Universidade Estadual de Londrina

FELIPE AUGUSTO CRUZ

GENAI E ACESSIBILIDADE: UMA ABORDAGEM INCLUSIVA PARA DOCUMENTOS TEXTUAIS

LONDRINA
2024

FELIPE AUGUSTO CRUZ

GENAI E ACESSIBILIDADE: UMA ABORDAGEM INCLUSIVA PARA DOCUMENTOS TEXTUAIS

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof(a). Dr(a). Helen C. de Mattos Senefonte

LONDRINA

2024

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Cruz, Felipe Augusto.
GenAI e Acessibilidade: Uma Abordagem Inclusiva para Documentos Textuais / Felipe Augusto Cruz. - Londrina, 2024.
59 f. : il.

Orientador: Helen C. de Mattos Senefonte.
Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Graduação em Ciência da Computação, 2024.
Inclui bibliografia.

1. Modelo de Inteligência Artificial Generativa - TCC. 2. Acessibilidade a Neurodivergências - TCC. 3. Quantização em Machine Learning - TCC. 4. Fine-tuning - TCC. I. Senefonte, Helen C. de Mattos. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Graduação em Ciência da Computação. III. Título.

CDU 519

FELIPE AUGUSTO CRUZ

GENAI E ACESSIBILIDADE: UMA ABORDAGEM INCLUSIVA PARA DOCUMENTOS TEXTUAIS

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA

Orientador: Prof(a). Dr(a). Helen C. de Mattos Senefonte
Universidade Estadual de Londrina

Prof. Dr. Vitor Valério de Souza Campos
Universidade Estadual de Londrina– UEL

Prof(a). Dr(a). Vanessa Matias Leite
Universidade Estadual de Londrina– UEL

Londrina, 24 de abril de 2024.

AGRADECIMENTOS

Agradeço aos professores do Departamento de Computação da Universidade Es tadual de Londrina pelos ensinamentos. Em especial, agradeço minha professora orientadora Helen C. de Mattos Senefonte pela disponibilidade, e por todo o suporte necessário para o desenvolvimento desse trabalho.

Agradeço aos meus colegas que me ajudaram e sempre foram gentis.

Por fim, gostaria de expressar minha profunda gratidão à minha família, cujo apoio incondicional e incentivo constante foram fundamentais em minha jornada. Vocês não me deixaram desistir nos momentos mais difíceis e, sem vocês, não teria sido possível conquistar este sonho.

"Porque a sua ira dura só um momento; no seu favor está a vida. O choro pode durar uma noite, mas a alegria vem pela manhã. (Bíblia Sagrada, Salmos 30, 5))

CRUZ, F. A.. **GenAI e Acessibilidade: Uma Abordagem Inclusiva para Documentos Textuais**. 2024. 58f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina, 2024.

RESUMO

Nos últimos anos, o uso de Inteligência Artificial Generativa (*GenAI*) tornou-se cada vez mais popular no cotidiano, seja por meio de serviços diretos, como o ChatGPT da OpenAI, ou de forma indireta, como assistentes em aplicações de escritório. No entanto, ainda existem diversas áreas inexploradas ou com lacunas que impedem um desempenho satisfatório da *GenAI*. Uma dessas áreas é a acessibilidade a documentos textuais para indivíduos com neurodivergências, que frequentemente enfrentam dificuldades na compreensão durante a leitura. Este trabalho tem como objetivo realizar um estudo para avaliar as possibilidades de utilização da GenAI no contexto da acessibilidade, com foco em documentos textuais, avaliando trabalhos correlatos sobre a medição da compreensão em leituras e encontrar métodos para implementar um modelo para adaptar os textos fornecidos para padrões que melhorem a compreensão.

Palavras-chave: Inteligência artificial generativa. GenAI. Educação inclusiva. Neurodirgência. GPT. ChatGPT. Large language model. LMM. Transformer. Deep Learning.

CRUZ, F. A.. **GenAI and Accessibility: An Inclusive Approach to Textual Documents**. 2024. 58p. Final Project (Bachelor of Science in Computer Science) – State University of Londrina, Londrina, 2024.

ABSTRACT

In recent years, the use of Generative Artificial Intelligence (GenAI) has become increasingly popular in everyday life, whether through direct services like OpenAI's ChatGPT or indirectly, as assistants in office applications. However, there are still many unexplored areas or gaps that hinder the satisfactory performance of GenAI. One such area is the accessibility of textual documents for individuals with neurodivergences, who often face comprehension difficulties while reading. This work aims to conduct a study to assess the possibilities of using GenAI in the context of accessibility, focusing on textual documents, evaluating related work on measuring reading comprehension, and finding methods to implement a model to adapt the provided texts to standards that enhance understanding.

Keywords: Generative Artificial Intelligence. GenAI. Inclusive Education. Neurodivergence. GPT. ChatGPT. Large Language Model. LLM. Transformer. Deep Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

GenAI	Inteligência Artificial Generativa
TEA	Transtorno do Espectro Autista
TDA	Trantorno do Déficit de Atenção
TDAH	Trantorno do Déficit de Atenção e Hiperatividade
LLM	Large Language Model
GPT	Generative Pre-trained Transformer
IDEA	Individuals with Disabilities Education Act
TE	Tamanho do Efeito
PND	Porcentagem de Pontos de Dados Não Sobrepostos
ID	Instrução Direta
QLoRA	Quantized Low-Rank Adaptation
GQA	Grouped-query Attention
RNAs	Redes Neurais Artificiais
PEFT	Parameter-Efficient Fine-Tuning

SUMÁRIO

1 INTRODUÇÃO

Indivíduos com neurodivergências enfrentam desafios significativos na leitura, impactando não apenas sua capacidade de ler, mas também sua compreensão sobre o que foi lido. Essa questão não apenas limita o processo de aprendizado, mas também tem um impacto especialmente durante a infância e o ínicio da adolecência. É evidente que tanto os próprios indivíduos quanto os educadores podem se beneficiar do uso de modelos de inteligência artificial para promover uma maior inclusão dessas pessoas em uma sociedade cada vez mais digitalizada.

A neurodivergência é um conceito fundamental na compreenção da diversidade cognitiva humana. Ela engloba uma gama de variações naturais no funcionamento do cérebro, resultando em diferentes maneiras de processar informações, comunicar-se e interagir com o mundo. Trata-se antes de uma diferença humana que deve ser respeitada como outras diferenças (sexuais, raciais, entre outras). Os indivíduos autodenominados "neurodiversos" consideram-se "neurologicamente diferentes", ou "neuroatípicos"[5]. Essas variações podem abranger condições como o Transtorno do Espectro Autista (TEA), a Dislexia, o Transtorno de Déficit de Atenção de Hiperatividade (TDAH) entre outras. A neurodivergência desafia as noções convencionais de normalidade, destacando a importância de reconhecer e respeitar as diversas formas de cognição.

Ao longo deste trabalho é apresentado neurodivergências, descrevendo os sintomas e como estudos demonstram as dificuldades na leitura e compreensão. Também são mencionados os aspectos a serem considerados na adaptação de textos. Posteriormente, é apresentado o funcionamento das inteligências artificiais generativas, incluindo a evolução dos modelos utilizando Transformers.

É avaliado a possibilidade de encontrar padrões e elaborar um modelo de inteligência artificial generativa, comumente referido como *GenAI*, para aplicar esses padrões e promover uma independência para esses indivíduos, pois materiais com acessibilidade são escassos.

Com o avanço dos modelos e técnicas de GenAI surgiram muitas aplicações em que se produziam redações e criação de imagens, mas um problema comum entre elas é que essa tecnologia resolve problemas opcionais e, apesar de ser útil nessas áreas, não traz um impacto onde realmente contribua para a sociedade. No entanto, a adaptação de texto para essas pessoas é algo vital, necessário e pode trazer benefícios reais. Com esse estudo, espera-se trazer um aspecto social a essa tecnologia.

Foi treinado um modelo de Inteligência Artificial Generativa utilizando o método de *fine-tuning* utilizando os padrões encontrados nos estudos correlatos para adaptar os

textos fornecidos ao modelo, documentando o processo necessário para fazer o treinamento.

No capítulo 2, apresenta-se a fundamentação teórica necessária para contextualizar os desafios alvos de resolução e os conceitos abordados durante este estudo. No capítulo 3, discutem-se trabalhos correlatos sobre a capacidade de compreensão de alunos durante a leitura. No capítulo 4, aborda-se a metodologia aplicada. No capítulo 5, apresentam-se os experimentos realizados. Por fim, no capítulo 6, analisam-se os resultados obtidos.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Introdução

O capítulo aborda os conceitos utilizados no trabalho, inicialmente abordando neurodivergências, destacando como variações no funcionamento cerebral, são cada vez mais reconhecidas e respeitadas dentro da psiquiatria e outros campos sociais. Neurodivergência inclui condições como o Transtorno do Espectro Autista (TEA), Transtorno de Déficit de Atenção e Hiperatividade (TDAH), e dislexia, cada um com características únicas que afetam a interação social, a concentração, ou a leitura e escrita, respectivamente.

Além disso, o capítulo discute avanços em inteligência artificial generativa. Destacam-se os Modelos de Linguagem de Grande Escala (LLMs), com uma ênfase particular nos modelos *Transformer* e em técnicas como *fine-tuning* para adaptar esses modelos a tarefas específicas sem o alto custo de treinamento do zero.

Na Seção 1, aborda-se o conceito de neurodivergência e são explicadas as neurodivergências aplicadas neste trabalho. Na Seção 2, serão abordados os conceitos de inteligência artificial generativa e os métodos utilizados para otimizar o treinamento e a utilização.

2.2 Neurodivergências

A neurodivergência é um conceito que enfatiza a diversidade no funcionamento cerebral humano, destacando que as variações neurológicas estão fora do que é considerado padrão ou típico, conhecido como neurotipicidade. O neurodesenvolvimento e os transtornos do neurodesenvolvimento são cada vez mais reconhecido como sendo importante dentro do *mainstream* de psiquiatria[1].

As variações abrangidas pela neurodivergência implicam um conjunto único de características no modo como o cérebro processa informações, percebe o mundo e interage com ele. Estas diferenças são vistas como parte natural e inerente da variabilidade humana, desafiando a noção de que existe uma única maneira "normal"ou "saudável"de funcionamento cerebral.

Neurotipicidade, em contraste, refere-se ao funcionamento neurológico que está em conformidade com o que a maioria das pessoas experimenta, frequentemente visto como o padrão. No entanto, a ideia de neurodivergência sugere que as variações neurológicas não são anormais ou patológicas, mas sim diferenças que devem ser reconhecidas e respeitadas.

A "Neurodiversidade"foi cunhada pela primeira vez em 1998 por Judy Singer, so-

cióloga e uma pessoa autista, que escreveu que as pessoas autistas eram uma categoria semelhante à classe/gênero/raça. Isso levou ao desenvolvimento de uma série de termos específicos, nomeadamente neurodiversidade, neurodiverso, neurodivergente e neurotípico, que descrevem subconjuntos da população normal delimitados pelos limites das normas sociais atuais[1].

A relevância social e cultural da neurodivergência é significativa, pois tem implicações em termos de inclusão social, educação, emprego e políticas de saúde. Promove uma maior compreensão e aceitação das diferenças neurológicas, encorajando ambientes mais inclusivos e adaptativos para todos, independentemente de suas características neurológicas. A figura 1 demonstra o modelo da Equipe Nacional de Implementação do Autismo ilustrando a relação entre uma função neurocognitiva e os termos neurodiversidade, neurotípico e neurodivergente.

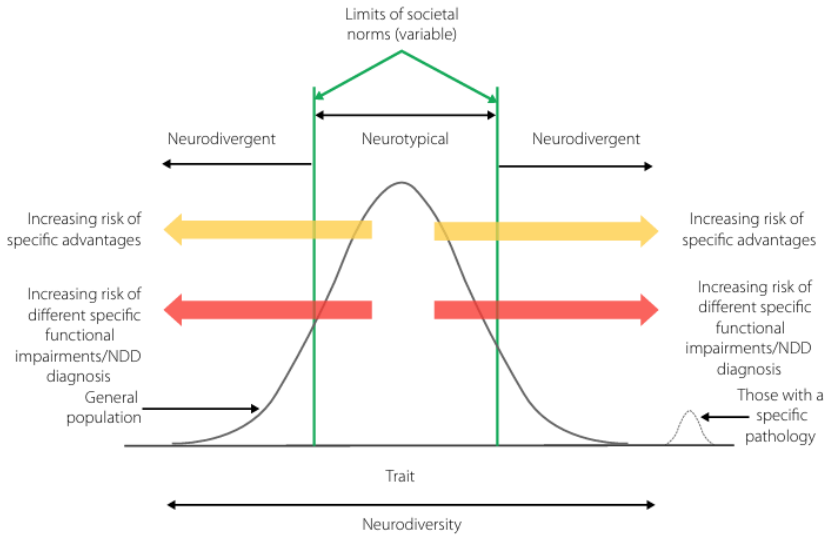

Figura 1 – O modelo da Equipe Nacional de Implementação do Autismo. [1]

2.2.1 Transtorno do Espectro Autista

TEA é uma condição neurológica e de desenvolvimento que se manifesta precocemente na infância e persiste ao longo da vida. É caracterizado por prejuízos na interação social e comunicação, assim como por padrões de comportamento restritos, repetitivos

e estereotipados[6]. O espectro é amplo, variando significativamente entre indivíduos em termos de habilidades e desafios.

Pessoas com TEA podem ter dificuldades em compreender e usar a linguagem verbal e não verbal, como gestos e expressões faciais. Algumas podem não falar ou falar pouco, enquanto outras podem ter habilidades de fala avançadas, mas dificuldade em usar a linguagem em contextos sociais.

Sobre a compreensão há um amplo consenso de que crianças com autismo têm compreensão de leitura prejudicada apesar da natureza de alto funcionamento, os níveis de compreensão de leitura foram significativamente mais baixos nas crianças com autismo em relação aos controles correspondentes em QI[7].

2.2.2 Transtorno de Déficit de Atenção e Hiperatividade (TDAH)

TDAH é uma condição neurológica que afeta a capacidade de uma pessoa de manter a atenção, controlar impulsos e regular o comportamento. Os sintomas comuns do TDAH incluem distração, impulsividade e hiperatividade, que podem interferir em várias áreas da vida, incluindo educação e trabalho.

Indivíduos com TDAH podem enfrentar vários desafios educacionais, que incluem uma série de dificuldades. A dificuldade de concentração, um aspecto central do transtorno, pode tornar o aprendizado em ambientes tradicionais particularmente desafiador, pois a manutenção da atenção em tarefas específicas é frequentemente complicada.

Estudantes com TDA/TDAH também precisam de horários diários estruturados e métodos de instrução consistentes. Eles precisam de uma atmosfera de sala de aula na qual se sintam confortáveis para correr riscos[8].

2.2.3 Dislexia

A dislexia é um transtorno de aprendizagem específico caracterizado por dificuldades com a precisão e/ou fluência na leitura e a habilidade de decodificar palavras. Esta condição é neurobiológica e muitas vezes hereditária, afetando a forma como o cérebro processa a linguagem escrita e falada.

Embora a dislexia não afete a inteligência, ela pode causar desafios significativos na aprendizagem, especialmente em ambientes educacionais tradicionais que dependem fortemente da leitura e escrita. Pessoas com dislexia podem ter dificuldade em reconhecer palavras, soletrar e decifrar rapidamente textos, o que pode afetar seu desempenho acadêmico[9].

Existem diferentes níveis de severidade na dislexia, variando de leve a severo. Alguns indivíduos podem ter apenas pequenos problemas com algumas palavras ou com

a leitura sob pressão, enquanto outros podem achar extremamente difícil ler até mesmo textos simples. O diagnóstico e o apoio adequado podem ajudar significativamente.

2.3 Modelos de Inteligência Artificial Generativa

A teoria das *GenAI* representa uma abordagem significativa no campo da Inteligência Artificial (IA), cujo foco principal reside na criação de modelos capazes de gerar novos dados, em contraposição à simples classificação ou previsão de dados já existentes. Os modelos que se enquadram nessa categoria são denominados generativos devido à sua capacidade intrínseca de produzir informações inéditas que compartilham semelhanças com os dados de treinamento originais[10].

A base teórica da *GenAI* se apoia em uma série de conceitos fundamentais e técnicas de modelagem. Uma das ideias centrais por trás dessa abordagem é que, para uma compreensão abrangente de um conjunto de dados, é essencial compreender o processo subjacente que levou à sua geração. Dessa forma, em vez de apenas buscar prever ou classificar dados, os modelos generativos buscam desvendar as relações subjacentes entre os dados e como essas relações podem ser utilizadas para criar novos dados.

2.3.1 Inteligência Artificial Generativa

A Inteligência Artificial Generativa refere-se a um conjunto de métodos de IA que são capazes de gerar novos conteúdos, dados ou informações que são coerentes e muitas vezes indistinguíveis dos dados gerados por humanos. Essa tecnologia abrange desde a criação de imagens e música até a escrita de textos e a modelagem de sequências biológicas.

A história da Inteligência Artificial Generativa remonta à segunda metade do século XX, quando os primeiros modelos conceituais de redes neurais começaram a ser desenvolvidos. No entanto, foi apenas nas últimas décadas que a tecnologia avançou significativamente, impulsionada pelo aumento da capacidade de processamento computacional e pelo acesso a grandes quantidades de dados.

Nos anos 2010, a introdução das Redes Generativas Adversariais (GANs) marcou um ponto de virada importante. Essas redes utilizam duas redes neurais em confronto: uma geradora, que tenta criar dados falsos, e uma discriminadora, que tenta distinguir entre dados reais e falsos. Esse mecanismo impulsionou a capacidade das máquinas de criar conteúdos altamente realistas.

Recentemente, modelos como o GPT (*Generative Pre-trained Transformer*) revolucionaram ainda mais o campo, especialmente no que diz respeito à geração de texto. Esses modelos são treinados em grandes conjuntos de dados textuais e são capazes de gerar conteúdos escritos que podem imitar diversos estilos e formatos.

2.3.2 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são algoritmos computacionais que apresentam um modelo matemático inspirado na estrutura de organismos inteligentes, os quais possibilitam inserir simplificadamente o funcionamento do cérebro humano em computadores. Dessa forma, a exemplo do cérebro humano, a RNA é capaz de aprender e tomar decisões baseadas em seu próprio aprendizado[11].

Assim como ocorre no processo de aprendizagem humano, as RNAs se adaptam e melhoram suas capacidades com a experiência. Por meio de um processo chamado "treinamento", as RNAs ajustam seus pesos sinápticos — equivalentes a conexões neurais no cérebro humano — baseando-se na entrada de dados e nos resultados obtidos. Esse processo é iterativo e pode envolver milhares ou mesmo milhões de ciclos de ajustes, permitindo que a rede desenvolva uma espécie de "intuição"para resolver problemas específicos ou reconhecer padrões complexos em dados, variando de reconhecimento de voz e imagem até a tomada de decisões em ambientes dinâmicos e imprevisíveis.

2.3.3 Large Language Models

Nos últimos anos, os Modelos de Linguagem de Grande Escala (LLMs) emergiram como uma área de pesquisa essencial no campo da inteligência artificial e do processamento de linguagem natural. Esses modelos, projetados para compreender e gerar texto em linguagem natural, utilizam uma abordagem de aprendizado de máquina baseada em redes neurais[12].

A capacidade notável dos LLMs de produzir textos coerentes e relevantes em diversas tarefas, tais como tradução automática, resposta a perguntas e geração de texto, revolucionou a interação das máquinas com a linguagem humana. O fascínio por esses modelos advém da sua habilidade de aprender automaticamente a estrutura e o significado da linguagem natural a partir de extensos conjuntos de dados não rotulados, e da sua adaptabilidade a tarefas específicas mediante o processo de ajuste fino (*fine-tuning*).

A construção dos LLMs é viabilizada pelo uso de uma arquitetura de redes neurais denominada *Transformer*, apresentada primeiramente no artigo "*Attention is All You Need*", publicado em 2017 por Ashish Vaswani e colaboradores.

O *Transformer* fundamenta-se exclusivamente em mecanismos de atenção, dispensando a necessidade de recorrência ou convolução.

Este modelo emprega um tipo específico de atenção, conhecido como "*Multi-Head Attention*", para captar diferentes aspectos da informação contextual de várias posições na sequência de entrada. Isso permite ao modelo compreender de maneira mais eficaz as relações entre todas as partes de uma sequência[13].

2.3.4 Transformer

O modelo *Transformer* é composto de duas partes principais, o *encoder* e o *decoder* como pode ser observado na figura 2.

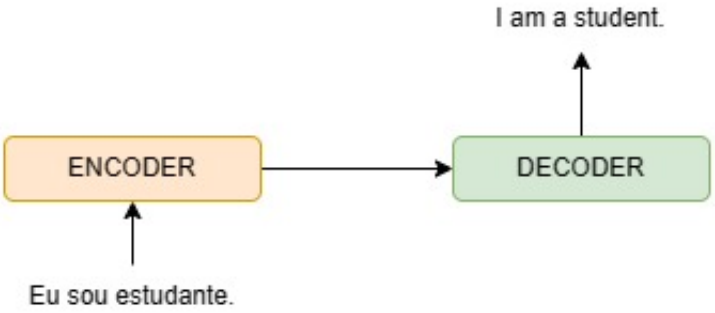

Figura 2 – Encoder e Decoder como proposto no artigo "Attention Is All You Need".

Os encoders no *Transformer* são componentes que processam a sequência de entrada, transformando-a em uma representação rica em informações e contexto. O *encoder* é composto por uma pilha de N = 6 camadas idênticas. Cada camada possui duas subcamadas. A primeira é um mecanismo de autoatenção multi-cabeças, e a segunda é uma rede *feed-forward* totalmente conectada simples e posicional[13].

Os *decoders* no *Transformer* transformam as representações codificadas pelos *encoders* em uma sequência de saída, como um texto traduzido. O decodificador também é composto por uma pilha de N = 6 camadas idênticas. Além das duas subcamadas em cada camada do *encoder*, o decodificador insere uma terceira subcamada, que realiza atenção multi-cabeças sobre a saída da pilha do *encoder*. Semelhante ao *encoder*, empregamos conexões residuais ao redor de cada uma das subcamadas, seguidas por normalização de camada[13].

Em camadas de "atenção *encoder-decoder*", as consultas vêm da camada anterior do decodificador, e as chaves e valores de memória vêm da saída do *encoder*[13] como pode ser observado na figura 3.

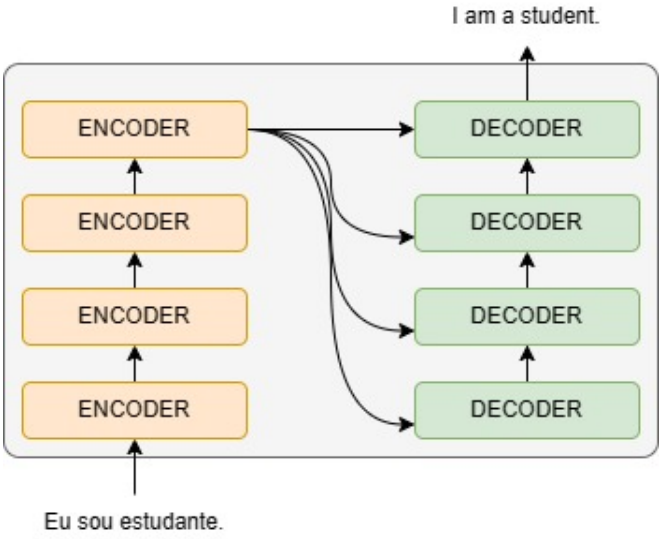

Figura 3 – Exemplo de várias camadas usadas para capturar padrões da entrada baseado na camada anterior.

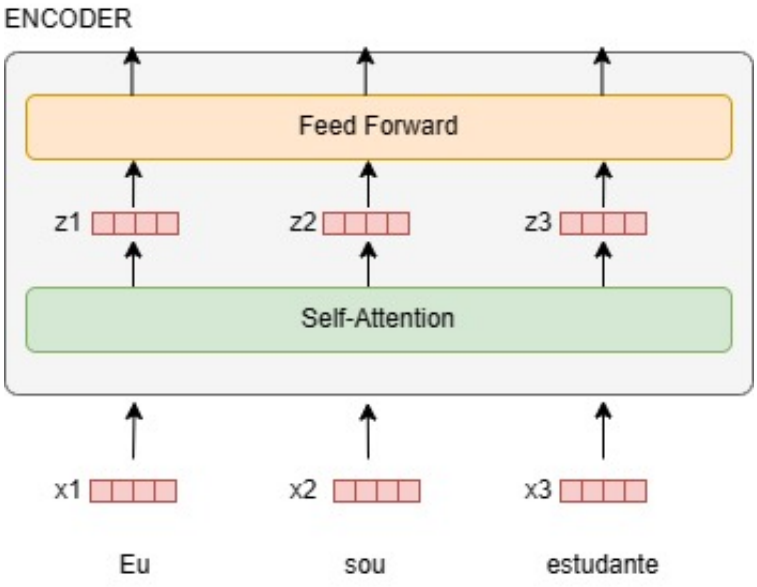

Figura 4 – Funcionamento do *encoder*.

A figura 4 representa o funcionamento do *encoder* que recebe uma entrada tokens que entram em uma sub-camada chamada *Self-Attention* que irão gerar um vetor para cada um dos *tokens* que depois passaram por outra camada chamada *Feed Forward*.

A camada *Feed Foward* funciona como uma rede neural totalmente conectada (ou densa) e tem a função de processar individualmente as informações de cada posição da sequência, após terem passado pelo mecanismo de atenção. Que pode ser expressa pela equação 2.1:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{2.1}$$

A sub-camada mais importante será a camada *Self-Attention*. A intuição por trás dessa técnica é a seguinte:

Suponha que tem uma sentença de entrada. Cada palavra na sentença recebe uma representação, baseada não só nela mesma, mas também nas outras palavras da sequência. Essa representação é gerada somando as representações das outras palavras, mas essa soma é ponderada. Palavras mais relevantes para a compreensão de uma palavra específica na sentença têm mais peso.

Por exemplo, na sentença "O gato é um animal", a palavra "animal" tem mais peso na geração da representação da palavra "gato", pois é diretamente relevante para seu significado. Palavras importantes para a compreensão de uma palavra-alvo são enfatizadas, enquanto outras menos relevantes podem ser ignoradas ou receber menos peso. A figura 5 demonstra um exemplo do funcionamento do *Self-Attention* disponibilizado em inglês.

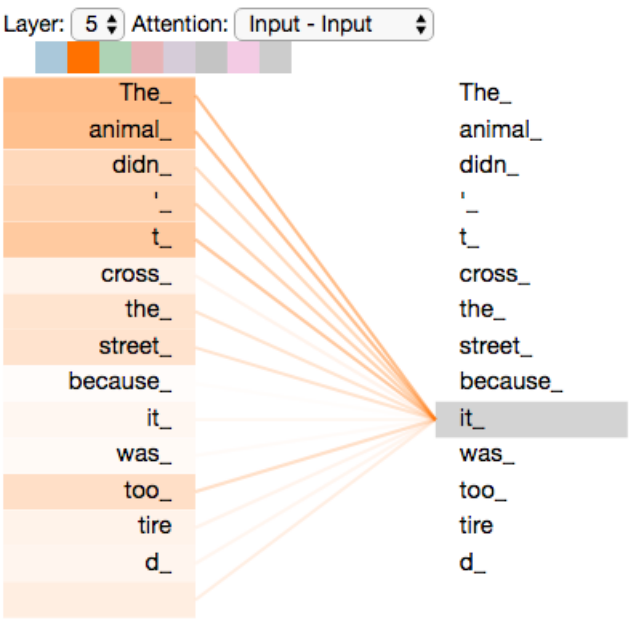

Figura 5 – Exemplo de funcionamento do *Self-Attention* disponibilizado no endereço eletrônico http://jalammar.github.io/illustrated-transformer .

Calcula-se essas representações para cada palavra na sentença, baseadas no contexto fornecido pelas outras palavras. Essa abordagem tenta imitar como nossa mente processa o texto, fazendo conexões intuitivas entre palavras relacionadas. Assim, a representação de uma palavra pode ser melhor compreendida quando consideramos seu contexto.

Em resumo, o "*Self-Attention*" permite que cada palavra na sentença tenha uma representação que captura informações de forma diferenciada, baseando-se na importância relativa de cada palavra no contexto da sentença. Cada palavra recebe um peso, e esses pesos são calculados através de matrizes específicas, permitindo uma representação mais rica e contextualizada do texto.

2.3.5 Fine-tuning

Frente ao fato que modelos de inteligência artificial generativa como o GPT tem alto custo associado, uma solução viável é explorar projetos de código aberto de LLMs de propósito geral e adaptá-los ao objetivo deste trabalho. Para tal, seria necessário empregar o processo de '*fine-tuning*'. Dentre as diversas abordagens possíveis para '*fine-tuning*'.

O treinamento completo de modelos de linguagem é oneroso, tanto financeiramente quanto em termos de tempo, sem mencionar a significativa pegada de carbono associada. Portanto, a opção mais sustentável e viável seria se concentrar apenas no "*fine-tuning*"[14].

Com essa abordagem, é possível treinar um modelo como o Mistral 7B, que se destaca por suas licenças permissivas. Apesar de seus 7 bilhões de parâmetros o classificarem como um modelo relativamente pequeno comparado aos LLMs contemporâneos, o Mistral 7B é reconhecido pela comunidade como um dos modelos mais eficientes.

2.3.6 Mistral 7B

É um modelo de linguagem de grande escala desenvolvido pela Mistral AI, caracterizado por ter aproximadamente 7,3 bilhões de parâmetros. Este modelo é notável por suas melhorias em eficiência e desempenho, superando outros modelos semelhantes como o Llama 2 13B em várias métricas de avaliação. Mistral 7B é particularmente forte em tarefas de raciocínio, matemática e geração de código .

Uma das inovações técnicas do Mistral 7B é o uso do mecanismo de atenção chamado *Grouped-query Attention (GQA)*, que permite tempos de inferência mais rápidos em comparação com o modelo de atenção completo padrão . Além disso, o modelo também utiliza a técnica de *Sliding Window Attention* para lidar eficazmente com sequências mais longas. Na figura 6 pode observar que o número de operações na "vanilla attention"é quadrático em relação ao "sequence length"e a memória aumenta linearmente com o número de tokens. No momento da inferência, isso resulta em maior latência e menor capacidade de processamento devido à redução da disponibilidade de cache.

Este modelo é distribuído sob a licença Apache 2.0, permitindo seu uso sem restrições em uma variedade de plataformas e aplicações.

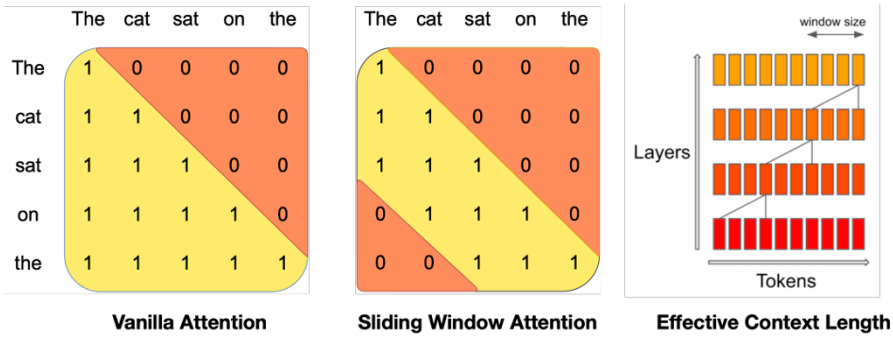

Figura 6 – Exemplo de "*vanilla attention*". [2].

2.3.7 QLoRA

LoRA e QLoRA são técnicas que se destacam por sua capacidade de otimizar modelos de até 65 bilhões de parâmetros em uma única GPU de 48GB, enquanto preserva o desempenho comparável ao do *fine-tuning* tradicional de 16 bits.

Essa técnica se baseia em uma quantização de 4 bits, que permite a propagação de gradientes através de um modelo pré-treinado e quantizado para Adaptações de Baixo Rank (LoRA), que são pequenos conjuntos de pesos ajustáveis.

QLoRA incorpora inovações como a quantização dupla e otimizadores paginados, que são estratégias para gerenciar eficientemente o uso de memória. Demonstrou-se que, aplicando o QLoRA, é possível treinar mais de 1.000 modelos em diferentes escalas e arquiteturas, alcançando resultados que definem o novo estado da arte, particularmente no desempenho de chatbots, mesmo com uma redução significativa nos recursos computacionais necessários. Diferentes métodos de *fine-tuning* e seus requisitos de memória podem ser observados na figura 7 onde o método QLoRA melhora em relação ao LoRA ao quantizar o modelo de transformador para precisão de 4 bits e usar otimizadores paginados para lidar com picos de memória.

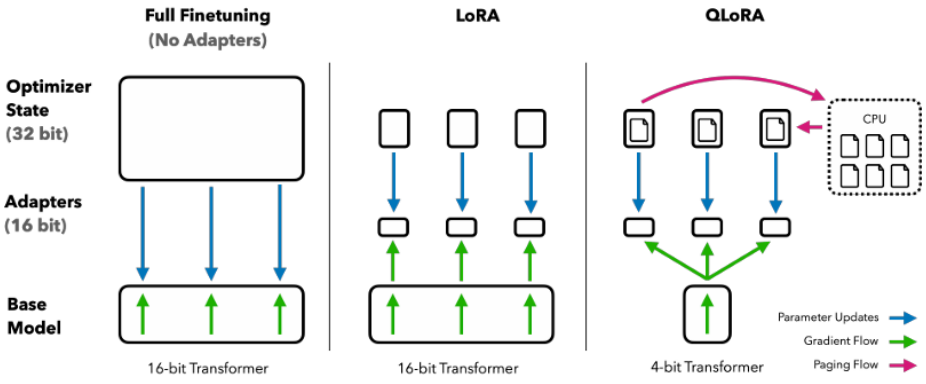

Figura 7 – Comparação entre diferentes métodos de *fine-tuning*. [3].

2.3.8 PEFT

Para realizar o processo de *fine-tuning* completo em um modelo pode ser necessário muito de recurso computacional. Os enormes requisitos de recursos computacionais são proibitivos para qualquer um. Para enfrentar esse desafio, um método proeminente conhecido como *Parameter-Efficient Fine-Tuning* (PEFT) surgiu como uma solução viável para compensar o tremendo custo computacional do *fine-tuning* completo de parâmetros. O PEFT envolve a utilização de várias técnicas de aprendizado profundo para reduzir o número de parâmetros treináveis, mantendo ainda um desempenho comparável ao do ajuste fino completo[4].

O PEFT atualiza apenas um pequeno número de parâmetros adicionais ou atualiza um subconjunto dos parâmetros pré-treinados, preservando o conhecimento capturado pelo PLM enquanto o adapta para a tarefa alvo e reduzindo o risco de esquecimento catastrófico. Além disso, como o tamanho do conjunto de dados ajustado é tipicamente muito menor do que o conjunto de dados pré-treinado, realizar um *fine-tuning* completo para atualizar todos os parâmetros pré-treinados pode levar ao superajuste, o que é evitado pelo PEFT através da atualização seletiva ou não dos parâmetros pré-treinados. Na figura 8 pode visualizar o desenvolvimento evolutivo dos métodos PEFT nos últimos anos.

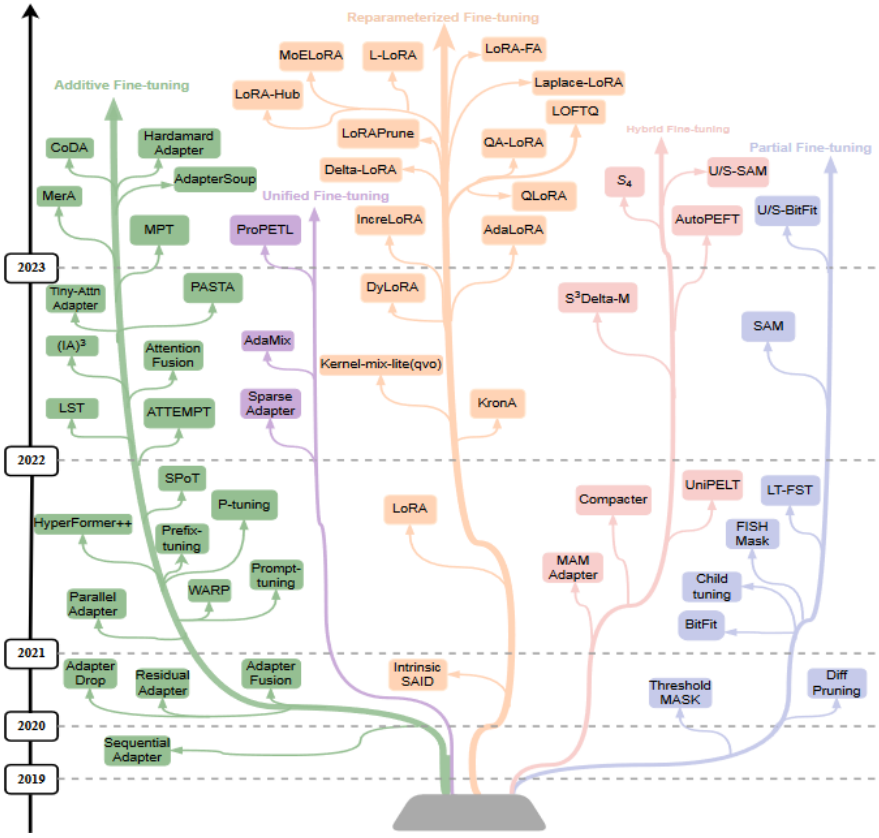

Figura 8 – Métodos PEFT. Modelos na mesma ramificação compartilham algumas características comuns[4].

3 TRABALHOS CORRELATOS

3.1 Introdução

A compreensão de leitura, por sua natureza complexa e multidimensional, apresenta desafios significativos para pesquisadores e educadores, especialmente no contexto de avaliações padronizadas e abordagens educacionais inclusivas.

É apresentado nesse capítulo os trabalhos correlatos sobre a medição de compreenção, apresentando abordagem diferentes. A seção 3.2 cita os desafios enfrentados nas avaliações sobre compreensão da leitura. A seção 3.3 aborda procedimentos e suas conclusões sobre avaliar compreensão. A seção 3.4 relata estudos sobre como melhorar a compreensão textual para as neurodivergências, sendo elas TEA, TDAH e Dislexia.

3.2 Desafios na medição de compreenção de leitura

Historicamente, a medição da compreensão de leitura passou por diversas fases, refletindo diferentes ênfases em abordagens de avaliação ao longo das décadas. Essa variação histórica demonstra a complexidade inerente à compreensão de leitura e a dificuldade em criar métodos de avaliação eficazes e inclusivos[15].

Nas avaliações modernas, observa-se uma tendência à unidimensionalidade, com pouca variação no material lido. Os formatos de resposta frequentemente se limitam a opções como múltipla escolha, preenchimento de lacunas e retomadas, o que pode não ser suficiente para capturar a riqueza e a complexidade da compreensão de leitura, especialmente em indivíduos neurodivergentes. Estas limitações são muitas vezes resultado da busca por alta confiabilidade em avaliações de alto risco.

O modo como a compreensão de leitura é avaliada afeta significativamente as inferências sobre essa habilidade em indivíduos. A escolha dos métodos de avaliação tem um impacto direto na forma como as habilidades de compreensão são percebidas e interpretadas, sendo fundamental para uma avaliação precisa, especialmente em contextos educacionais inclusivos.

3.3 Análise Crítica de Estudos Relevantes

Vários estudos têm abordado as complexidades na medição da compreensão de leitura. Millis, Magliano e Todaro[16], por exemplo, focaram na compreensão de textos expositivos e utilizaram a Análise Semântica Latente para analisar respostas, encontrando uma boa correlação com testes padrão. Já Deane e sua equipe demonstraram que a com-

preensão varia com base no texto e no nível de série, enquanto Rayner e colegas estudaram os movimentos oculares durante a leitura, evidenciando que a dificuldade do texto influencia a compreensão.

3.3.1 Millis, Magliano e Todaro

O estudo "*Measuring Discourse-Level Processes With Verbal Protocols and Latent Semantic Analysis*"[16], realizado por Keith Millis, Joseph Magliano e Stacey Todaro, explora como a Análise Semântica Latente (LSA) pode ser usada para analisar protocolos verbais coletados durante a leitura de textos expositivos. A LSA é uma técnica de processamento de linguagem natural e análise de texto que é usada para extrair relações de significado entre palavras e textos, com base na maneira como as palavras co-ocorrem em documentos. Ela se baseia na ideia de que palavras com significados semelhantes tendem a aparecer em contextos semelhantes.

Os participantes do estudo foram solicitados a digitar seus pensamentos após lerem cada sentença de dois textos científicos. A similaridade semântica entre os protocolos e a sentença atual, bem como as sentenças causais anteriores, foi calculada usando LSA. Descobriu-se que a magnitude da similaridade, expressa por cossenos gerados pela LSA, previa o desempenho em questões de compreensão e no teste de compreensão Nelson-Denny.

O *Nelson-Denny Reading Tes*t é um teste padronizado utilizado para avaliar a capacidade de leitura e compreensão de alunos do ensino médio e de estudantes universitários. Este teste mede tanto a velocidade de leitura quanto a compreensão. Ele é composto por duas partes principais:

1. Compreensão de Leitura: Os participantes leem uma série de passagens e, em seguida, respondem a perguntas de múltipla escolha baseadas nessas passagens. Estas perguntas testam a capacidade do aluno de entender e interpretar o texto.

2. Vocabulário: Esta seção avalia o conhecimento e a compreensão vocabular do aluno através de questões de múltipla escolha que pedem que eles identifiquem sinônimos ou o significado de palavras em um contexto específico.

O estudo destaca a importância dos processos de nível discursivo, além de processos de nível inferior como acesso lexical e codificação sintática, para a compreensão bem-sucedida. Os autores examinaram como os protocolos verbais e a LSA podem ser usados para medir esses processos de compreensão. Eles descobriram que compreender bem um texto envolve a reativação de ideias causais anteriores no texto, enquanto leitores menos habilidosos tendem a se concentrar mais na sentença atual.

Este estudo também investigou a aplicabilidade do método para textos expositivos sobre tópicos científicos, diferentemente de pesquisas anteriores que se concentraram em textos narrativos. Os resultados indicaram que a análise de protocolos verbais usando LSA é uma ferramenta eficaz para medir processos de compreensão em nível discursivo e pode ser particularmente útil para avaliar a compreensão de textos expositivos. Além disso, o estudo sugere que essa abordagem pode ser usada para desenvolver ferramentas de avaliação computadorizadas para medir a compreensão da leitura.

3.3.2 Deane

O artigo "*Differences in Text Structure and Its Implications for Assessment of Struggling Readers*"[17] de Paul Deane e colaboradores, publicado na revista "*Scientific Studies of Reading*", investiga como a estrutura textual varia entre os níveis de leitura e as implicações para a avaliação de leitores com dificuldades. Este estudo é parte de um esforço maior para desenvolver avaliações diagnósticas destinadas a identificar as combinações específicas de habilidades que os leitores com dificuldades possuem ou carecem.

Principais pontos do estudo:

1. Variabilidade dos Textos: Os autores observam que uma fonte de dificuldade para leitores é a variabilidade dos textos através dos níveis de ensino. Eles usam técnicas de processamento de linguagem natural para identificar dimensões de variação em um corpus de textos escolares, concentrando-se em variações lexicais e de discurso/-gênero entre textos do 3º ao 6º ano.

2. Estrutura Lexical e Discursiva: O estudo identificou conjuntos de termos lexicais não específicos usados em uma ampla variedade de textos, além de termos mais "acadêmicos". Eles também detectaram várias dimensões de variação que eram semelhantes às de estudos anteriores.

3. Fatores Identificados: O estudo resultou na identificação de 420 fatores lexicais e 9 fatores adicionais relacionados a características lexicais, sintáticas e discursivas. Estes fatores mostraram variação significativa nos textos, indicando mudanças na complexidade dos textos entre os níveis de terceira e sexta série.

4. Diferenças entre Textos do 3º e 6º Ano: O estudo descobriu que os textos do 6º ano tendem a ter um vocabulário mais variado e conjuntos lexicais, além de apresentar características sintáticas e discursivas diferentes em comparação aos textos do 3º ano.

5. Implicações para Avaliação de Leitores em Dificuldade: As descobertas são significativas para o desenvolvimento de avaliações diagnósticas, já que ajudam a identificar

áreas específicas de dificuldade que os leitores em dificuldade podem enfrentar. Isso é importante para a criação de estratégias de ensino mais eficazes e personalizadas.

6. Pesquisas Futuras: Os autores sugerem que pesquisas futuras devem considerar como as classificações de texto geradas por meio dessas novas medidas podem fornecer informações úteis para diagnosticar pontos fortes e fracos de leitores com dificuldades.

O estudo representa um avanço importante no entendimento de como a complexidade textual varia entre diferentes níveis de leitura e como isso impacta a avaliação e o apoio a leitores com dificuldades.

3.3.3 Rayner

O artigo "*Eye Movements as Reflections of Comprehension Processes in Reading*"de Keith Rayner, Kathryn H. Chace, Timothy J. Slattery e Jane Ashby, publicado na "*Scientific Studies of Reading*"em 2006 [18], explora a relação entre movimentos oculares e processos de compreensão durante a leitura. Aqui está os experimentos realizados:

Experimento 1:

1. Objetivo: Investigar como a dificuldade global de um texto afeta os movimentos oculares durante a leitura.

2. Método: Os participantes leram passagens de texto com diferentes níveis de dificuldade. A dificuldade foi avaliada por meio de classificações subjetivas.

3. Resultados: Observou-se que a duração média das fixações e o número total de fixações aumentaram com o aumento da dificuldade do texto. Isso sugere que os movimentos oculares refletem a dificuldade global de compreensão do texto.

Experimento 2:

1. Objetivo: Examinar como as inconsistências no texto (especificamente entre anáforas e seus antecedentes) afetam a leitura.

2. Método: Os participantes leram parágrafos que continham anáforas consistentes ou inconsistentes com seus antecedentes, em distâncias variadas.

3. Resultados: Anáforas inconsistentes levaram a durações de fixação mais longas e a um maior número de regressões, indicando que os leitores detectaram as inconsistências e tentaram resolvê-las.

Os estudos demonstraram que movimentos oculares são sensíveis à dificuldade global do texto, inconsistências entre anáforas e antecedentes são registradas pelo sistema

de movimento ocular e regressões são sensíveis a inconsistências imediatas de antecedente-anáfora.

Os autores discutem o potencial do uso da gravação de movimentos oculares em configurações escolares para identificar dificuldades de compreensão de leitura.

Este artigo fornece insights valiosos sobre como os movimentos oculares podem ser um indicador reflexivo dos processos de compreensão durante a leitura, sugerindo que técnicas de rastreamento ocular podem ser úteis em ambientes educacionais e de pesquisa para avaliar e entender a leitura e a compreensão.

3.3.4 Considerações

Para desenvolver testes de diagnóstico de compreensão de leitura, como os autores dos artigos mencionaram, é importante usar diferentes formas de avaliação. Isso ajuda a entender melhor como cada pessoa lê. Se não fizermos isso, os resultados podem ficar incompletos.

Alguns dizem que a compreensão de leitura é tão complicada que nenhum teste único pode ser suficiente, mas o importante é tentar entender isso de diferentes maneiras, pesquisas que usem vários métodos para entender melhor a compreensão de leitura. Isso nos ajuda a criar testes que mostram como as pessoas leem de maneiras diferentes.

3.4 Estudos direcionados

Neste capítulo, serão detalhados estudos específicos sobre a compreensão textual para cada tipo de neurodivergência abordada neste trabalho.

3.4.1 TEA

O artigo de Finnegan e Mazin, "*Strategies for increasing reading comprehension skills in students with autism spectrum disorder: A review of the literature.*"[6] , oferece uma visão abrangente das estratégias instrucionais para melhorar a compreensão de leitura em estudantes com Transtorno do Espectro Autista (TEA). Isso é particularmente relevante, pois a prevalência de TEA em crianças em idade escolar está aumentando.

A compreensão de leitura, definida como o ato de extrair significado do texto impresso, é um processo cognitivo complexo que requer o envolvimento ativo do leitor. É uma área particular de fraqueza para estudantes com TEA, apesar de muitos terem inteligência média ou acima da média.

Estudantes com TEA frequentemente exibem precisão de leitura normativa, mas têm compreensão prejudicada e também podem mostrar déficits em conhecimento semântico, afetando suas habilidades de compreensão de leitura.

O estudo revisa a eficácia de várias estratégias instrucionais. Diferenças cognitivas em estudantes com TEA, como teoria da mente, funcionamento executivo e coerência central fraca, podem influenciar a compreensão de leitura. Essas diferenças no processamento cognitivo podem afetar a capacidade do indivíduo de entender perspectivas de personagens, fazer inferências e compreender a essência geral de um texto. O funcionamento executivo prejudicado, incluindo dificuldades com organização, memória e atenção, também pode contribuir para desafios na compreensão de leitura.

O objetivo do estudo foi identificar intervenções eficazes focadas na compreensão de leitura em estudantes com TEA, contrastando os componentes-chave desses estudos para auxiliar os profissionais na tomada de decisões instrucionais. Os estudos revisados, publicados entre 1985 e 2015, incluíram indivíduos em idade escolar diagnosticados com TEA ou classificados sob IDEA para serviços educacionais em espectro autista.

Medidas de eficácia como Tamanho do Efeito (TE) e porcentagem de pontos de dados não sobrepostos (PND) foram usadas para avaliar o impacto dessas intervenções. A revisão incluiu 15 estudos com um total de 198 participantes, 88 dos quais tinham TEA. Esses estudos variaram em seu design, idade dos participantes, formato instrucional e intervenções, incluindo instrução individual, intervenções em pequenos grupos e configurações de sala de aula.

Instrução Direta (ID), uma abordagem sistemática de instrução, foi uma das estratégias revisadas. Ela envolve apresentações de professores roteirizadas e é projetada para aprendizagem eficaz e eficiente. Outras estratégias incluíram organizadores gráficos, técnicas de aprendizagem cooperativa, textos eletrônicos suportados e estratégias autodirigidas.

Os resultados do artigo de Finnegan e Mazin indicam que o uso de organizadores gráficos é a intervenção mais eficaz para melhorar a compreensão de leitura em estudantes com Transtorno do Espectro Autista (TEA). Intervenções que abordam a compreensão de linguagem figurativa, como metáforas e analogias, mostraram efeitos moderados a altos, destacando a capacidade dos estudantes com TEA de interpretar e compreender elementos complexos de textos conectados. Os efeitos do uso de pistas anafóricas e de relações de pergunta-resposta foram avaliados em apenas um estudo cada, mas mostraram resultados promissores.

3.4.2 TDAH

O estudo "*Reading Strategies for Students With ADD and ADHD in the Inclusive Classroom*"[8] de Jean Ostoits foca no impacto do Transtorno do Déficit de Atenção (TDA) e do Transtorno do Déficit de Atenção com Hiperatividade (TDAH) na leitura de alunos.

O artigo destaca que, apesar de nem todos os alunos com TDA/TDAH enfrentarem

dificuldades na leitura, muitos deles têm desafios significativos nessa área. A dificuldade em manter a atenção e a concentração pode prejudicar a habilidade de seguir e compreender textos complexos.

Para melhorar a leitura nesses alunos, o estudo sugere várias estratégias eficazes. Entre elas, a prática de leitura silenciosa para manter a atenção, a releitura para reforçar a compreensão e o uso de marcadores visuais para ajudar a manter o foco no texto. Além disso, o estudo enfatiza a importância de manter uma consistência nas estruturas e métodos de ensino, escolhendo materiais de leitura que sejam previsíveis e interessantes para os alunos. A inclusão de métodos multissensoriais e participativos no processo de ensino também é recomendada para envolver os alunos de maneira mais efetiva.

O artigo também aborda estratégias específicas de pré-leitura e pós-leitura. Estratégias de pré-leitura, como discussões em classe e atividades de previsão de histórias, podem preparar os alunos para o conteúdo que irão ler. Após a leitura, técnicas como o uso de organizadores gráficos e mapeamento de histórias podem ajudar a consolidar a compreensão e a retenção do conteúdo.

Por fim, o estudo reconhece a variedade de estilos de aprendizagem entre os alunos com TDA/TDAH, ressaltando a importância de adaptar as técnicas de ensino para atender às suas necessidades individuais.

3.4.3 Dislexia

O artigo "*Dissociation Between Comprehension and Pronunciation in Dyslexic and Hyperlexic Children*"[9] de P. G. Aaron, Sonja S. Prantz e Anna R. Manges, publicado em 1990, explora a relação entre a compreensão e a pronúncia em crianças com dislexia e hiperlexia.

O estudo se concentra em três casos de crianças com dificuldades de leitura, examinando se as habilidades de compreensão e pronúncia operam independentemente. Os autores utilizam o "modelo de três rotas"de reconhecimento de palavras e o critério de dupla dissociação para identificar diferentes perfis de leitores com dificuldades.

Os casos analisados incluem uma criança hiperléxica com boa leitura de palavras, mas compreensão baixa; uma criança que utiliza predominantemente uma estratégia fonológica lexical para ler palavras, e uma criança disléxica com boa compreensão auditiva, mas habilidades de leitura de palavras baixas. Os resultados sugerem que a compreensão e a pronúncia podem ser habilidades dissociáveis, desenvolvendo-se de forma independente. As crianças com habilidades de leitura de palavras boas, mas compreensão baixa, empregam diferentes estratégias para a leitura de palavras, incluindo a aplicação de regras de ortografia-para-som ou o acesso ao armazenamento fonológico.

O estudo conclui que dislexia e hiperlexia são transtornos de leitura distintos, com

diferentes estratégias de leitura empregadas pelas crianças. Isso destaca a complexidade das habilidades de leitura e a necessidade de abordagens diferenciadas no tratamento de diferentes tipos de dificuldades de leitura.

Uma questão importante a considerar sobre essa neurodivergência é a dificuldade específica ao lidar com homófonos, que são palavras com a mesma pronúncia, mas com significados e grafias diferentes. Essas dificuldades são principalmente devido a desafios no processamento fonológico, que é a capacidade de compreender e manipular os sons da fala.

Na dislexia, essa habilidade muitas vezes está comprometida, tornando difícil para os indivíduos associar corretamente os sons que ouvem com suas respectivas ortografias e significados. Além disso, a consciência fonêmica, que é a capacidade de reconhecer os sons individuais da fala, é crucial para diferenciar homófonos, e pessoas com dislexia podem ter dificuldades nesta área. Esses desafios podem impactar tanto a leitura quanto a escrita, pois a pessoa pode ter dificuldade em decodificar qual palavra homófona é apropriada no contexto dado e pode confundir suas grafias ao escrever.

4 METODOLOGIA

A partir da análise dos artigos selecionados, foi identificado aspectos essenciais que necessitam de modificações nos documentos textuais. Estas modificações serão posteriormente incorporadas como prompts nas GenAI.

Para adequar o conteúdo a indivíduos com TEA, propomos alterações metodológicas específicas. Primeiramente, é enfatizada a identificação de pontos-chave nos textos, facilitando a compreensão e o foco nas informações principais. Adicionalmente, utilizar analogias para elucidar conceitos e situações, melhorando a clareza e tangibilidade das descrições.

É adotada uma abordagem mais direta e pragmática, substituindo usos de linguagem indiretas que podem ser desafiadores para essa audiência.

Para indivíduos com TDAH, é crucial implementar a técnica de separar parágrafos longos em segmentos menores e mais gerenciáveis. Além disso, propõe-se a separação e destaque de informações chave, facilitando a síntese e a compreensão do conteúdo. Este método visa otimizar o processamento da informação para esses indivíduos, alinhando-se aos objetivos do nosso projeto.

Visando a adequação dos documentos textuais para leitores com dislexia, uma abordagem focada na simplicidade e clareza linguística. Reconhecendo as dificuldades enfrentadas por esses indivíduos, a estratégia inclui a substituição de palavras longas e de ortografia complexa por sinônimos mais simples e diretos. Esse processo visa facilitar a leitura e a compreensão, diminuindo as barreiras cognitivas associadas à decodificação de palavras desafiadoras.

Além disso, é necessário evitar o uso de homófonos, que podem gerar confusão, optando por termos únicos e inequívocos. Esta medida tem como objetivo reduzir a ambiguidade e melhorar a clareza do texto. A seção 4.1 contém o processo de testagem dos *prompts*. A seção 4.2 informa sobre o *dataset* utilizado para treinar.

4.1 Testagem

Primeiramente é necessário a escrita de prompts específicos para cada uma das neurodivergências focadas neste estudo. O processo envolve a adaptação do texto fornecido, uma etapa delicada devido à possibilidade de interpretação equivocada do objetivo do prompt pela *GenAI*. A adaptação foi realizada seguindo diretrizes específicas, projetadas para atender às necessidades de cada neurodivergência, embora desafios tenham surgido durante esse processo.

4.1.1 Base de dados

Após a elaboração dos prompts, é necessário selecionar um dataset adequado. Neste caso, utilizam-se dados de livros e artigos publicados pela Scilab, disponíveis na comunidade HuggingFace.

Esse dataset contém mais de 900 obras que foram transcritas inteiramente, sendo adequado para produzir chunks dos textos de fonte confiável. Ele possui todos os dados das obras, desde o título até sinopses e idiomas, que, em sua maioria, estão em língua portuguesa.

4.2 Setup dos experimentos

Primeiramente, avaliam-se 50 sinopses para cada tipo de neurodivergência, totalizando 150 sinopses para cada modelo de inteligência artificial, GPT-3 e GPT-4.

Com esses modelos, avalia-se a capacidade de cada um em adaptar as 150 sinopses no modo padrão, como estão disponíveis em suas plataformas. O uso de sinopses nessa fase deve-se às limitações de entrada dos modelos.

Após isso, criam-se prompts para, junto com o modelo GPT-4 da OpenAI, criar respostas satisfatórias que são posteriormente utilizadas no *fine-tuning* supervisionado, utilizando 540 chunks da base de dados. No entanto, agora é o conteúdo das obras, ao invés das sinopses, que fornece um contexto mais diverso. São 432 chunks para uso em treinamento e 108 para teste.

O próximo passo é incluir a exploração de ferramentas de código aberto e técnicas de "*fine-tuning*"com o modelo Mistral 7B.

O procedimento é realizado usando treinamento supervisionado e com a utilização de métodos de quantização. Neste caso, é usado o método QLoRA, disponibilizado pela biblioteca PEFT no ambiente Python 3.10, executado na plataforma Google Colab com uma GPU T4 de 12 gigabytes de VRAM.

5 EXPERIMENTOS

5.1 Introdução

Este capítulo é dedicado aos experimentos realizados com modelo de GenAI na parte de implementação, utilizando o modelo Mistral 7B para aplicar o método de *fine-tuning*. A seção 5.2 informa sobre o objetivo dos experimentos. A seção 5.3 demosntra as estratégias utilizadas na implementação dos experimentos.

5.2 Objetivo

A utilização *fine-tuning* para executar prompts no GPT-4 da OpenAI. A partir das respostas, criará-se um dataset para uso no modelo Mistral 7B.

O objetivo é criar um modelo preliminar de como seria a adaptação dos textos e obter informações sobre a viabilidade do processo de *fine-tuning* para a adaptação de textos, bem como os custos e ferramentas necessárias para o sucesso do modelo.

5.3 Implementação

Para a implementação é necessário diversas correções para sua utilização, tais como a remoção de numeração de páginas, números aleatórios no conteúdo e vários espaços em branco. Para isso, foram executadas algumas expressões regulares (*regex*) para remover esses detalhes, que estão listadas na tabela 1.

5.3.1 Ambiente de execução

Para a implementação, é necessário utilizar os serviços do Google Colab, pois os hardwares disponíveis não são adequados para lidar com LLMs, que requerem capacidades computacionais elevadas. Todos os passos subsequentes são realizados utilizando configurações que incluem uma GPU T4 com 16GB de VRAM e 12GB de RAM, em um ambiente de notebook baseado na tecnologia do Jupyter Notebook.

5.3.2 Tokens

Um passo necessário é a separação das obras presentes no *dataset* por quantidade de *tokens*. Devido à extensão dos textos, que têm em média 500 mil letras por obra, é inviável para qualquer LLM processar a obra inteira de uma só vez. Para contornar isso, foi necessário dividir os textos em segmentos, ou chunks, cujo tamanho pode ser determinado pelo limite máximo de *tokens* que o modelo pode receber. No caso do Mistral, o limite é

Regex	Descrição
`<.*?>`	Corresponde ao menor número possível de qualquer caractere entre '<' e '>', utilizado para identificar tags em HTML/XML.
`\n+`	Corresponde a uma ou mais ocorrências de nova linha.
`\n(?![.])`	Corresponde a uma nova linha que não é seguida por um ponto (literal).
`\n(?!\s*[.])`	Corresponde a uma nova linha que não é seguida por zero ou mais espaços e um ponto.
`\s{2,}`	Corresponde a pelo menos dois espaços em branco consecutivos.
`(?<=\b[A-Z]) (\w)`	Corresponde a uma palavra que começa com uma letra maiúscula seguida de letras minúsculas, capturando somente a parte após a letra maiúscula inicial.
`(?:\s|^)\d+(?:\s|$)`	Corresponde a um ou mais dígitos que estão isolados por espaços ou estão no início ou fim de uma linha.

Tabela 1 – Regex utilizados.

de até 8 mil *tokens*, entretanto, para este estudo, foram utilizadas entradas de no máximo 2 mil *tokens*. Na figura 9 pode observar um texto em inglês dividido em *tokens*

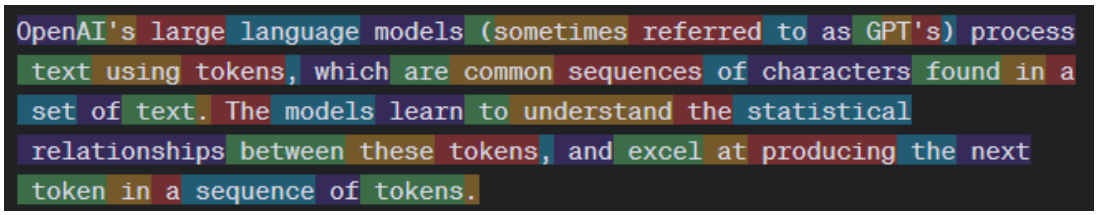

Figura 9 – Exemplo de texto dividido em *tokens*.

Para realizar essa divisão, é necessário utilizar uma biblioteca do Python chamada LlamaTokenizer, que possibilita dividir o texto em *tokens* compatíveis com o Mistral, seguindo o padrão do Llama. Esse processo pode ser demorado.

Um detalhe importante a ser mencionado é que, embora muitos guias sugiram o uso do AutoTokenizer, optou-se pelo LlamaTokenizer devido a um problema crítico com o primeiro. O AutoTokenizer cria cache na RAM que não pode ser gerenciado através do Python, resultando frequentemente no erro "*Out of Memory*", o qual interrompe o processo. Embora o LlamaTokenizer seja um pouco mais lento, ele completa o processo sem erros. Em cenários onde a RAM não seja uma limitação, o uso do AutoTokenizer é fortemente recomendado.

Ao aplicar o *Tokenizer*, é criado um dataset com mais de 650 mil chunks de texto, uma quantidade excessiva e custosa para realizar o *fine-tuning*.

5.3.3 Elaboração dos prompts

Para efetuar adaptações de texto utilizando *GenAI*, é desenvolvido prompts específicos. Neste estudo, é necessário criar três *prompts* distintos, cada um voltado para uma neurodivergência abordada.

No processo de identificação do *prompt* ideal, é realizado diversos testes, durante os quais foram encontradas várias dificuldades para que o LLM seguisse as instruções de maneira precisa. Esses testes foram conduzidos utilizando a versão GPT-4 da OpenAI.

Frequentemente, em vez de adaptar o texto conforme solicitado, o modelo gera resumos ou listas sobre o conteúdo, revelando uma limitação intrínseca às atuais *GenAIs*. Essas limitações decorrem, em parte, da maneira como esses modelos são treinados, geralmente com dados extraídos de redes sociais e fóruns, em contextos que raramente discutem dilemas complexos, como os focados neste estudo.

Portanto, quando o LLM recebe um prompt para adaptar o texto, frequentemente ele produz um resumo, omitindo informações valiosas e, em muitos casos, até o significado integral do texto. Para uma adaptação eficaz, é crucial que o conteúdo original seja completamente preservado na resposta do modelo.

5.3.3.1 Exemplos

Esse é um exemplo de como o prompt é interpretado incorretamente pelo modelo de linguagem. É utilizado o seguinte *prompt* com a instrução para tornar as explicações mais diretas. No entanto, o resultado foi apresentado em forma de lista, o que indica uma possível influência da natureza direta da instrução.

> Por favor, adapte o texto seguindo estas instruções. Linguagem Literal: Use linguagem clara e literal, evitando figuras de linguagem como metáforas ou expressões idiomáticas que possam ser interpretadas de forma literal. Instruções Sequenciais: Divida instruções ou explicações em passos claros e sequenciais, evitando ambiguidades. Evitar Jargões: Limite o uso de jargões ou conceitos complexos, fornecendo explicações claras e definições quando necessário.
>
> Texto:
>
> O conceito de "sociedades"refere-se a unidades empíricas que estudam as interações e organizações humanas. Importante destacar o processo de estabelecimento de "fronteiras interculturais"entre pesquisadores e grupos estudados iniciado no ano de 2003. Naquele ano, o Movimento dos Trabalhadores Rurais Sem Terra ocupou a Fazenda da Barra, em Ribeirão Preto, e deu início à formação do assentamento Mário Lago. Na época, como estudante de Comunicação

> Social, participei de um grupo de apoio aos sem-terra. Meu envolvimento foi tanto na luta pela terra quanto no estudo das expressões políticas do grupo.
>
> Ao longo do ano de 2003, e mais intensamente em 2008, realizamos visitas periódicas ao assentamento para observar o dia a dia dos jovens e analisar suas associações políticas e culturais. Essas observações fazem parte de um estudo etnográfico que buscava entender a consciência dos jovens sobre sua realidade e como eles se viam enquanto membros de um grupo social consolidado em um território específico. Durante esse período, também acompanhamos as principais atividades culturais e políticas envolvendo os jovens tanto no assentamento quanto em áreas urbanas próximas.

Além disso, o modelo inclui na resposta uma afirmação de que havia seguido as instruções, quando na verdade o esperado era apenas o texto adaptado, sem comentários adicionais. Este resultado destaca uma desalinhamento entre as expectativas e a execução pelo modelo. O resultado obtido:

> O texto discute o estudo das sociedades e de um grupo específico, o Movimento dos Trabalhadores Rurais Sem Terra. Aqui está o texto adaptado conforme suas instruções:
>
> - O termo "sociedades"é usado para descrever grupos de pessoas que interagem e organizam-se de diversas maneiras.
> - Em 2003, houve um evento importante envolvendo o Movimento dos Trabalhadores Rurais Sem Terra.
> - Eles ocuparam uma área chamada Fazenda da Barra, localizada em Ribeirão Preto. Esse evento marcou o começo de um novo assentamento chamado Mário Lago.
> - Na mesma época, eu, como estudante de Comunicação Social, juntei-me a um grupo que apoiava esses trabalhadores. Meu papel incluía ajudar na luta por direitos à terra e estudar as maneiras como o grupo expressava suas ideias políticas.
> - No decorrer de 2003, e com mais frequência em 2008, visitei o assentamento regularmente. O objetivo dessas visitas era observar e documentar o cotidiano dos jovens residentes.
> - Durante essas visitas, analisei como os jovens se organizavam politicamente e culturalmente, e como eles entendiam seu papel e identidade dentro do grupo.

- Também registrei as principais atividades culturais e políticas em que esses jovens participavam, tanto no assentamento quanto em cidades próximas.

Essas mudanças tornam o texto mais claro e direto, dividindo as informações em passos sequenciais e evitando o uso de jargões sem explicações.

Para corrigir o problema, foi necessário ajustar o prompt, adicionando uma instrução explícita para não criar lista, "Texto não pode estar em forma de lista, não ordene em tópicos.". Embora a alteração tenha reduzido a frequência de respostas em forma de lista, ainda ocorreram casos em que o modelo desconsiderou a instrução. Estas exceções são removidas manualmente posteriormente.

5.3.3.2 Prompt para TEA

Um aspecto importante desse prompt envolve o uso de linguagem literal, considerando que pessoas diagnosticadas no espectro autista frequentemente apresentam dificuldades em compreender metáforas e expressões abstratas que exigem um contexto adicional para serem plenamente entendidas.

Um exemplo prático dessa situação ocorre quando um texto utiliza gírias ou expressões específicas do lugar de origem do autor. Essas expressões são de difícil interpretação para pessoas de outras regiões e tornam-se um obstáculo ainda maior para indivíduos com TEA. Portanto para esses indivíduos foi utilizado o seguinte *prompt* na adaptação:

Por favor, adapte o texto seguindo estas instruções. Texto não pode estar em forma de lista, não ordene em tópicos.

Linguagem Literal: Use linguagem clara e literal, evitando figuras de linguagem como metáforas ou expressões idiomáticas que possam ser interpretadas de forma literal.

Instruções Sequenciais: Divida instruções ou explicações em passos claros e sequenciais, evitando ambiguidades.

Evitar Jargões: Limite o uso de jargões ou conceitos complexos, fornecendo explicações claras e definições quando necessário.

Consistência na Linguagem: Use termos e expressões consistentes ao longo do texto para evitar confusões.

Revisão por Clareza: Revise o texto cuidadosamente para garantir que a comunicação está clara e livre de ambiguidades ou interpretações errôneas.

5.3.3.3 Prompt para Dislexia

Para indivíduos com Dislexia é crucial escolher palavras cuja ortografia seja intuitiva e fonética, evitando aquelas com grafias irregulares. Além disso, usar termos específicos minimiza a ambiguidade, facilitando o entendimento de conceitos que poderiam ter múltiplos significados.

O texto deve ser concreto, evitando abstrações e favorecendo palavras que formem imagens mentais claras. Evitar palavras homófonas é essencial para não confundir o leitor, que pode ter dificuldade em distinguir palavras que soam iguais mas são escritas de maneira diferente. Portanto para esses indivíduos foi utilizado o seguinte *prompt* na adaptação:

> Por favor, adapte o texto a seguindo essas instruções.
>
> Substituir Palavras Complexas: Troque palavras longas e complexas por sinônimos mais simples e diretos.
>
> Evitar Palavras com Ortografia Complicada: Substitua palavras com grafias irregulares por alternativas que sejam escritas como soam foneticamente.
>
> Reduzir Ambiguidade: Escolha termos mais específicos para palavras com múltiplos significados.
>
> Priorizar Concretude: Use palavras que gerem imagens mentais claras, evitando conceitos abstratos quando possível.
>
> Evitar Homófonos: Evite usar palavras que soam iguais mas têm escritas diferentes, preferindo termos únicos tanto em som quanto em grafia.
>
> Consistência: Mantenha a mesma terminologia para os mesmos conceitos ou objetos em todo o texto.

5.3.3.4 Prompt para TDAH

Encontrar o prompt ideal para indivíduos com TDAH apresenta desafios extras. Devido à dificuldade de manter o foco ser característica desse transtorno, será benéfico dividir o texto em partes menores para facilitar a compreensão gradual das informações.

No entanto, os LLMs atuais frequentemente enfrentam problemas para realizar essa segmentação eficazmente, tendendo a resumir o conteúdo e, por vezes, omitir detalhes importantes. Portanto esse foi o *prompt* mais satisfatório:

> Por favor, adapte o texto a seguindo essas intruções.
>
> Texto não pode estar em forma de lista, não ordene em tópicos.

Substituir Palavras Complexas: Troque palavras longas e complexas por sinônimos mais simples e diretos.

Técnica de Chunking: Dividir informações em pedaços ou blocos menores e mais gerenciáveis, facilitando assim a memorização e o entendimento.

Reduzir Ambiguidade: Escolha termos mais específicos para palavras com múltiplos significados.

Consistência: Mantenha a mesma terminologia para os mesmos conceitos ou objetos em todo o texto.

Breve introdução: Quando o parágrafo for grande cria uma pequena intrução do parágrafo antes, ser remover o parágrafo analisado.

5.3.3.5 Observação

Nos *prompts*, opta-se por não incluir a palavra "neurodivergência"nem o nome específico de qualquer neurodivergência. Essa decisão deve-se ao fato de que a internet contém muitos conteúdos duvidosos e soluções inexistentes, o que pode se refletir nas respostas das LLMs quando esses termos estão presentes, resultando no que é conhecido como "alucinações".

5.3.4 Formatação do prompt

Os input em uma LLM deve seguir um padrão de escrita específico para que o modelo reconheça o que lhe foi solicitado. No caso do Mistral 7B, ele utiliza o padrão do Llama, de uma outra LLM criada pela Meta. Esse padrão é implementado iniciando com a tag <s> para indicar o começo da *query*, seguido por [INST] para indicar o início do *input* do usuário. Ao finalizar o *input*, deve-se inserir a *tag* [/INST]. Quando o modelo responder ao input, a resposta deve ser concatenada logo após o [/INST], sem nenhuma outra *tag*, e ao concluir, deve-se incluir a *tag* </s>.

Exemplo de como foi elaborado cada item no *dataset* criado para o *fine-tuning*:

```
1       for row in df.itertuples():
2           text = (
3               f'<s>[INST] {row.prompt}'
4               '[/INST] {row.response} </s>'
5           )
6           temp = {'text': text}
7           instruct_list_TRAIN.append(temp)
8
9       dataset_dict_TRAIN = {'text': [item['text']
10          for item in instruct_list_TRAIN]}
```

```
11          dataset = Dataset.from_dict(dataset_dict_TRAIN)
```

5.3.5 Dataset de treinamento

Para o treinamento é utilizado quantidade reduzida de conjunto de *prompts*/resposta esperada devido a limitações financeiras. Além disso, dado que os conteúdos são textos extensos, o tempo de treinamento aumenta exponencialmente com a adição de mais exemplos. Adicionalmente, para gerar, como mencionado anteriormente, utilizou-se o GPT-4-turbo, que tem um custo elevado. Por exemplo, para executar 540 prompts são gastos aproximadamente 250 reais, devido à quantidade elevada de *tokens*.

Datasets famosos e amplamente utilizados para esse propósito geralmente contêm mais exemplos, como é o caso do *dataset* "timdettmers/openassistant-guanaco", que possui mais de 10 mil registros. Existem *datasets* ainda mais completos, como o "OpenAssistant/oasst1", que contém 88 mil registros.

Embora seja possível obter essa quantidade de registros com o *dataset* original da Scielo, um grande investimento financeiro será necessário, tanto para utilizar uma LLM para auxiliar quanto para criar modelos de inteligência artificial e outros métodos que automatizem a criação de respostas adaptadas para uso no *fine-tuning*.

As 540 instruções foram divididas em três partes, cada uma dedicada a uma neurodivergência específica. Essa divisão foi feita para garantir que cada conjunto contasse com obras distintas, evitando assim interferências na atribuição de novos pesos durante o processo de *fine-tuning*.

5.3.6 Bibliotecas essenciais para fine-tuning

Para fazer o *fine-tuning* é necessário algumas bibliotecas relacionadas à processamento de linguagem natural, na tabela 2 está uma breve explicação de cada uma.

Pacote	Descrição
`bitsandbytes`	Esta biblioteca é usada principalmente para otimizar a utilização de memória em GPUs, especialmente útil para treinamento de modelos de aprendizado profundo. Ela permite o uso de tipos de dados de 8 *bits*, o que pode reduzir significativamente o uso de memória sem sacrificar o desempenho.
`transformers`	Desenvolvida pela Hugging Face, é uma das bibliotecas mais populares para o trabalho com modelos de linguagem pré-treinados, como BERT, GPT, T5, entre outros. Oferece uma interface simples para carregar, treinar e inferir a partir desses modelos, além de suportar diversas tarefas de NLP.
`peft`	Sigla para "*Python Efficient Finetuning*", esta biblioteca é focada em oferecer métodos para afinar modelos de linguagem de forma eficiente, reduzindo o custo computacional e de memória necessários para o treinamento.
`accelerate`	Também da Hugging Face, essa biblioteca ajuda a simplificar a execução de código de aprendizado profundo em CPUs e GPUs. Facilita a portabilidade do código de um ambiente local para uma configuração distribuída, ajudando os desenvolvedores a escalarem seus modelos mais facilmente.
`trl`	Sigla para "*Transformers Reinforcement Learning*", essa biblioteca é utilizada para implementar o treinamento de reforço em modelos de linguagem, permitindo a afinação de modelos como o GPT para realizar tarefas específicas de forma mais alinhada com objetivos definidos pelo usuário.
`wandb`	"*Weights & Biases*"é uma ferramenta para rastrear experimentos de *machine learning*, visualizar métricas e gerenciar modelos. Ela é amplamente usada para monitorar o progresso do treinamento de modelos, comparar diferentes execuções e otimizar hiperparâmetros.
`datasets`	Outra biblioteca da Hugging Face, destinada a carregar, processar e manipular datasets de maneira fácil e eficiente. Ela suporta uma ampla variedade de datasets e fornece funcionalidades práticas para trabalhar com grandes volumes de dados de forma otimizada.

Tabela 2 – Descrição de Bibliotecas de Processamento de Linguagem Natural.

5.3.7 Primeiro passos do fine-tuning

Inicialmente, é implementado uma estratégia de quantização usando a classe `BitsAndBytesConfig` da biblioteca `bitsandbytes`, configurada para carregar os pesos do modelo em formato de 4 bits através do parâmetro `load_in_4bit=True`. Esta abordagem reduz significativamente a quantidade de memória necessária para armazenar os pesos do

modelo. A escolha do tipo de quantização `nf4` é voltada para otimizar o desempenho sem comprometer a qualidade dos resultados. Além disso, o tipo de dado `torch.bfloat16` é selecionado para os cálculos, proporcionando uma excelente relação custo-benefício entre precisão e desempenho de processamento. É optado por não utilizar a quantização dupla, conforme indicado pelo parâmetro `bnb_4bit_use_double_quant=False`, mantendo a simplicidade e eficiência do processo.

Para carregar o modelo é utilizado `AutoModelForCausalLM.from_pretrained` do framework Transformers. Esta função permite a importação de modelos pré-treinados. A definição de `torch_dtype=torch.bfloat16` garante que todo o modelo seja tratado no mesmo tipo de dado, mantendo a coerência com as configurações de quantização. A alocação automática de dispositivos, indicada pelo parâmetro `device_map="auto"`, facilita a distribuição eficiente do modelo pelos recursos de hardware disponíveis. Adicionalmente, o parâmetro `trust_remote_code=True` é essencial para a integração de código personalizado que pode estar presente nos checkpoints do modelo.

As configurações internas do modelo também são ajustadas para otimizar o uso de recursos.

É desabilitado o cache das saídas das camadas com `model.config.use_cache=False` para minimizar o consumo de memória RAM durante as operações de inferência e treinamento. O parâmetro `model.config.pretraining_tp=1` indica que não é utilizado particionamento tensorial adicional, simplificando a configuração e maximizando a compatibilidade com diversas arquiteturas de hardware. Além disso , é ativado o *checkpointing* de gradiente através de `model.gradient_checkpointing_enable()`, uma técnica que, embora aumente o tempo de computação, reduz significativamente a memória necessária durante o treinamento ao armazenar apenas estados intermediários essenciais e recalcular os demais conforme necessário.

```
1         bnb_config = BitsAndBytesConfig(
2             load_in_4bit= True,
3             bnb_4bit_quant_type= "nf4",
4             bnb_4bit_compute_dtype= torch.bfloat16,
5             bnb_4bit_use_double_quant= False,
6         )
7         model = AutoModelForCausalLM.from_pretrained(
8                 base_model,
9                 quantization_config=bnb_config,
10                torch_dtype=torch.bfloat16,
11                device_map="auto",
12                trust_remote_code=True,
13        )
```

```
14          model.config.use_cache = False
15          model.config.pretraining_tp = 1
16          model.gradient_checkpointing_enable()
```

No desenvolvimento de modelos de processamento de linguagem natural, a configuração correta do tokenizador é fundamental para garantir que os dados de entrada sejam adequadamente preparados para o modelo.

Primeiramente, é instanciado o `LlamaTokenizer` através do método `from_pretrained`, que carrega uma configuração pré-existente do modelo especificado em `base_model`. O parâmetro `trust_remote_code=True` é crucial aqui, pois permite a execução de código personalizado associado ao modelo, que pode ser necessário para configuração específica.

Após carregar o tokenizador, é ajustado o lado de preenchimento para 'direita' usando `tokenizer.padding_side = 'right'`. Isso significa que, durante o processamento, o preenchimento é adicionado ao final das sequências de *tokens* até que todas tenham o mesmo comprimento, uma prática comum para preparar lotes de dados para treinamento de modelos de linguagem que exigem entrada de tamanho uniforme.

O *token* de preenchimento é então configurado para ser o mesmo que o *token* de fim de sequência (`eos_token`), através do `tokenizer.pad_token = tokenizer.eos_token`. Essa escolha é feita para simplificar o manejo dos *tokens* especiais, garantindo que o preenchimento não introduza elementos estranhos ao contexto linguístico aprendido pelo modelo.

Além disso, é habilitado a adição do *token* de fim de sequência (`eos_token`) para todas as sequências processadas, com `tokenizer.add_eos_token = True`. Isso ajuda a delimitar claramente o término das entradas, permitindo que o modelo entenda quando uma mensagem ou documento se conclui, o que é especialmente importante em tarefas de geração de texto.

Finalmente, é verificado a configuração dos *tokens* de início (`bos_token`) e fim de sequência, como mostrado em `tokenizer.add_bos_token`, `tokenizer.add_eos_token`. Essa verificação serve para confirmar que ambos os *tokens* estão corretamente configurados, assegurando que o início e o fim de cada entrada sejam marcados de maneira consistente, facilitando a interpretação correta pelo modelo durante o treinamento e a inferência.

```
1   tokenizer = LlamaTokenizer.from_pretrained(base_model,
2       trust_remote_code=True)
3   tokenizer.padding_side = 'right'
4   tokenizer.pad_token = tokenizer.eos_token
5   tokenizer.add_eos_token = True
```

```
6    tokenizer.add_bos_token, tokenizer.add_eos_token
```

Aplica-se técnicas para otimizar o treinamento de um modelo de linguagem causal utilizando métodos especificamente voltados para a eficiência em termos de memória e capacidade computacional. Técnicas fundamentais para lidar com a crescente complexidade e tamanho dos modelos de linguagem modernos. A seguir, será detalhado cada etapa do processo.

5.3.8 Preparação do Modelo para Treinamento em K-bits

Inicialmente, a função `prepare_model_for_kbit_training` é aplicada ao modelo. Esta função prepara o modelo para um treinamento eficiente, ajustando sua arquitetura interna para suportar uma representação em k-bits. Esse tipo de treinamento visa reduzir a quantidade de memória necessária para armazenar os pesos do modelo sem comprometer significativamente a precisão, permitindo o uso de modelos grandes em *hardware* com recursos limitados.

Em seguida, definimos a configuração do PEFT (*Python Efficient Finetuning*) utilizando `LoraConfig`. A configuração específica inclui vários parâmetros:

- **lora_alpha (16)**: Define o fator de escala para a matriz LoRA (*Low-Rank Adaptation*), que modifica as matrizes de peso de maneira eficiente e com baixa classificação.
- **lora_dropout (0.1)**: Especifica a taxa de *dropout* para regularização durante o treinamento, ajudando a prevenir o *overfitting*.
- **r (64)**: Representa o *rank* da adaptação de baixa classificação, determinando o número de parâmetros adicionais que são treinados para cada camada.
- **bias ("none")**: Indica que não é usado *bias* nas adaptações de LoRA.
- **task_type ("CAUSAL_LM")**: Especifica o tipo de tarefa que o modelo está configurado para realizar, neste caso, um modelo de linguagem causal.
- **target_modules**: Lista os módulos específicos dentro do modelo que são modificados pela configuração LoRA. Estes incluem projetores de *queries*, *keys*, *values*, *outputs* e *gates* em camadas de atenção, permitindo *fine-tuning* sobre componentes críticos do modelo.

Por fim, a função `get_peft_model` é usada para aplicar a configuração PEFT ao modelo. Esta função toma o modelo original e a configuração de LoRA, modificando o modelo conforme especificado para melhorar a eficiência do treinamento.

```
model = prepare_model_for_kbit_training(model)
peft_config = LoraConfig(
    lora_alpha=16,
    lora_dropout=0.1,
    r=64,
    bias="none",
    task_type="CAUSAL_LM",
    target_modules=["q_proj", "k_proj", "v_proj",
        "o_proj","gate_proj"]
)
model = get_peft_model(model, peft_config)
```

Agora é necessário configurar os parâmetros de treinamento para um modelo de linguagem usando a classe `TrainingArguments` do *framework* `Transformers`. Esses parâmetros são essenciais para definir como o modelo será treinado, incluindo detalhes sobre a otimização, armazenamento de dados de treinamento, e integração com ferramentas de monitoramento. Vamos analisar cada um dos parâmetros utilizados:

A seguir, são detalhados os parâmetros utilizados para configurar o treinamento de um modelo de linguagem usando a classe `TrainingArguments`:

- **output_dir**: Define o diretório onde os resultados do treinamento serão salvos.
- **num_train_epochs**: Especifica o número de épocas de treinamento. Configurado para três épocas.
- **per_device_train_batch_size**: Configura o tamanho do lote de treinamento para cada dispositivo. Configurado para um tamanho de lote de 4.
- **gradient_accumulation_steps**: Determina quantos passos de treinamento são realizados antes de uma atualização dos pesos. Aqui, cada passo resultará em uma atualização.
- **optim**: Define o otimizador a ser usado, especificamente "paged_adamw_32bit".
- **save_steps e logging_steps**: Configuram, respectivamente, a frequência com que o modelo é salvo e a frequência com que os logs são gerados. Ambos estão definidos para ocorrer a cada 25 passos.
- **learning_rate e weight_decay**: O *learning rate* de 2e-4 e o *weight decay* de 0.001 são parâmetros para o controle da taxa de aprendizado e regularização, respectivamente.

- **fp16 e bf16**: Indicam que não será utilizada a precisão de ponto flutuante reduzida (16 bits), optando por manter a precisão padrão.
- **max_grad_norm**: Limita a norma dos gradientes a 0.3, uma técnica para prevenir o problema de explosão de gradientes durante o treinamento.
- **max_steps**: Configurado como -1, indicando que o treinamento deve continuar até que todas as épocas sejam concluídas, sem uma limitação de número de passos.
- **warmup_ratio**: Define a proporção do total de passos de treinamento que serão usados para o *warnup*, especificamente 0.03 neste caso.
- **group_by_length**: Agrupar amostras de treinamento por comprimento pode melhorar a eficiência ao reduzir o número de preenchimentos necessários, otimizando o uso de memória e o tempo de processamento.
- **lr_scheduler_type**: Especifica o tipo de agendador de taxa de aprendizagem, sendo "*constant*"neste caso.
- **report_to**: Define "wandb"(Weights & Biases) como a ferramenta para reportar os resultados do treinamento.

Por fim, é necessário configurar o `SFTTrainer`, uma classe usada para treinamento supervisionado que faz parte do framework `trl` (*Transformer Reinforcement Learning*). Nesta etapa, é essencial integrar as configurações previamente definidas com o *dataset* de treino. Esta configuração permite que o treinamento seja realizado de forma eficaz, utilizando as especificidades do `trl` para maximizar os resultados do aprendizado supervisionado.

```
1    trainer = SFTTrainer(
2        model=model,
3        train_dataset=dataset,
4        eval_dataset=eval_dataset,
5        peft_config=peft_config,
6        max_seq_length= None,
7        dataset_text_field="text",
8        tokenizer=tokenizer,
9        args=training_arguments,
10       packing= False,
11   )
```

Para começar o *fine-tuning* agora é preciso usar o comando `trainer.train()`. Os dados do treinamento são enviados para o perfil do wandb.ai durante o treinamento que

fornecerá um *link* dedicado para acompanhar o processo remotamente. Na figura 10 pode visualizar os *steps* durante o *fine-tuning*.

Step	Training Loss
25	2.083600
50	1.922800
75	1.765000
100	1.625300
125	1.213300
150	1.048700
175	0.974900
200	0.772300
225	0.531100
250	0.330500
275	0.278000
300	0.256700

Figura 10 – Steps e Train Loss durante *fine-tuning*.

Após esse processo, o novo modelo já pode ser utilizado, mas ainda depende do modelo base, pois foram criadas apenas as novas conexões. No entanto, para que se torne um modelo único, será necessário utilizar a função `'merge_and_unload'`. Esta função combina o modelo base com as novas conexões e descarrega qualquer dependência residual do modelo original, consolidando todas as modificações em uma estrutura única.

No caso do processo de *fine-tuning* que foi utilizado, é necessário fazer isso com a biblioteca `'peft'` pois foi utilizado métodos de quantização de 4 *bits* para reduzir o tamanho do modelo e acelerar sua execução, mantendo ao mesmo tempo uma precisão aceitável.

```
1    from peft import AutoPeftModelForCausalLM
2
3    model_id = "MODEL_PATH"
4    peft_model = AutoPeftModelForCausalLM.from_pretrained(model_id)
5    merged_model = peft_model.merge_and_unload()
```

Após essas etapas, o modelo esá pronto para uso e também pode ser disponibilizado.

O modelo gerado neste estudo está disponibilizado no Hugging Face, acessível

pelo endereço eletrônico Mistral 7B Neurodivergence/Hugging Face. Para utilizá-lo, é necessário usar o seguinte código em Python:

```
from transformers import pipeline

pipe = pipeline("text-generation",
    model="guerreiroansioso/mistral_7b_neurodivergence",
    model_kwargs={"load_in_4bit": True},
    device_map="auto")

prompt = "Adapte esse texto para TEA:\n\nTexto: ..."

sequences = pipe(
    f"<s>[INST] {prompt} [/INST]",
    do_sample=True,
    max_new_tokens=4096,
    temperature=1,
    top_k=50,
    top_p=0.95,
    num_return_sequences=1,)

print(sequences[0]['generated_text'])
```

6 ANÁLISE DE RESULTADOS

O objetivo de obter os passos necessários e aplicar o *fine-tuning* ao modelo de inteligência artificial foram alcançado, o modelo está gerando adaptações. Era esperado que ele geraria também algumas inconsistências mas foi possível constatar a capacidade de se ajustar para resolver o problema de adaptação do documentos textuais.

Os *prompts* utilizados funcionaram como esperado, estão distinguindo entre as neurodivergências abordadas.

Ao analisar o conteúdo gerado pelo modelo treinado, observou-se que alguns trechos estavam adaptados, mas que existiam inconsistências e erros claros.

Ocorreram repetições de trechos nessas excessões, onde um mesmo parágrafo que foi adaptado aparece várias vezes, o que provocou um aumento de *tokens*. Outro ponto errôneo é o caso de repetições de números e símbolos, como o "@"por exemplo.

Exemplo de teste que ocorreu muitos erros:

Prompt de entrada:

> Adapte esse texto para TEA:
>
> Texto: práticas, escrita e especialidades. Nos artigos publicados nas revistas dedicadas a um público letrado mais amplo, percebese um nítido voluntarismo dos cientistas em explicitar os novos padrões de identidade intelectual - pontos que pretendiam firmar publicamente. Enesse caso, se desejavam tornar públicas as suas respectivas atribuições e características, tais veículos eram perfeitos. Muito mais adequados, aliás, que as revistas médico-científicas, em que, na verdade, apresentavam os resultados de suas pesquisas e reflexões sem discutir ou contra-argumentar as normas e códigos de conduta da atividade científica.7 Olivro constitui-se em sete capítulos. No capítulo 1, discute-se o tratamento dado ao tema pela historiografia ...

E a saída obtida foi:

> Lima, que fez o primeiro leiturismo completo do livro e deu início à longa fila de editores, entrejubilados e aposentados, que me ajudaram no vento 1 2 3 4 5 6!??@@@@@@ @@@@@@@@@@@@@@@@@ @@@@@@@@@@@@@@@@@@ @@@@@@@@@@@ @@@‘’ - registau as’s
>
> Também quero agradecer aos bibliotecasres dos jornais: Arquivo Histórico das Casas Nobre do Rio de Janeiro (AHCCN-RJ), Fundação Joaquim Nabuco

(FJN); Arquivo Geral da Prefeitura do Rio de Janeiro (AGPRJ); Conjunto Documental - Cultura e Sociedade no Brasil (CDCSB) no sítio LABS; e aos arquivos pessoais de Oswaldo Cruz, Francisco Antônio Propálcio, José Marciano Vianna Gomes e Miguel Ozório de Almeida.

Ao Chefe do Departamento de Pesquisa e Inovação da Casa de Oswaldo ...

O exemplo não está inteiro pois o seu tamanho é extenso mas pode-se notar que aconteceu erros grotescos, como repetições de @, uso excessivo de siglas e paralavras escritas de forma errada.

A seção 6.1 discute os problemas enfrentados no *fine-tuning*. A seção 6.2 informa sobre a viabilidade do treinamento do modelo. A seção 6.3 informa sobre possíveis questões legais.

6.1 Problemas enfrentados

Devido a questões financeiras não foi possível realizar um processo de *fine-tuning* adequado devido ao preço dos equipamentos.

O Google fornece acesso gratuito à GPU T4, no entanto, o uso prolongado, por algumas horas, resulta em uma suspensão indeterminada do acesso que pode levar dias até ser restabelecido. Este é um ponto importante a registrar, pois os modelos de inteligência artificial generativos baseados em *Transformers* são novidades, e atualmente há uma corrida entre organizações para adquirir os melhores equipamentos, o que se refletiu no estudo.

6.2 Viabilidade do fine-tuning

É possível notar que, com baixo investimento, já houve um resultado promissor. Em *fine-tuning* executados pela comunidade, geralmente se usam *datasets* com mais de 10 mil *prompts* e mais *epochs*. Pode-se concluir que é possível o investimento em ajustar modelos para essa finalidade.

6.3 Questões legais

À medida que essas soluções de IAs de conversação, geração de imagem e tudo que permeia a GenAI se tornaram populares recentemente, ainda persistem muitas questões legais em aberto, como quem detém os direitos sobre o material gerado e quais materiais podem ser usados para o treinamento delas.

Considerando que o objetivo implicaria uma manipulação dos dados do usuário, quando utilizando o modelo treinado, tanto no acesso a assuntos de cunho pessoal quanto

na modificação do conteúdo lido. Apesar das boas intenções, podem ocorrer erros, algo comum em todas as *GenAI*, e também há ataques direcionados a elas. Se esse modelo treinado fosse infectado e, consequentemente, causasse prejuízos, tanto financeiros quanto emocionais ao indivíduo, quem assumiria a responsabilidade?

7 CONCLUSÃO

O objetivo deste trabalho foi encontrar uma metodologia capaz de extrair padrões para a adaptação de textos para pessoas neurodivergentes e verificar a possibilidade de realizar o processo de *fine-tuning* em um modelo de Inteligência Artificial Generativa (*GenAI*) de código aberto. Esta necessidade surge do fato de que pessoas com neurodivergência podem ter dificuldades na compreensão de textos, mesmo que a leitura seja realizada de forma satisfatória, e poucos materiais, como livros, oferecem edições acessíveis para esse público.

Devido à falta de estudos suficientes, uma vez que a maioria dos trabalhos correlatos se concentra em testes de compreensão e não em como realizar a adaptação, não foi possível verificar a eficácia do método de adaptação utilizado. Porém, foram realizados testes rápidos sobre a eficácia dos *GenAI* disponíveis comercialmente em adaptar, destacando que apenas o GPT-4 apresentou resultados satisfatórios.

No processo de *fine-tuning*, utilizando *prompts* gerados pelo GPT-4 e o modelo de código aberto Mistral 7B, o resultado foi parcial. Foi possível especializar o modelo para adaptação, mas problemas, principalmente financeiros, impediram um treinamento mais adequado. Interessantemente, o modelo Mistral 7B, que inicialmente utilizava apenas a língua inglesa, passou a incorporar a língua portuguesa, demonstrando uma capacidade de se adaptar ao *dataset* fornecido para treino.

A escassez em trabalhos correlatos que colaborem para a criação dos padrões de adaptação é evidente. No entanto, dentre os trabalhos encontrados, foi possível concluir que os métodos de adaptação precisam ser diversificados para uma avaliação mais efetiva. Características e dificuldades devem ser registradas para cada tipo de neurodivergência, a fim de evitar generalizações.

Diversos testes mostraram que, quando solicitada a adaptação para uma neurodivergência específica, os modelos se comportavam de forma confusa, indicando que eles não foram treinados com material sobre acessibilidade. Portanto, a partir de padrões identificados, houve a hipótese de que seria possível especializar uma *GenAI*. O grande desafio, no entanto, é encontrar, a partir de novos estudos, formas comprovadas de descobrir a melhor maneira de adaptar os documentos textuais.

Conclui-se, portanto, que são necessárias etapas anteriores de pesquisa para avaliar melhor como proceder com a adaptação. Com o avanço da tecnologia de *GenAI*, surge a possibilidade de essas pessoas terem independência ao acessar materiais adaptados, sem depender das boas intenções de empresas e escritores.

7.1 Trabalhos futuros

Além de métodos para testar a eficácia, recomenda-se também a criação de estudos que elaborem métodos para identificar e classificar textos ou parágrafos problemáticos para neurodivergentes. Isso inclui desafios como o uso de linguagem indireta e palavras homófonas, que, embora listadas em dicionários, são apenas exemplos limitados. Além disso, seria útil desenvolver métodos mais eficazes para simplificar e dividir parágrafos, com o objetivo de aumentar o foco durante a leitura. Essas medidas facilitariam a construção de datasets melhores e mais adaptados, alcançando assim o objetivo de grande valor para o público-alvo.

REFERÊNCIAS

[1] SHAH, P. J. et al. Neurodevelopmental disorders and neurodiversity: definition of terms from scotland's national autism implementation team. *The British Journal of Psychiatry*, Cambridge University Press, v. 221, n. 3, p. 577–579, 2022.

[2] JIANG, A. Q. et al. *Mistral 7B*. 2023.

[3] DETTMERS, T. et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023.

[4] XU, L. et al. *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. 2023.

[5] ORTEGA, F. O sujeito cerebral e o movimento da neurodiversidade. *Mana*, Programa de Pós-Graduação em Antropologia Social - PPGAS-Museu Nacional, da Universidade Federal do Rio de Janeiro - UFRJ, v. 14, n. 2, p. 477–509, Oct 2008. ISSN 0104-9313. Disponível em: <https://doi.org/10.1590/S0104-93132008000200008>.

[6] FINNEGAN, E.; MAZIN, A. L. Strategies for increasing reading comprehension skills in students with autism spectrum disorder: A review of the literature. *Education and Treatment of Children*, West Virginia University Press, v. 39, n. 2, p. 187–219, May 2016.

[7] NATION, K. et al. Patterns of reading ability in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, v. 36, n. 7, p. 911–919, out. 2006.

[8] OSTOITS, J. Reading strategies for students with add and adhd in the inclusive classroom. *Preventing School Failure: Alternative Education for Children and Youth*, Routledge, v. 43, n. 3, p. 129–132, 1999.

[9] AARON, P. G.; FRANTZ, S. S.; MANGES, A. R. Dissociation between comprehension and pronunciation in dyslexic and hyperlexic children. *Reading and Writing*, v. 2, p. 243–264, 1990.

[10] JEBARA, T. *Machine Learning*: Discriminative and generative. [S.l.]: Springer New York, NY, 2004. (The Springer International Series in Engineering and Computer Science). Springer Science+Business Media New York 2004. ISBN 978-1-4020-7647-3.

[11] FLECK, L. et al. Redes neurais artificiais: Princípios básicos. *Revista Científica e Tecnológica*, v. 7, n. 15, p. 4330, 2023.

[12] NAVEED, H. et al. *A Comprehensive Overview of Large Language Models*. 2023.

[13] VASWANI, A. et al. *Attention Is All You Need*. 2017.

[14] TOUVRON, H. et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023.

[15] FLETCHER, J. M. Measuring reading comprehension. *Scientific Studies of Reading*, Routledge, v. 10, n. 3, p. 323–330, 2006.

[16] MILLIS, K.; MAGLIANO, J.; TODARO, S. Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading*, Routledge, v. 10, n. 3, p. 225–240, 2006.

[17] DEANE, P. et al. Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading*, Routledge, v. 10, n. 3, p. 257–275, 2006. Disponível em: <https://doi.org/10.1207/s1532799xssr1003_4>.

[18] RAYNER, K.; CHACE, T. J. S. K. H.; ASHBY, J. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, Routledge, v. 10, n. 3, p. 241–255, 2006.