível em: <https://doi.org/10.1145/3458817.3476209>.

OPENAI; ACHIAM, J.; ADLER, S.; AGARWAL, S.; AHMAD, L.; AKKAYA, I.; ALEMAN, F. L.; ALMEIDA, D.; ALTENSCHMIDT, J.; ALTMAN, S.; ANADKAT, S.; AVILA, R.; BABUSCHKIN, I.; BALAJI, S.; BALCOM, V.; BALTESCU, P.; BAO, H.; BAVARIAN, M.; BELGUM, J.; BELLO, I.; BERDINE, J.; BERNADETT-SHAPIRO, G.; BERNER, C.; BOGDONOFF, L.; BOIKO, O.; BOYD, M.; BRAKMAN, A.-L.; BROCKMAN, G.; BROOKS, T.; BRUNDAGE, M.; BUTTON, K.; CAI, T.; CAMPBELL, R.; CANN, A.; CAREY, B.; CARLSON, C.; CARMICHAEL, R.; CHAN, B.; CHANG, C.; CHANTZIS, F.; CHEN, D.; CHEN, S.; CHEN, R.; CHEN, J.; CHEN, M.; CHESS, B.; CHO, C.; CHU, C.; CHUNG, H. W.; CUMMINGS, D.; CURRIER, J.; DAI, Y.; DECAREAUX, C.; DEGRY, T.; DEUTSCH, N.; DEVILLE, D.; DHAR, A.; DOHAN, D.; DOWLING, S.; DUNNING, S.; ECOFFET, A.; ELETI, A.; ELOUNDOU, T.; FARHI, D.; FEDUS, L.; FELIX, N.; FISHMAN, S. P.; FORTE, J.; FULFORD, I.; GAO, L.; GEORGES, E.; GIBSON, C.; GOEL, V.; GOGINENI, T.; GOH, G.; GONTIJO-LOPES, R.; GORDON, J.; GRAFSTEIN, M.; GRAY, S.; GREENE, R.; GROSS, J.; GU, S. S.; GUO, Y.; HALLACY, C.; HAN, J.; HARRIS, J.; HE, Y.; HEATON, M.; HEIDECHE, J.; HESSE, C.; HICKEY, A.; HICKEY, W.; HOESCHELE, P.; HOUGHTON, B.; HSU, K.; HU,

S.; HU, X.; HUIZINGA, J.; JAIN, S.; JAIN, S.; JANG, J.; JIANG, A.; JIANG, R.; JIN, H.; JIN, D.; JOMOTO, S.; JONN, B.; JUN, H.; KAFTAN, T.; KAISER Łukasz; KAMALI, A.; KANITSCHIEDER, I.; KESKAR, N. S.; KHAN, T.; KILPATRICK, L.; KIM, J. W.; KIM, C.; KIM, Y.; KIRCHNER, J. H.; KIROS, J.; KNIGHT, M.; KOKOTAJLO, D.; KONDRACIUK Łukasz; KONDRICH, A.; KONSTANTINIDIS, A.; KOSIC, K.; KRUEGER, G.; KUO, V.; LAMPE, M.; LAN, I.; LEE, T.; LEIKE, J.; LEUNG, J.; LEVY, D.; LI, C. M.; LIM, R.; LIN, M.; LIN, S.; LITWIN, M.; LOPEZ, T.; LOWE, R.; LUE, P.; MAKANJU, A.; MALFACINI, K.; MANNING, S.; MARKOV, T.; MARKOVSKI, Y.; MARTIN, B.; MAYER, K.; MAYNE, A.; MCGREW, B.; MCKINNEY, S. M.; MCLEAVEY, C.; MCMILLAN, P.; MCNEIL, J.; MEDINA, D.; MEHTA, A.; MENICK, J.; METZ, L.; MISHCHENKO, A.; MISHKIN, P.; MONACO, V.; MORIKAWA, E.; MOSSING, D.; MU, T.; MURATI, M.; MURK, O.; MÉLY, D.; NAIR, A.; NAKANO, R.; NAYAK, R.; NEELAKANTAN, A.; NGO, R.; NOH, H.; OUYANG, L.; O'KEEFE, C.; PACHOCKI, J.; PAINO, A.; PALERMO, J.; PANTULIANO, A.; PARASCANDOLO, G.; PARISH, J.; PARPARITA, E.; PASSOS, A.; PAVLOV, M.; PENG, A.; PERELMAN, A.; PERES, F. de A. B.; PETROV, M.; PINTO, H. P. de O.; MICHAEL; POKORNY; POKRASS, M.; PONG, V. H.; POWELL, T.; POWER, A.; POWER, B.; PROEHL, E.; PURI, R.; RADFORD, A.; RAE, J.; RAMESH, A.; RAYMOND, C.; REAL, F.; RIMBACH, K.; ROSS, C.; ROTSTED, B.; ROUSSEZ, H.; RYDER, N.; SALTARELLI, M.; SANDERS, T.; SANTURKAR, S.; SASTRY, G.; SCHMIDT, H.; SCHNURR, D.; SCHULMAN, J.; SELSAM, D.; SHEPPARD, K.; SHERBAKOV, T.; SHIEH, J.; SHOKER, S.; SHYAM, P.; SIDOR, S.; SIGLER, E.; SIMENS, M.; SITKIN, J.; SLAMA, K.; SOHL, I.; SOKOLOWSKY, B.; SONG, Y.; STAUDACHER, N.; SUCH, F. P.; SUMMERS, N.; SUTSKEVER, I.; TANG, J.; TEZAK, N.; THOMPSON, M. B.; TILLET, P.; TOOTOONCHIAN, A.; TSENG, E.; TUGGLE, P.; TURLEY, N.; TWOREK, J.; URIBE, J. F. C.; VALLONE, A.; VIJAYVERGIYA, A.; VOSS, C.; WAINWRIGHT, C.; WANG, J. J.; WANG, A.; WANG, B.; WARD, J.; WEI, J.; WEINMANN, C.; WELIHINDA, A.; WELINDER, P.; WENG, J.; WENG, L.; WIETHOFF, M.; WILLNER, D.; WINTER, C.; WOLRICH, S.; WONG, H.; WORKMAN, L.; WU, S.; WU, J.; WU, M.; XIAO, K.; XU, T.; YOO, S.; YU, K.; YUAN, Q.; ZAREMBA, W.; ZELLERS, R.; ZHANG, C.; ZHANG, M.; ZHAO, S.; ZHENG, T.; ZHUANG, J.; ZHUK, W.; ZOPH, B. **GPT-4 Technical Report**. 2024. Disponível em: <https://arxiv.org/abs/2303.08774>.

OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; WAINWRIGHT, C.; MISHKIN, P.; ZHANG, C.; AGARWAL, S.; SLAMA, K.; RAY, A.; SCHULMAN, J.; HILTON, J.; KELTON, F.; MILLER, L.; SIMENS, M.; ASKELL, A.; WELINDER, P.; CHRISTIANO, P. F.; LEIKE, J.; LOWE, R. Training language models to follow instructions with human feedback. In: KOYEJO, S.; MOHAMED, S.; AGARWAL, A.; BELGRAVE, D.; CHO, K.; OH, A. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2022. v. 35, p. 27730–27744. Disponível em: [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf).

PALMA, D. D. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In: **Proceedings of the 17th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2023. (RecSys '23), p. 1369–1373. ISBN 9798400702419. Disponível em: <https://doi.org/10.1145/3604915.3608889>.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. USA: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. Disponível em: <https://doi.org/10.3115/1073083.1073135>.

PEREZ, L. G.; BARRANCO, M.; MARTINEZ, L. Building user profiles for recommender systems from incomplete preference relations. In: **2007 IEEE International Fuzzy Systems Conference**. [S. l.: s. n.], 2007. p. 1–6.

QIN, Y.; LI, G.; LI, Z.; XU, Z.; SHI, Y.; LIN, Z.; CUI, X.; LI, K.; SUN, X. **Incentivizing Reasoning for Advanced Instruction-Following of Large Language Models**. 2025. Disponível em: <https://arxiv.org/abs/2506.01413>.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. *et al.* Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, v. 21, n. 140, p. 1–67, 2020. Disponível em: <http://jmlr.org/papers/v21/20-074.html>.

RAZA, M.; JAHANGIR, Z.; RIAZ, M. B.; SAEED, M. J.; SATTAR, M. A. Industrial applications of large language models. **Scientific Reports**, Nature Publishing Group UK London, v. 15, n. 1, p. 13755, 2025.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019. Disponível em: <https://arxiv.org/abs/1908.10084>.

SALEMI, A.; MYSORE, S.; BENDERSKY, M.; ZAMANI, H. **LaMP: When Large Language Models Meet Personalization**. 2024. Disponível em: <https://arxiv.org/abs/2304.11406>.

SANNER, S.; BALOG, K.; RADLINSKI, F.; WEDIN, B.; DIXON, L. Large language models are competitive near cold-start recommenders for language- and item-based preferences. In: **Proceedings of the 17th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2023. (RecSys '23), p. 890–896. ISBN 9798400702419. Disponível em: <https://doi.org/10.1145/3604915.3608845>.

SANTURKAR, S.; DURMUS, E.; LADHAK, F.; LEE, C.; LIANG, P.; HASHIMOTO, T. Whose opinions do language models reflect? In: KRAUSE, A.; BRUNSKILL, E.; CHO, K.; ENGELHARDT, B.; SABATO, S.; SCARLETT, J. (Ed.). **Proceedings of the 40th International Conference on Machine Learning**. PMLR, 2023. (Proceedings of Machine Learning Research, v. 202), p. 29971–30004. Disponível em: <https://proceedings.mlr.press/v202/santurkar23a.html>.

SCHICK, T.; DWIVEDI-YU, J.; DESSÍ, R.; RAILEANU, R.; LOMELI, M.; HAMBRO, E.; ZETTLEMOYER, L.; CANCEDDA, N.; SCIALOM, T. Toolformer: language models can teach themselves to use tools. In: **Proceedings of the 37th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2023. (NIPS '23).

SHAO, M.; BASIT, A.; KARRI, R.; SHAFIQUE, M. Survey of different large language model architectures: Trends, benchmarks, and challenges. **IEEE Access**, v. 12, p. 188664–188706, 2024.

SHI, F.; CHEN, X.; MISRA, K.; SCALES, N.; DOHAN, D.; CHI, E. H.; SCHÄRLI, N.; ZHOU, D. Large language models can be easily distracted by irrelevant context. In: KRAUSE,

A.; BRUNSKILL, E.; CHO, K.; ENGELHARDT, B.; SABATO, S.; SCARLETT, J. (Ed.). **Proceedings of the 40th International Conference on Machine Learning**. PMLR, 2023. (Proceedings of Machine Learning Research, v. 202), p. 31210–31227. Disponível em: <https://proceedings.mlr.press/v202/shi23a.html>.

SHI, Y.; ZI, X.; SHI, Z.; ZHANG, H.; WU, Q.; XU, M. Enhancing retrieval and managing retrieval: A four-module synergy for improved quality and efficiency in rag systems. In: \_\_\_\_\_. **ECAI 2024**. IOS Press, 2024. ISBN 9781643685489. Disponível em: <http://dx.doi.org/10.3233/FAIA240748>.

SHIN, T.; RAZEGHI, Y.; IV, R. L. L.; WALLACE, E.; SINGH, S. **AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts**. 2020. Disponível em: <https://arxiv.org/abs/2010.15980>.

SHINN, N.; CASSANO, F.; GOPINATH, A.; NARASIMHAN, K.; YAO, S. Reflexion: language agents with verbal reinforcement learning. In: OH, A.; NAUMANN, T.; GLOBERSON, A.; SAENKO, K.; HARDT, M.; LEVINE, S. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2023. v. 36, p. 8634–8652. Disponível em: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf).

SUN, C.; YANG, K.; REDDY, R. G.; FUNG, Y.; CHAN, H. P.; SMALL, K.; ZHAI, C.; JI, H. Persona-DB: Efficient large language model personalization for response prediction with collaborative data refinement. In: RAMBOW, O.; WANNER, L.; APIDIANAKI, M.; AL-KHALIFA, H.; EUGENIO, B. D.; SCHOCKAERT, S. (Ed.). **Proceedings of the 31st International Conference on Computational Linguistics**. Abu Dhabi, UAE: Association for Computational Linguistics, 2025. p. 281–296. Disponível em: <https://aclanthology.org/2025.coling-main.20/>.

TONG, Y.; LI, D.; WANG, S.; WANG, Y.; TENG, F.; SHANG, J. Can LLMs learn from previous mistakes? investigating LLMs' errors to boost for reasoning. In: KU, L.-W.; MARTINS, A.; SRIKUMAR, V. (Ed.). **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 3065–3080. Disponível em: <https://aclanthology.org/2024.acl-long.169/>.

TOUVRON, H.; LAVRIL, T.; IZACARD, G.; MARTINET, X.; LACHAUX, M.-A.; LACROIX, T.; ROZIÈRE, B.; GOYAL, N.; HAMBRO, E.; AZHAR, F.; RODRIGUEZ, A.; JOULIN, A.; GRAVE, E.; LAMPLE, G. **LLaMA: Open and Efficient Foundation Language Models**. 2023. Disponível em: <https://arxiv.org/abs/2302.13971>.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).

WANG, L.; LIM, E.-P. **Zero-Shot Next-Item Recommendation using Large Pretrained Language Models**. 2023. Disponível em: <https://arxiv.org/abs/2304.03153>.

WANG, X.; WEI, J.; SCHUURMANS, D.; LE, Q. V.; CHI, E. H.; NARANG, S.; CHOWDHURY, A.; ZHOU, D. Self-consistency improves chain of thought reasoning in language models. In: **The Eleventh International Conference on Learning Representations**. [S. n.], 2023. Disponível em: <https://openreview.net/forum?id=1PL1NIMMrw>.

WANG, X.; XIAO, Y.; HUANG, J. tse; YUAN, S.; XU, R.; GUO, H.; TU, Q.; FEI, Y.; LENG, Z.; WANG, W.; CHEN, J.; LI, C.; XIAO, Y. **InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews**. 2024. Disponível em: <https://arxiv.org/abs/2310.17976>.

WANG, Z. M.; PENG, Z.; QUE, H.; LIU, J.; ZHOU, W.; WU, Y.; GUO, H.; GAN, R.; NI, Z.; YANG, J.; ZHANG, M.; ZHANG, Z.; OUYANG, W.; XU, K.; HUANG, S. W.; FU, J.; PENG, J. **RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2310.00746>.

WEI, J.; WANG, X.; SCHUURMANS, D.; BOSMA, M.; ICHTER, b.; XIA, F.; CHI, E.; LE, Q. V.; ZHOU, D. Chain-of-thought prompting elicits reasoning in large language models. In: KOYEJO, S.; MOHAMED, S.; AGARWAL, A.; BELGRAVE, D.; CHO, K.; OH, A. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2022. v. 35, p. 24824–24837. Disponível em: [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf).

WOZNIAK, S.; KOPTYRA, B.; JANZ, A.; KAZIENKO, P.; KOCON, J. **Personalized Large Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2402.09269>.

WU, Q.; BANSAL, G.; ZHANG, J.; WU, Y.; LI, B.; ZHU, E.; JIANG, L.; ZHANG, X.; ZHANG, S.; LIU, J.; AWADALLAH, A. H.; WHITE, R. W.; BURGER, D.; WANG, C. **AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation**. 2023. Disponível em: <https://arxiv.org/abs/2308.08155>.

XI, Y.; LIU, W.; LIN, J.; CAI, X.; ZHU, H.; ZHU, J.; CHEN, B.; TANG, R.; ZHANG, W.; YU, Y. Towards open-world recommendation with knowledge augmentation from large language models. In: **Proceedings of the 18th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2024. (RecSys '24), p. 12–22. ISBN 9798400705052. Disponível em: <https://doi.org/10.1145/3640457.3688104>.

XU, X.; LIU, C.; SONG, D. **SQLNet: Generating Structured Queries From Natural Language Without Reinforcement Learning**. 2017. Disponível em: <https://arxiv.org/abs/1711.04436>.

XU, Z.; LIU, Y.; DENG, G.; LI, Y.; PICEK, S. **A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2402.13457>.

YAO, S.; ZHAO, J.; YU, D.; DU, N.; SHAFRAN, I.; NARASIMHAN, K. R.; CAO, Y. React: Synergizing reasoning and acting in language models. In: **The Eleventh International Conference on Learning Representations**. [S. n.], 2023. Disponível em: [https://openreview.net/forum?id=WE\\_vluYUL-X](https://openreview.net/forum?id=WE_vluYUL-X).

ZERHOUDI, S.; GRANITZER, M. Personarag: Enhancing retrieval-augmented generation systems with user-centric agents. **arXiv preprint arXiv:2407.09394**, 2024.

ZHANG, Z.; ROSSI, R. A.; KVETON, B.; SHAO, Y.; YANG, D.; ZAMANI, H.; DERNONCOURT, F.; BARROW, J.; YU, T.; KIM, S.; ZHANG, R.; GU, J.; DERR, T.; CHEN, H.; WU, J.; CHEN, X.; WANG, Z.; MITRA, S.; LIPKA, N.; AHMED, N.; WANG, Y. **Personalization of Large Language Models: A Survey**. 2025. Disponível em: <https://arxiv.org/abs/2411.00027>.

ZHAO, W. X.; ZHOU, K.; LI, J.; TANG, T.; WANG, X.; HOU, Y.; MIN, Y.; ZHANG, B.; ZHANG, J.; DONG, Z.; DU, Y.; YANG, C.; CHEN, Y.; CHEN, Z.; JIANG, J.; REN, R.; LI, Y.; TANG, X.; LIU, Z.; LIU, P.; NIE, J.-Y.; WEN, J.-R. **A Survey of Large Language Models**. 2025. Disponível em: <https://arxiv.org/abs/2303.18223>.

ZHONG, V.; XIONG, C.; SOCHER, R. **Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning**. 2017. Disponível em: <https://arxiv.org/abs/1709.00103>.

ZHOU, D.; SCHÄRLI, N.; HOU, L.; WEI, J.; SCALES, N.; WANG, X.; SCHUURMANS, D.; CUI, C.; BOUSQUET, O.; LE, Q.; CHI, E. **Least-to-Most Prompting Enables Complex Reasoning in Large Language Models**. 2023. Disponível em: <https://arxiv.org/abs/2205.10625>.

ZIEMS, C.; HELD, W.; SHAIKH, O.; CHEN, J.; ZHANG, Z.; YANG, D. Can large language models transform computational social science? **Computational Linguistics**, MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA . . . , v. 50, n. 1, p. 237–291, 2024.

## APÊNDICES

## APÊNDICE A – CONJUNTO DE DADOS INICIAL

Os Quadros 14, 15, 16, 17, 18, 19 e 20 apresentam os atributos que compõem o conjunto de dados biométricos e físicos utilizados no Estudo de Caso 1 desta tese. Os dados foram coletados por meio de dispositivos vestíveis durante sessões de treino e jogos realizados por atletas profissionais, e estão organizados em sete categorias principais: (1) identificação do atleta e da sessão, (2) dados temporais da atividade, (3) informações sobre frequência cardíaca (HR), (4) métricas de distância e velocidade, (5) eventos de aceleração e desaceleração, (6) indicadores de carga fisiológica e esforço, e (7) variabilidade da frequência cardíaca (HRV) e intervalos RR.

Quadro 14 – Identificação do atleta e sessão.

Atributo	Unidade de Medida
Número do jogador	–
Nome do jogador	–
Nome da sessão	–
Tipo de sessão	–
Fase da sessão	–

Fonte: elaborado pelo autor.

Quadro 15 – Dados temporais da sessão.

Atributo	Unidade de Medida
Duração da sessão	minutos
Horário de início	timestamp
Horário de término	timestamp

Fonte: elaborado pelo autor.

Quadro 16 – Frequência cardíaca (HR).

Atributo	Unidade de Medida
Frequência cardíaca mínima	bpm
Frequência cardíaca média	bpm
Frequência cardíaca máxima	bpm
Frequência cardíaca mínima	%
Frequência cardíaca média	%
Frequência cardíaca máxima	%
Tempo na zona 1 (50–59%)	minutos
Tempo na zona 2 (60–69%)	minutos
Tempo na zona 3 (70–79%)	minutos
Tempo na zona 4 (80–89%)	minutos
Tempo na zona 5 (90–100%)	minutos

Fonte: elaborado pelo autor.

Os Quadros 21, 22 e 23 apresentam as informações do conjunto de dados hospitalar utilizado no Estudo de Caso 2. Os dados são públicos e divulgados pelo Hospital Sírio-Libanês.

Quadro 17 – Distância e velocidade.

Atributo	Unidade de Medida
Distância total	metros
Distância por minuto	m/min
Velocidade máxima	km/h
Velocidade média	km/h
Número de sprints	–
Distância na zona de velocidade 1 (7.20–10.19)	metros
Distância na zona de velocidade 2 (10.20–14.39)	metros
Distância na zona de velocidade 3 (14.40–19.79)	metros
Distância na zona de velocidade 4 (19.80–25.19)	metros
Distância na zona de velocidade 5 (25.20+)	metros

Fonte: elaborado pelo autor.

Quadro 18 – Acelerações e desacelerações.

Atributo	Unidade de Medida
Acelerações (-50.00 a -3.00)	contagem
Acelerações (-2.99 a -2.00)	contagem
Acelerações (-1.99 a -1.00)	contagem
Acelerações (-0.99 a -0.50)	contagem
Acelerações (0.50 a 0.99)	contagem
Acelerações (1.00 a 1.99)	contagem
Acelerações (2.00 a 2.99)	contagem
Acelerações (3.00 a 50.00)	contagem

Fonte: elaborado pelo autor.

Quadro 19 – Carga fisiológica e esforço.

Atributo	Unidade de Medida
Calorias	kcal
Pontuação de carga de treino	–
Carga cardiovascular	–
Tempo de recuperação	horas

Fonte: elaborado pelo autor.

Quadro 20 – Variabilidade da frequência cardíaca (HRV) e intervalos RR.

Atributo	Unidade de Medida
Intervalo RR mínimo	milissegundos
Intervalo RR máximo	milissegundos
Intervalo RR médio	milissegundos
HRV (RMSSD)	milissegundos

Fonte: elaborado pelo autor.

Quadro 21 – Informações sobre os pacientes.

Atributo	Descrição	Conteúdo
ID_PACIENTE	Identificação única do paciente (correlaciona todos os dados do paciente)	String, anonimizado
IC_SEXO	Sexo do Paciente	F - Feminino / M - Masculino
AA_NASCIMENTO	Ano de nascimento do Paciente	Os 4 dígitos do ano do nascimento ou AAAA - desconhecido
CD_PAIS	País de residência do Paciente	BR ou XX (país estrangeiro)
SG_UF	Estado de residência do Paciente	Sigla da UF ou XX (estado estrangeiro)
CD_MUNICIPIO	Município de residência do Paciente	Código do IBGE do município ou 999999

Fonte: elaborado pelo autor.

Quadro 22 – Dados dos exames coletados.

<b>Atributo</b>	<b>Descrição</b>	<b>Conteúdo</b>
ID_PACIENTE	Identificação única do paciente (correlaciona todos os dados do paciente)	String, anonimizado
ID_ATENDIMENTO	Identificação única do atendimento. Correlação entre atendimento e exame	String, anonimizado
DT_COLETA	Data em que o material foi coletado do paciente	Exemplo: 24/08/2020
DE_ORIGEM	Local de Coleta daquele exame	Exemplo: LAB, UPA, etc.
CD_EXAME	Código do exame	Exemplo: 406
DE_EXAME	Descrição do exame	Texto livre

Fonte: elaborado pelo autor.

Quadro 23 – Informações gerais sobre os desfechos dos pacientes.

<b>Atributo</b>	<b>Descrição</b>	<b>Conteúdo</b>
ID_PACIENTE	Identificação única do paciente (correlaciona todos os dados do paciente)	String, anonimizado
ID_ATENDIMENTO	Identificação única do atendimento. Cada atendimento possui um código único.	String, anonimizado
DT_ATENDIMENTO	Data de realização do atendimento	DD = Dia / MM = Mês / AAAA = Ano
DE_TIPO_ATENDIMENTO	Descrição do tipo de atendimento realizado	Exemplo: Pronto atendimento
ID_CLINICA	Identificação da clínica onde o evento aconteceu	Exemplo: 1013
CD_EVENTO	Código do evento realizado	Código padronizado
DE_EVENTO	Descrição do evento realizado	Exemplo: Administração de medicamento

Fonte: elaborado pelo autor.

## APÊNDICE B – CRITÉRIOS TEXTUAIS UTILIZADOS NA AVALIAÇÃO POR MODELOS DE LINGUAGEM

Este apêndice reúne os critérios textuais utilizados para orientar a avaliação automática das respostas nos experimentos conduzidos nesta pesquisa. Tais critérios foram aplicados no contexto do *framework G-Eval*, que implementa a abordagem conhecida como *LLM-as-a-Judge*, na qual uma LLM é utilizada como avaliadora com base em descrições linguísticas dos aspectos a serem avaliados. Cada critério é acompanhado de uma rúbrica textual que detalha os parâmetros esperados em diferentes níveis de qualidade, servindo como referência explícita para orientar os julgamentos realizados pela LLM.

Diferentemente de métricas automatizadas convencionais (como BLEU, ROUGE ou Distância de Edição), os critérios aqui apresentados foram formulados em linguagem natural, com instruções explícitas sobre como a LLM avaliadora deveria interpretar cada dimensão qualitativa da resposta, como personalização e alinhamento com o perfil. Esses critérios foram fundamentais para garantir uma avaliação contextualizada e alinhada aos objetivos de cada experimento.

Os Códigos-Fonte 1, 2 e 3 apresentam as implementações dos critérios utilizados nas avaliações dos experimentos.

## Código-fonte 1 – Critério “Linguagem e Estilo Apropriados”.

```
1 GEval(  
2     name="Linguagem e Estilo Apropriados",  
3     criteria="Avalie se o estilo da resposta soa natural para o  
4     perfil informado, com palavras, tom e jeito de falar que combinem  
5     com esse tipo de perfil.",  
6     evaluation_params=[  
7         LLMTestCaseParams.INPUT,  
8         LLMTestCaseParams.ACTUAL_OUTPUT,  
9         LLMTestCaseParams.CONTEXT,  
10    ],  
11    verbose_mode=True,  
12    rubric=[  
13        Rubric(score_range=(0, 2), expected_outcome="A linguagem esta  
14        fora do tom esperado, usando palavras difíceis, jargões ou um  
15        estilo que não combina com o perfil."),  
16        Rubric(score_range=(3, 5), expected_outcome="Ha tentativa de  
17        adaptar o estilo, mas ainda soa generico ou forçado em varias  
18        partes."),  
19        Rubric(score_range=(6, 8), expected_outcome="O estilo e  
20        geralmente bom e condiz com o perfil, mesmo que tenha pequenos  
21        momentos de desalinhamento."),  
22        Rubric(score_range=(9, 10), expected_outcome="O estilo e  
23        totalmente natural e fluido, combinando perfeitamente com o perfil  
24        do inicio ao fim."),  
25    ],  
26 )
```

Fonte: elaborado pelo autor.

## Código-fonte 2 – Critério “Interpretação e Explicação dos Dados”.

```
1 GEval(  
2     name="Interpretacao e Explicacao dos Dados",  
3     criteria="Avalie se a resposta interpreta corretamente os dados  
4     apresentados e fornece explicacoes uteis e contextualizadas para o  
5     perfil informado.",  
6     evaluation_params=[  
7         LLMTestCaseParams.INPUT,  
8         LLMTestCaseParams.ACTUAL_OUTPUT,  
9         LLMTestCaseParams.CONTEXT,  
10    ],  
11    verbose_mode=True,  
12    rubric=[  
13        Rubric(score_range=(0, 2), expected_outcome="A resposta  
14        apenas menciona dados brutos ou incorretos, sem interpretar ou  
15        contextualizar."),  
16        Rubric(score_range=(3, 5), expected_outcome="Ha tentativa de  
17        interpretacao, mas ela e vaga, superficial ou generica, sem  
18        agregar valor real."),  
19        Rubric(score_range=(6, 8), expected_outcome="Os dados sao bem  
20        interpretados, com explicacoes que ajudam o perfil a compreender  
21        sua relevancia."),  
22        Rubric(score_range=(9, 10), expected_outcome="A resposta  
23        combina dados especificos, interpretacao correta e sugestoes  
24        relevantes, oferecendo uma explicacao clara e estrategica adaptada  
25        ao perfil."),  
26    ],  
27 )
```

Fonte: elaborado pelo autor.

## Código-fonte 3 – Critério “Coerência do Estilo ao Longo do Texto”.

```
1 GEval(  
2     name="Coerencia do Estilo ao Longo do Texto",  
3     criteria="Avalie se o estilo adotado e mantido de forma  
4     consistente ao longo da resposta, sem mudancas abruptas ou  
5     incoerentes no tom, vocabulario ou estrutura.",  
6     evaluation_params=[  
7         LLMTestCaseParams.INPUT,  
8         LLMTestCaseParams.ACTUAL_OUTPUT,  
9         LLMTestCaseParams.CONTEXT,  
10    ],  
11    verbose_mode=True,  
12    rubric=[  
13        Rubric(score_range=(0, 2), expected_outcome="A resposta  
14        apresenta variacoes bruscas no estilo, alternando entre registros  
15        diferentes ou incoerentes."),  
16        Rubric(score_range=(3, 5), expected_outcome="O estilo tenta  
17        ser consistente, mas ha mudancas perceptíveis de tom ou  
18        vocabulario em partes relevantes."),  
19        Rubric(score_range=(6, 8), expected_outcome="O estilo e  
20        predominantemente coerente, com apenas uma ou outra leve oscilacao  
21        que nao compromete o todo."),  
22        Rubric(score_range=(9, 10), expected_outcome="A resposta  
23        mantem o mesmo padrao de estilo do inicio ao fim, com tom e  
24        vocabulario uniformes e adequados."),  
25    ],  
26 )
```

Fonte: elaborado pelo autor.

## APÊNDICE C – PROMPTS DOS AGENTES

Este apêndice apresenta o prompt completo utilizado pelo Agente de Personalização, responsável por adaptar as respostas com base no perfil do usuário. O prompt explicita o raciocínio esperado em cada etapa da personalização, desde a análise do perfil até a escolha do tom e apresentação dos dados.

```
1 personalization_prompt = PromptTemplate.from_template("""
2 Sua tarefa e gerar uma resposta final personalizada a partir de uma
   pergunta feita por um usuario com um perfil especifico. Voce deve
   considerar:
3
4 - O perfil do usuario ({profile}) para adaptar a linguagem, o foco, o
   nivel de formalidade e o grau de detalhamento tecnico da resposta
   .
5 - O conteudo da pergunta original ({question}) para compreender o
   contexto e estilo.
6 - O historico de dados disponiveis ({history}), que traz informacoes
   factuais relevantes.
7
8 ---
9
10 ### Etapa 1 - Reflexao sobre o Perfil do Usuario
11
12 Antes de comecar a resposta, reflita cuidadosamente sobre o perfil
   informado em {profile}. Em particular, avalie:
13
14 - Que tipo de linguagem essa pessoa espera? (informal, neutra,
   tecnica, narrativa?)
15 - Qual o nivel de formalidade adequado? (gurias ou termos tecnicos?
   frases diretas ou envolventes?)
16 - O usuario valoriza interpretacoes subjetivas ou dados tecnicos
   precisos?
17 - Que tipo de conteudo e mais util para ela?
18
19 ---
20
21 ### Etapa 2 - Analise da Pergunta
22
```

```
23 Analise o estilo e o tom da pergunta ({question}). Se ela for
    informal, mantenha esse tom. Se for tecnica, preserve o foco e a
    precisao. Use esse estilo como referencia na resposta.
24
25 ---
26
27 ### Etapa 3 - Selecao e Apresentacao dos Dados
28
29 Voce tem acesso ao historico de dados e analises ({history}). Use-o
    da seguinte forma:
30
31 - Selecione apenas os dados mais relevantes para o perfil.
32 - Se o perfil exigir interpretacao subjetiva e envolvente (ex:
    torcedor do time ou narrador de jogo de futebol), evite numeros
    crus e foque em transformar os dados em descricoes
    significativas.
33 - Se o perfil exigir precisao tecnica e formalidade (ex: tecnico
    ou fisiologista), apresente os valores exatos com clareza (ex:
    "velocidade media de 7.23 km/h", "tempo total em alta intensidade
    : 436s", etc).
34
35 ---
36
37 ### Etapa 4 - Escrita e Revisao da Resposta
38
39 Redija a resposta final com base nas decisoes anteriores. Antes de
    finalizar:
40
41 - Verifique se a linguagem esta adequada ao perfil (informal vs.
    tecnica);
42 - Confirme se os dados estao apresentados no formato certo (
    interpretacao subjetiva ou metrica bruta);
43 - Certifique-se de que a personalizacao ocorre ao longo de toda a
    resposta e nao apenas na conclusao.
44
45 Simule mentalmente que voce e o usuario descrito. Pergunte-se:
46
47 > Essa resposta soa natural, clara e util para alguem com esse perfil
    ?
```

```
48  
49 Retorne apenas a resposta final personalizada.  
50 """ )
```

Código-fonte 4 – Prompt do Agente de Personalização.

## **APÊNDICE D – EXEMPLOS DE PERGUNTAS E RESPOSTAS**

### **Variações de Perguntas por Perfil de Usuário**

O Quadro 24 apresenta o um subconjunto de perguntas utilizadas nos experimentos em linguagem neutra, juntamente com suas respectivas variações adaptadas a diferentes perfis funcionais de usuário. O objetivo é evidenciar como o mesmo conteúdo informacional pode ser solicitado com estilos comunicacionais distintos, dependendo do papel desempenhado pelo interlocutor.

### **Varição de Respostas para uma mesma Pergunta**

Os Quadros 25, 26, 27, 28 apresentam exemplos exemplos de perguntas e suas respectivas respostas para os diferentes perfis utilizados.

Quadro 24 – Variações das perguntas por perfil de usuário.

<b>Perfil</b>	<b>Variações das Perguntas</b>
<b>Linguagem Neutra</b>	<ol style="list-style-type: none"> <li>1. Qual foi a velocidade média de Bruno Pacheco no jogo contra o Ceará no dia 1º de abril de 2023?</li> <li>2. Qual jogador apresentou a maior quantidade de sprints durante o jogo contra o Ceará no dia 1º de abril de 2023?</li> <li>3. Qual foi o tempo total que Bruno Pacheco passou sprintando durante a partida contra o Ceará em 1º de abril de 2023?</li> </ol>
<b>Torcedor do Time</b>	<ol style="list-style-type: none"> <li>1. O Bruno Pacheco correu bem naquele jogo contra o Ceará do dia 1 de abril ou ficou devendo na velocidade?</li> <li>2. Quem foi o cara que mais deu arrancadas durante o jogo contra o Ceará no começo de abril?</li> <li>3. Quanto tempo o Bruno Pacheco passou correndo no gás total naquele jogo contra o Ceará no começo de abril?</li> </ol>
<b>Técnico do Time</b>	<ol style="list-style-type: none"> <li>1. Qual foi a velocidade média registrada por Bruno Pacheco no jogo contra o Ceará em 1º de abril de 2023?</li> <li>2. Qual jogador apresentou a maior frequência de sprints durante o jogo contra o Ceará no dia 1º de abril de 2023?</li> <li>3. Quantos segundos Bruno Pacheco permaneceu em atividade de sprint na partida contra o Ceará no dia 1º de abril de 2023?</li> </ol>
<b>Narrador de Futebol</b>	<ol style="list-style-type: none"> <li>1. Torcedor ligado na nossa transmissão quer saber: o quão rápido o Bruno Pacheco correu contra o Ceará no jogo do dia 1 de abril?</li> <li>2. Quem foi o verdadeiro raio em campo, disparando sprints sem parar naquele clássico contra o Ceará no dia 1º de abril de 2023?</li> <li>3. Quantos minutos o incansável Bruno Pacheco passou em ritmo acelerado, sprintando com tudo naquela partida contra o Ceará no começo de abril?</li> </ol>
<b>Influenciador Digital</b>	<ol style="list-style-type: none"> <li>1. O Bruno Pacheco tava rápido naquele jogo contra o Ceará no começo de abril? Tô pensando em postar algo legal sobre isso!</li> <li>2. Quem foi que mais disparou em campo naquele jogo contra o Ceará no começo de abril? Quero destacar quem mais botou fogo na partida!</li> <li>3. Sabe me dizer quanto tempo o Bruno Pacheco ficou no modo turbo naquele jogo contra o Ceará no dia 1º de abril de 2023? Pode virar um bom reels!</li> </ol>
<b>Fisioterapeuta do Time</b>	<ol style="list-style-type: none"> <li>1. Qual foi a velocidade média do Bruno Pacheco no jogo contra o Ceará no dia 1º de abril de 2023? Quero avaliar a demanda mecânica sobre ele.</li> <li>2. Qual jogador apresentou maior número de sprints no jogo contra o Ceará no dia 1º de abril de 2023? Essa informação ajuda no controle de carga excêntrica.</li> <li>3. Por quanto tempo o Pacheco ficou sprintando durante a partida contra o Ceará no dia 1º de abril de 2023? Isso impacta diretamente no planejamento de recuperação.</li> </ol>
<b>Jornalista Esportivo</b>	<ol style="list-style-type: none"> <li>1. Qual foi a velocidade média do Bruno Pacheco no jogo contra o Ceará no dia 1º de abril de 2023? Quero incluir esse dado na matéria.</li> <li>2. Quem liderou em número de sprints no clássico contra o Ceará do dia 1º de abril? Pode ser um bom destaque de performance.</li> <li>3. Quanto tempo Bruno Pacheco passou sprintando nessa partida contra o Ceará do dia 1º de abril de 2023? Pode revelar o nível de intensidade dele.</li> </ol>

Fonte: elaborado pelo autor.

Quadro 25 – Respostas geradas pelo PersonaReAct para diferentes perfis a partir da mesma pergunta sobre sprints no jogo contra o Ceará.

<b>Pergunta:</b> <i>Qual jogador apresentou a maior quantidade de sprints durante o jogo contra o Ceará no dia 1º de abril de 2023?</i>	
<b>Perfil do usuário</b>	<b>Resposta</b>
<b>Resposta sem Personalização</b>	<i>O jogador que apresentou a maior quantidade de sprints durante o jogo contra o Ceará no dia 1º de abril de 2023 foi Tomas Pochettino, com uma média de 1.0176470588235293 sprints.</i>
<b>Torcedor do Time</b>	<i>No jogo contra o Ceará no começo de abril, quem se destacou com suas arrancadas foi o nosso guerreiro, Tomas Pochettino! O cara fez bonito em campo, mostrando todo seu ímpeto e garra, com uma média de mais de uma arrancada por minuto. Realmente, ele se entregou demais nesse jogo, não teve pra ninguém! É isso que a gente gosta de ver: jogador dando o sangue pela vitória! Vamos em frente, com ele no time, a esperança é grande!</i>
<b>Técnico do Time</b>	<i>O jogador que apresentou a maior frequência de sprints durante o jogo contra o Ceará no dia 1º de abril de 2023 foi Tomas Pochettino, com uma média de 1.02 sprints por minuto. Esses dados indicam que ele teve um desempenho físico destacado ao longo da partida, sendo fundamental na dinâmica de jogo e na pressão em campo. É uma estatística interessante para avaliarmos o impacto desse jogador na partida, especialmente em funções que demandam alta intensidade e movimentação constante.</i>
<b>Narrador de Jogo de Futebol</b>	<i>No emocionante clássico contra o Ceará no dia 1º de abril de 2023, o verdadeiro raio que iluminou o gramado com sua velocidade e determinação foi Tomas Pochettino. Com uma performance de dar inveja, ele disparou sprints sem parar, desfilando sua habilidade e energia, e garantindo que sua presença fosse notada em cada canto do campo. Com uma média impressionante de sprints que deixaram os adversários atordoados, Pochettino foi, sem dúvida, o destaque do jogo, mostrando a todos quem realmente faz a diferença em campo. O torcedor vibrou com cada arrancada e cada jogada, e não é à toa que ele foi considerado o grande ícone daquele dia.</i>
<b>Influenciador Digital</b>	<i>'Galera, olha só quem incendiou o jogo contra o Ceará no início de abril: foi Tomas Pochettino! Ele mandou ver e foi o cara que mais disparou no campo, com uma média de 1.02 sprints por partida! O hõmi parecia um foguete, tudo pra ajudar a equipe! Sem dúvidas, a performance dele foi crucial para o desempenho do time. O que acham? Vamos dar os créditos a esse cara que realmente botou fogo na partida?'</i>
<b>Fisioterapeuta do Time</b>	<i>No jogo contra o Ceará em 1º de abril de 2023, o jogador que se destacou com o maior número de sprints foi Tomas Pochettino, alcançando uma média de 1,02 sprints. Essa informação é bastante valiosa para o controle de carga excêntrica, pois sprints são indicativos de intensidade e requerem uma análise cuidadosa para garantir a recuperação e performance adequada dos atletas. Monitorar esses dados ajuda a ajustar o treinamento e a prevenir lesões, garantindo a integridade física do jogador ao longo da temporada.</i>
<b>Jornalista Esportivo</b>	<i>No clássico contra o Ceará, realizado no dia 1º de abril, Tomas Pochettino foi o destaque em campo, liderando em número de sprints com uma média impressionante de 1,02 sprints por minuto. Esse desempenho não apenas evidencia sua intensidade e empenho durante a partida, mas também o posiciona como um jogador crucial nas dinâmicas de movimentação e pressão no campo. Certamente, essa é uma informação que pode trazer luz à sua performance e importância no jogo.</i>

Fonte: elaborado pelo autor.

Quadro 26 – Respostas geradas pelo PersonaReAct para diferentes perfis sobre o tempo de sprint de Bruno Pacheco no jogo contra o Ceará.

<b>Pergunta:</b> <i>Qual foi o tempo total que Bruno Pacheco passou sprintando durante a partida contra o Ceará em 1º de abril de 2023?</i>	
<b>Resposta sem Personalização</b>	<i>Bruno Pacheco passou 115 segundos sprintando durante a partida contra o Ceará em 1º de abril de 2023.</i>
<b>Torcedor do Time</b>	<i>Olha, meu amigo, o Bruno Pacheco realmente deu o sangue naquela partida contra o Ceará no começo de abril! Foram impressionantes 424 minutos em que ele manteve o gás total, correndo como um verdadeiro leão em campo. O cara não parou nem um segundo, mostrando todo o seu empenho e garra. Essa dedicação é o que a gente espera de um jogador que veste a nossa camisa! Que atuação!</i>
<b>Técnico do Time</b>	<i>Bruno Pacheco permaneceu em atividade de sprint por 115 segundos na partida contra o Ceará, realizada no dia 1º de abril de 2023. Esse dado é relevante para entender seu nível de intensidade e participação em momentos-chave do jogo, evidenciando sua contribuição na dinâmica da equipe durante as transições ofensivas.</i>
<b>Narrador de Jogo de Futebol</b>	<i>Naqueles momentos intensos, nosso destemido Bruno Pacheco se destacou de forma impressionante na partida contra o Ceará no começo de abril. Ele passou cerca de 21,76 segundos do jogo em pura velocidade, sprintando com tudo, como um verdadeiro leão em campo. Essa intensidade mostra o quanto ele é incansável e sempre disposto a dar o máximo, seja atacando, seja se defendendo. Um jogador desse calibre, que entrega tanto desgaste físico em tantos momentos, realmente faz a diferença no time, certo? Espectacular!</i>
<b>Influenciador Digital</b>	<i>E aí, pessoal? Olha só essa info que talvez vire um reels matador! No jogo contra o Ceará, no dia 1º de abril de 2023, o Bruno Pacheco acionou o 'modo turbo' e ficou nessa pegada intensa por cerca de 21 minutos! Ele realmente deu tudo em campo, né? Corre lá e já pensa em como usar isso no seu conteúdo! #Futebol #BrunoPacheco #ModoTurbo</i>
<b>Fisioterapeuta do Time</b>	<i>Durante a partida contra o Ceará no dia 1º de abril de 2023, o Pacheco ficou sprintando por aproximadamente 21,8 minutos. Essa informação é fundamental para o nosso planejamento de recuperação, visto que o tempo de esforço intenso pode impactar diretamente nos protocolos de recuperação a serem utilizados, como a intensidade e o tipo de fisioterapia que podemos implementar. Recomendo que consideremos essa carga em nossa avaliação de desgaste e nas estratégias de recuperação, para otimizar o retorno do atleta às atividades.</i>
<b>Jornalista Esportivo</b>	<i>Na partida contra o Ceará, no dia 1º de abril de 2023, Bruno Pacheco teve um desempenho interessante, com cerca de 0,36 minutos dedicados a sprints. Isso demonstra que ele não ficou apenas na defesa, mas se aventurou à frente quando necessário, apresentando um nível de intensidade moderado em seu jogo. Essa quantidade de tempo sprintando mostra que, embora não tenha se destacado como um atleta que correu sem parar, ele efetuou algumas investidas rápidas, contribuindo para o movimento do time em campo. Vale ressaltar que, em momentos críticos, a presença de um jogador como Pacheco em situações de velocidade pode ser crucial para a dinâmica de jogo.</i>

Fonte: elaborado pelo autor.

Quadro 27 – Respostas geradas pelo PersonaReAct para diferentes perfis sobre o maior coeficiente de variação da velocidade no jogo contra o Ceará.

<b>Pergunta:</b> <i>Qual jogador apresentou o maior coeficiente de variação da velocidade no jogo contra o Ceará no dia 1º de abril de 2023?</i>	
<b>Resposta sem Personalização</b>	<i>O jogador que apresentou o maior coeficiente de variação da velocidade no jogo contra o Ceará no dia 1º de abril de 2023 foi Tomas Pochettino.</i>
<b>Torcedor do Time</b>	<i>No joguinho contra o Ceará no dia 1 de abril, quem andou oscilando bastante na velocidade foi o Tomas Pochettino. Teve momentos em que parecia um foguete e outros em que ficou meio perdido, sabe? A média dele foi de quase 1.94, mas o que impressiona mesmo é o desvio, que mostrou que ele não estava muito seguro ali na correria. Chato ver um jogador tão talentoso assim sem estabilidade, mas faz parte do jogo. Vamos torcer pra ele se encontrar mais nos próximos jogos!</i>
<b>Técnico do Time</b>	<i>O jogador que apresentou o maior coeficiente de variação da velocidade em campo no jogo contra o Ceará, realizado em 1º de abril de 2023, foi Tomas Pochettino. Ele alcançou um coeficiente de 0.5012, o que demonstra uma grande variação em sua intensidade de corrida durante a partida. Esses dados são importantes para analisarmos o desempenho dinâmico do atleta e sua capacidade de adaptação às necessidades do jogo.</i>
<b>Narrador de Jogo de Futebol</b>	<i>E aí, galera! No embate contra o Ceará no dia 1º de abril de 2023, o jogador que mais oscilou na velocidade foi o Titi. Aquele dia foi pra lá de movimentado! Ele teve uma média de velocidade que, embora não seja das mais altas, mostrou que ele estava em campo funcionando como um verdadeiro ioiô, ora esticando e ora encolhendo. Com 577 corridas registradas, dava pra ver que Titi não parou um segundo, mas as oscilações foram notáveis. Isso só comprova que o futebol às vezes é drama e emoção, e nesse dia o Titi foi um grande exemplo disso!</i>
<b>Influenciador Digital</b>	<i>E aí, galera! No jogo contra o Ceará no dia 1º de abril de 2023, o nosso craque mais instável na velocidade foi o Tomas Pochettino! Ele estava meio que acelerando e desacelerando o tempo todo, mostrando uma velocidade média de 1.94, com um desvio de 1.07. Então, dá pra notar que ele teve aqueles momentos de correr muito e depois dar uma segurada! É o famoso efeito 'freio e acelerador' em campo! Vamos ficar de olho nele, porque quando ele tá a mil, pode ser um grande diferencial!</i>
<b>Fisioterapeuta do Time</b>	<i>No jogo contra o Ceará em 1º de abril de 2023, o jogador que apresentou o maior coeficiente de variação da velocidade foi Tomas Pochettino, com um coeficiente de variação de 0.5012. Esse valor indica uma variabilidade significativa na sua performance, o que pode ser interpretado como um sinal de fadiga potencial. Em termos práticos, isso sugere que a execução dos movimentos dele teve uma oscilação considerável, o que geralmente ocorre em atletas que estão experimentando um nível alto de esforço, possivelmente associada a um período de alto desgaste físico.</i>
<b>Jornalista Esportivo</b>	<i>O jogador que teve a maior oscilação de velocidade no duelo contra o Ceará, realizado em 1º de abril de 2023, foi Bruno Pacheco. Ele apresentou uma notável variação de 5.26 km/h, o que demonstra seu intenso esforço tanto na marcação quanto ao ascender para o ataque. A habilidade de alternar entre essas funções foi crucial para a dinâmica do jogo, mostrando que Bruno não apenas esteve presente defensivamente, mas também se lançou ao campo ofensivo com grande intensidade e entrega.</i>

Fonte: elaborado pelo autor.

Quadro 28 – Respostas geradas pelo PersonaReAct para diferentes perfis sobre as acelerações e desacelerações de Bruno Pacheco no jogo contra o Ceará.

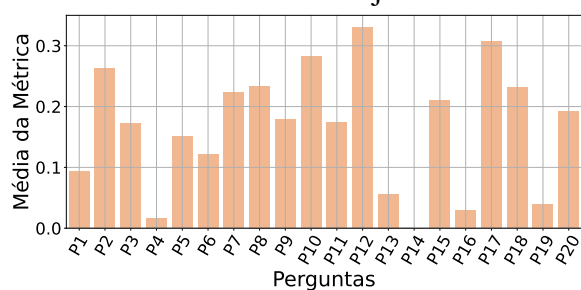
<b>Pergunta:</b> <i>Quantas vezes Bruno Pacheco atingiu altos níveis de aceleração e desaceleração durante a partida contra o Ceará em 1º de abril de 2023?</i>	
<b>Resposta sem Personalização</b>	<i>Bruno Pacheco atingiu altos níveis de aceleração 249 vezes e desaceleração 211 vezes durante a partida contra o Ceará em 1º de abril de 2023.</i>
<b>Torcedor do Time</b>	<i>Olha, na partida contra o Ceará no dia 1 de abril, o Bruno Pacheco foi um verdadeiro leão em campo! Ele não só fez 44 arrancadas fortes, mostrando sua velocidade e determinação, mas também teve 79 paradas, o que comprova o quanto ele se esforçou para marcar presença e controlar a jogada. Ele realmente deu o sangue pela nossa equipe! Sem dúvida, uma exibição de alto nível que fez a torcida vibrar!</i>
<b>Técnico do Time</b>	<i>Na partida contra o Ceará, realizada no dia 1º de abril de 2023, Bruno Pacheco executou um total de 460 ações de alta aceleração e desaceleração, sendo 249 delas referentes a acelerações e 211 a desacelerações. Esses números refletem uma performance intensa e dinâmica, crucial para a movimentação e o posicionamento em campo. Essa quantidade de ações é um indicativo do ritmo do jogo e do papel ativo que o jogador teve durante a partida.</i>
<b>Narrador de Jogo de Futebol</b>	<i>No duelo contra o Ceará no dia 1º de abril de 2023, Bruno Pacheco mostrou-se um verdadeiro maestro da aceleração e frenagem em campo. Ele não hesitou em fazer 103 acelerações explosivas, sempre buscando ganhar espaço e proporcionar dinamismo ao ataque da sua equipe. Além disso, foi preciso em 95 das suas frenagens, sabendo exatamente o momento de segurar o ritmo e evitar que a jogada se tornasse um contra-ataque do adversário. Uma performance que deixou a torcida vibrando e a comissão técnica satisfeita!</i>
<b>Influenciador Digital</b>	<i>E aí! Com certeza, o Pacheco deu um show de arrancadas e freadas no jogo contra o Ceará! Ele apresentou uma média alta de aceleração de 0.7855, o que significa que ele realmente se lançou com tudo para frente, como um foguete! E não parou por aí, a média de desaceleração dele ficou em 0.6656, o que indica que ele também mandou bem nas freadas, conseguindo controlar a velocidade com bastante habilidade. Resumindo: o desempenho dele foi bem robusto, mostrando que estava afim de jogo e fazendo valer cada minuto em campo. Dados como esses são ouro mesmo para avaliar a performance! Vamos acompanhar de perto os próximos jogos!</i>
<b>Fisioterapeuta do Time</b>	<i>Na partida contra o Ceará no dia 1º de abril de 2023, Bruno Pacheco demonstrou um alto nível de atividade física, realizando um total de 249 acelerações e 211 desacelerações intensas. Esses números são cruciais para a nossa análise de carga neuromuscular, pois indicam o grau de exigência que o atleta enfrentou durante o jogo. A elevada quantidade de acelerações e desacelerações intensas pode estar diretamente relacionada à sua capacidade de resposta rápida e mudanças de direção, características essenciais no futebol. É fundamental monitorarmos esses dados para avaliarmos a fadiga e a recuperação do jogador; bem como para planejarmos o treinamento e a prevenção de lesões. Dessa forma, conseguimos otimizar o desempenho e a saúde do atleta ao longo da temporada.</i>
<b>Jornalista Esportivo</b>	<i>No jogo contra o Ceará, realizado no dia 1º de abril de 2023, Bruno Pacheco teve um desempenho notável, acelerando ou freando bruscamente impressionantes 317 vezes. Essa quantidade não apenas revela a intensidade do seu esforço físico, mas também exemplifica como ele se envolveu ativamente na dinâmica da partida, sempre buscando contribuir para a movimentação do time. Esses dados são fundamentais para entender a participação dele no jogo, pois refletem não apenas a capacidade atlética, mas também a necessidade de manter a fluidez ofensiva e defensiva em um jogo tão competitivo.</i>

Fonte: elaborado pelo autor.

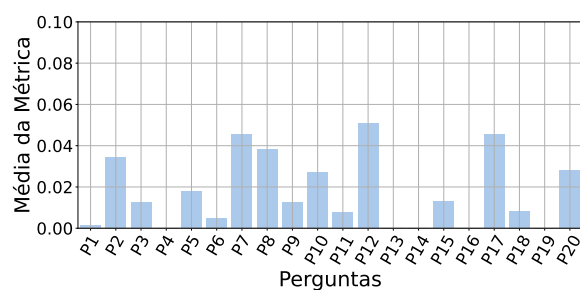
## APÊNDICE E – RESULTADOS DA ANÁLISE DE SIMILARIDADE TEXTUAL PARA ESTUDO DE CASO 2

Neste apêndice são apresentados apenas os gráficos referentes ao Estudo de Caso 2 (domínio médico). As análises detalhadas, já discutidas no texto principal para o Estudo de Caso 1 (domínio esportivo), aplicam-se de forma bastante similar neste contexto. Dessa forma, optou-se por não repetir as interpretações, preservando a concisão do trabalho e evitando redundâncias.

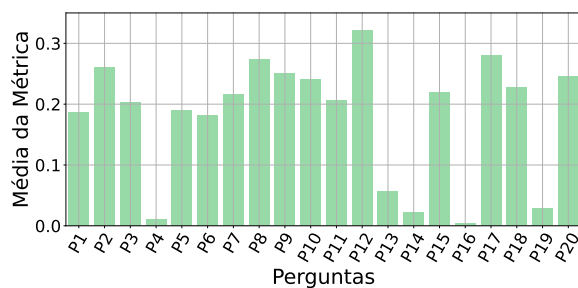
Figura 27 – Gráfico de barras das métricas de similaridade textual por pergunta para o modelo o3 mini no contexto do conjunto de dados do domínio médico.



(a) ROUGE-L



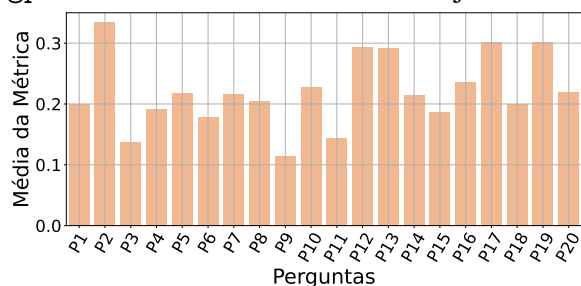
(b) BLEU



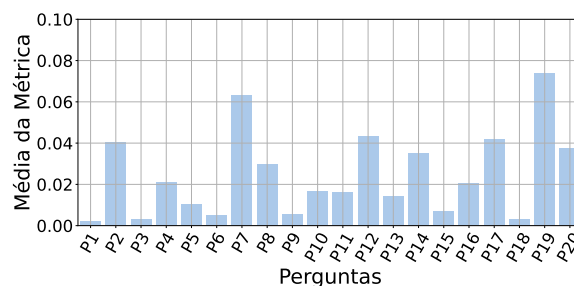
(c) Distância de Edição

Fonte: elaborado pelo autor.

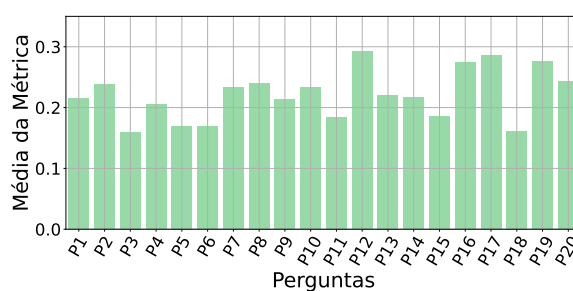
Figura 28 – Gráfico de barras das métricas de similaridade textual por pergunta para o modelo gpt-4.1-mini no contexto do conjunto de dados do domínio médico.



(a) ROUGE-L



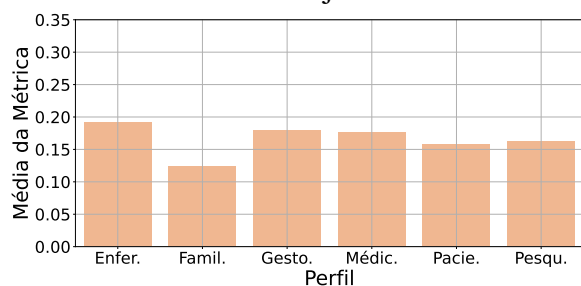
(b) BLEU



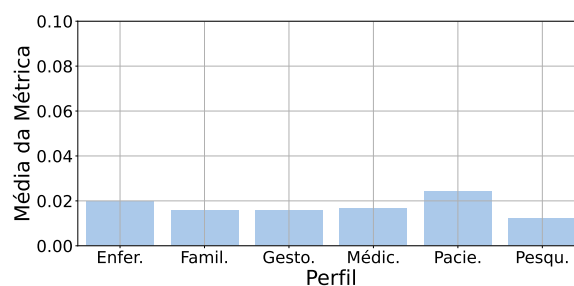
(c) Distância de Edição

Fonte: elaborado pelo autor.

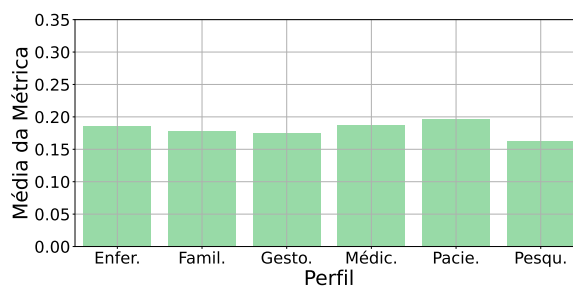
Figura 29 – Gráfico de barras das métricas de similaridade textual por perfil para o modelo o3 mini no contexto do conjunto de dados do domínio médico.



(a) ROUGE-L



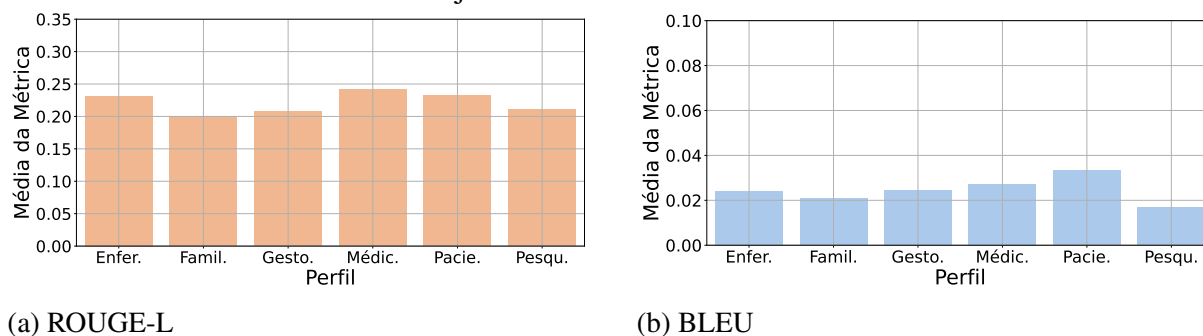
(b) BLEU



(c) Distância de Edição

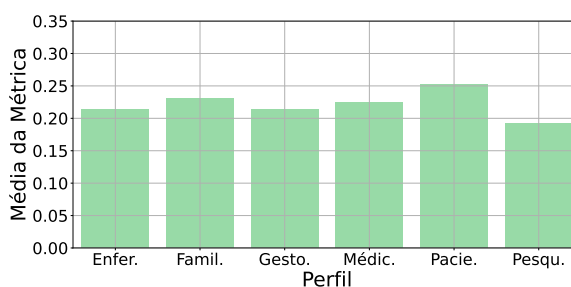
Fonte: elaborado pelo autor.

Figura 30 – Gráfico de barras das métricas de similaridade textual por perfil para o modelo GPT-4.1 mini no contexto do conjunto de dados do domínio médico.



(a) ROUGE-L

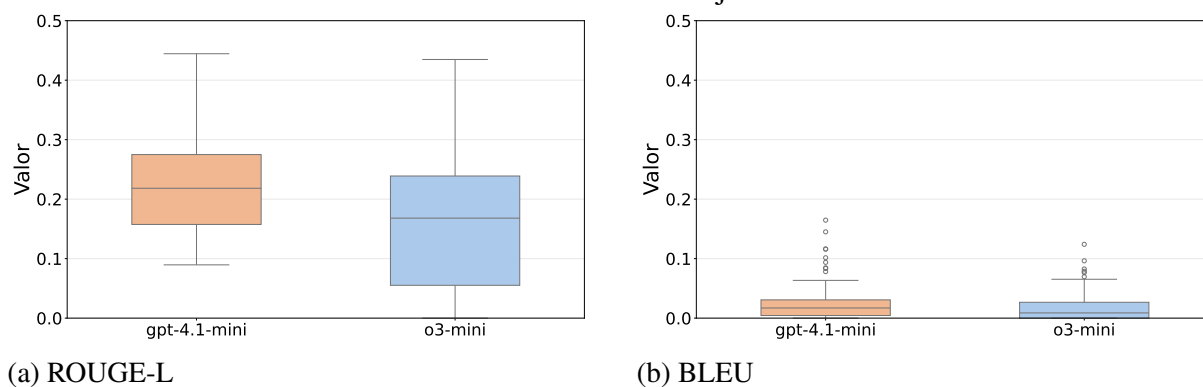
(b) BLEU



(c) Distância de Edição

Fonte: elaborado pelo autor.

Figura 31 – Distribuição Geral das Métricas (ROUGE-L, BLEU, Distância de Edição) para os modelos o3 mini e GPT-4.1 mini no contexto do conjunto de dados do domínio médico.



(a) ROUGE-L

(b) BLEU

(c) Distância de Edição

Fonte: elaborado pelo autor.