



UNIVERSIDADE FEDERAL DE SANTA CATARINA CENTRO DE COMUNICAÇÃO E  
EXPRESSÃO DEPARTAMENTO DE LÍNGUA E LITERATURA VERNÁCULAS CURSO  
DE GRADUAÇÃO EM LETRAS - LÍNGUA PORTUGUESA E LITERATURAS DE  
LÍNGUA PORTUGUESA

Camila de Oliveira Muniz

**Inteligência Artificial e os riscos à Diversidade Linguística na Comunidade Lusófona.**

Florianópolis  
2024

Camila de Oliveira Muniz

**Inteligência Artificial e os Riscos à Diversidade Linguística na Comunidade Lusófona.**

Trabalho de Conclusão de Curso submetido ao curso de Letras - Língua Portuguesa e Literaturas de Língua Portuguesa do Centro de Comunicação e Expressão da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Bacharel em Letras - Língua Portuguesa e Literaturas de Língua Portuguesa.

Orientador(a): Prof.(a): Dr.(a) Gilvan Müller de Oliveira

Florianópolis  
2024

Muniz, Camila de Oliveira  
Inteligência Artificial e os Riscos à Diversidade  
Linguística na Comunidade Lusófona. / Camila de Oliveira  
Muniz ; orientador, Gilvan Muller de Oliveira, 2024.  
53 p.

Trabalho de Conclusão de Curso (graduação) -  
Universidade Federal de Santa Catarina, Centro de  
Comunicação e Expressão, Graduação em Letras - Língua  
Portuguesa, Florianópolis, 2024.

Inclui referências.

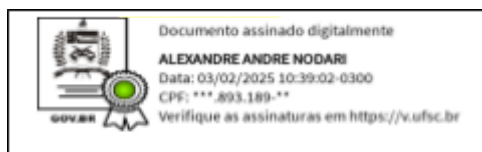
1. Letras - Língua Portuguesa. 2. Diversidade  
Linguística. 3. Inteligência Artificial . I. Oliveira,  
Gilvan Muller de. II. Universidade Federal de Santa  
Catarina. Graduação em Letras - Língua Portuguesa. III.  
Título.

Camila de Oliveira Muniz

Inteligência Artificial e os Riscos à Diversidade Linguística na Comunidade Lusófona.

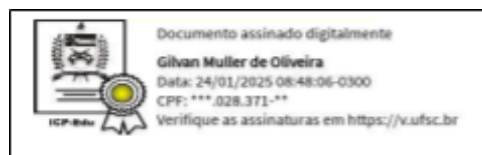
Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Inteligência Artificial e os riscos à Diversidade Linguística na Comunidade Lusófona. e aprovado em sua forma final pelo Curso de Letras - Língua Portuguesa e Literaturas de Língua Portuguesa.

Florianópolis, 11 de Dezembro de 2024.

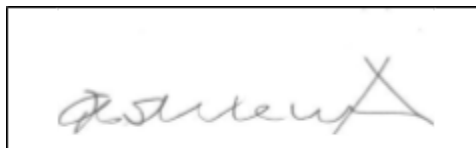


Coordenação do Curso

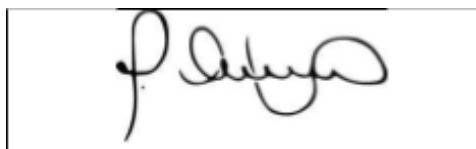
**Banca examinadora**



Prof. Dr. Gilvan Muller de Oliveira  
Orientador

A handwritten signature in black ink, enclosed in a rectangular box. The signature appears to be 'Cláudio Menezes'.

Prof. Dr. Cláudio Menezes Universidade  
de Brasília

A handwritten signature in black ink, enclosed in a rectangular box. The signature appears to be 'Thaianne Moreira de Oliveira'.

Prof.(a) Dr.(a). Thaianne Moreira de Oliveira  
Universidade Federal Fluminense

Florianópolis, 2024.

## RESUMO

O desenvolvimento acelerado dos últimos anos na área de inteligência artificial (IA) desencadeou impactos no campo da comunicação e linguagem, especialmente no que se refere à diversidade linguística. O presente estudo busca analisar os potenciais riscos do uso de aplicações em IA na supressão da diversidade linguística, com ênfase na língua portuguesa e suas variantes, por meio da investigação das implicações culturais de uma possível homogeneização linguística do português e seus desdobramentos na pluralidade dos dialetos, regionalismos e expressões culturais particulares no contexto brasileiro. A pesquisa, de abordagem exploratória-descritiva, possui como fundamento teórico a revisão bibliográfica sistemática, contemplando produções acadêmicas, documentações técnicas e recomendações internacionais. Os resultados visam destacar a necessidade de um planejamento linguístico e políticas de incentivo que assegurem a equidade digital da língua portuguesa perante as tecnologias de IA, com objetivo de favorecer a inclusão linguístico-cultural dos falantes de língua portuguesa no ecossistema tecnológico contemporâneo. Como conclusão, o estudo ressalta a importância da preservação linguística como um ativo estratégico para a manutenção da identidade cultural e a representatividade social da comunidade lusófona em ambientes digitais.

Palavras-chave: Inteligência Artificial; Diversidade Linguística; Língua Portuguesa.

## **ABSTRACT**

The accelerated development of artificial intelligence (AI) in recent years has had an impact on the field of communication and language, especially with regard to linguistic diversity. This study seeks to analyse the potential risks of using AI applications to suppress linguistic diversity, with an emphasis on the Portuguese language and its variants, by investigating the cultural implications of a possible linguistic homogenization of Portuguese and its consequences for the plurality of dialects, regionalisms and particular cultural expressions in the Brazilian context. The research, with an exploratory-descriptive approach, has a systematic bibliographic review as its theoretical basis, including academic productions, technical documentation and international recommendations. The results aim to highlight the need for linguistic planning and incentive policies to ensure the digital equity of the Portuguese language in the face of AI technologies, with the aim of favoring the linguistic-cultural inclusion of Portuguese speakers in contemporary technological ecosystem. In conclusion, the study highlights the importance of linguistic preservation as a strategic asset for maintaining the cultural identity and social representativeness of the Portuguese-speaking community in digital environments.

Keywords: Artificial Intelligence; Linguistic Diversity; Portuguese Language.

## LISTA DE ABREVIATURAS E SIGLAS

<b>AIPI</b>	Índice de Prontidão para IA
<b>AWS</b>	Amazon Web Services
<b>CNN</b>	Redes Neurais Convolucionais
<b>FMI</b>	Fundo Monetário Internacional
<b>GPT-3</b>	Generative Pre-trained Transformer 3
<b>GPU</b>	Unidade de Processamento Gráfico
<b>IA</b>	Inteligência Artificial
<b>INDL</b>	Inventário Nacional da Diversidade Linguística
<b>IPHAN</b>	Instituto do Patrimônio Histórico e Artístico Nacional
<b>LLMs</b>	Large Language Models (Grandes Modelos de Linguagem)
<b>Llama</b>	Modelo de Linguagem da Meta
<b>NIC</b>	Núcleo de Informação e Coordenação do Ponto BR
<b>NLP</b>	Natural Language Processing (Processamento de Linguagem Natural)
<b>OBDILCI</b>	Observatório da Diversidade Linguística e Cultural na Internet
<b>PLN</b>	Processamento de Linguagem Natural
<b>RNN</b>	Redes Neurais Recorrentes
<b>UNESCO</b>	Organização das Nações Unidas para a Educação, a Ciência e a Cultura

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>7</b>
1.1	METODOLOGIA	8
<b>2</b>	<b>INTELIGÊNCIA ARTIFICIAL: CONTEXTUALIZAÇÃO HISTÓRICA E FUNDAMENTOS DOS GRANDES MODELOS DE LINGUAGEM</b>	<b>11</b>
2.1	SUBÁREAS DO CAMPO DE INTELIGÊNCIA ARTIFICIAL	12
2.2	OS GRANDES MODELOS DE LINGUAGEM ( <i>LLMs</i> )	14
2.3	TOKENIZAÇÃO E EMBEDDING	17
2.4	A ARQUITETURA TRANSFORMER	18
<b>3</b>	<b>MACROECONOMIA: OS IMPACTOS GEOPOLÍTICOS DA INTELIGÊNCIA ARTIFICIAL</b>	<b>21</b>
3.1	ÍNDICE DE PRONTIDÃO PARA IA (AIPI) - FUNDO MONETÁRIO INTERNACIONAL	22
3.2	A COMUNIDADE LUSÓFONA E A PRONTIDÃO PARA IA	24
<b>4</b>	<b>GOVERNANÇA E LÍNGUA: A PRESENÇA DO PORTUGUÊS NA ERA DIGITAL</b>	<b>29</b>
4.1	DESAFIOS DE INFRAESTRUTURA NO BRASIL	30
4.2	GOVERNANÇA DA INFORMAÇÃO	32
4.3	TREINAMENTO DE GRANDES MODELOS DE LINGUAGEM	34
4.4	A PRESENÇA DA LÍNGUA PORTUGUESA NA INTERNET	35
4.1	LÍNGUA E PODER	38
<b>5</b>	<b>DIVERSIDADE LINGUÍSTICA</b>	<b>40</b>
6.1	EXEMPLOS DE PROBLEMAS DE ALINHAMENTO LINGUÍSTICO E CULTURAL - PORTUGUÊS BRASILEIRO	43
<b>6</b>	<b>CONCLUSÃO</b>	<b>48</b>

## 1 INTRODUÇÃO

Os avanços recentes na área da Inteligência Artificial (IA) têm afetado diversos campos, inclusive possui impactos diretos no domínio da comunicação e da linguagem. No contexto contemporâneo de interação entre humanos e máquinas, algumas comunidades linguísticas enfrentam desafios estruturais de equidade no acesso ao ambiente digital e tecnológico, cenário que compromete a diversidade linguística e a preservação de línguas minoritárias. Essa assimetria tecnológica apresenta-se como uma barreira de acesso, implicando em um potencial apagamento linguístico-cultural no contexto de tecnologias emergentes. A partir desse cenário, a pesquisa a seguir busca analisar os potenciais riscos do uso de sistemas de Inteligência Artificial na supressão da diversidade linguística, com ênfase em especial nos falantes de língua portuguesa.

O estudo visa elencar pontos de tensão relativos ao apagamento linguístico-cultural da língua portuguesa, discutindo como as tecnologias de Inteligência Artificial podem comprometer a pluralidade do português, tendo como desdobramentos a invisibilidade de dialetos, regionalismos e expressões culturais dentro do contexto do português brasileiro. Além disso, o estudo propõe-se a analisar a posição da comunidade lusófona perante os países que possuem maior preparo e prontidão para uso de ferramentas de IA ...buscando diagnosticar os desafios e caminhos possíveis para que os falantes do português não sofram com a uniformização linguística, mas que, sobretudo, mantenha a língua portuguesa bem articulada em um ecossistema tecnológico.

A investigação sobre a preservação da diversidade linguística do português frente às tecnologias de inteligência artificial é estratégica, pois evidencia os processos de homogeneização que as línguas podem sofrer no âmbito digital. Considerando que as línguas são, acima de tudo, um patrimônio cultural (IPHAN, 2014) carregado de identidades, histórias e memórias, abordar problemas como esses contribui para proteger um sistema linguístico e preservar a existência cultural de comunidades inteiras que se expressam por meio das especificidades de suas línguas.

Assim, essa pesquisa contribui para o desenvolvimento de políticas que assegurem o uso inclusivo das tecnologias emergentes, respeitando as particularidades culturais e linguísticas. Busca-se um olhar atento para a equidade no acesso e uso da IA, evitando

apagamento digital e garantindo a representatividade da diversidade linguística em cenário global. O estudo discute o fortalecimento da identidade cultural da comunidade lusófona ao apontar argumentos técnicos e sociais, trazendo para o centro da análise soluções que sejam sustentáveis e respeitadas com a pluralidade cultural das línguas.

## 1.1 METODOLOGIA

A pesquisa adota uma abordagem qualitativa de caráter exploratório-descritivo, consistindo em uma revisão bibliográfica como a principal estratégia metodológica. A escolha desta abordagem justifica-se pela necessidade de construir um entendimento amplo sobre as complexas interações entre inteligência artificial e a diversidade linguística, com foco na língua portuguesa e suas variantes. Conforme mencionado por Gil (2008) a revisão bibliográfica sistemática constitui-se como instrumento importante para mapear o estado da arte do conhecimento, identificar lacunas teóricas e práticas na literatura existente, portanto, aplicada a temática, estabelece conexões entre diferentes perspectivas para fundamentar análises críticas sobre as implicações sociais, culturais e tecnológicas associadas à preservação das línguas no ciberespaço.

O corpus da pesquisa contempla diversos formatos de documentos, incluindo produção acadêmica, documentação técnica e recomendações éticas em IA por agentes internacionais, como a UNESCO. O estudo reconhece suas limitações metodológicas, visto o dinamismo no campo da IA, entretanto, para Gil (2008), a pesquisa qualitativa tem como característica principal a exploração e compreensão de fenômenos complexos a partir de uma análise contextualizada com as informações disponíveis.

Tendo em vista que nessa abordagem é necessário buscar interpretar os dados considerando os significados subjacentes e as relações presentes nos textos analisados, visando um olhar integrado do tema investigado. A metodologia foi estruturada para alcançar um panorama atualizado do tema, identificar tendências e espaços para aprimoramento, ao mesmo tempo que, dialoga com possíveis caminhos para contribuição com políticas linguísticas voltadas para o âmbito de inteligência artificial e o processamento de línguas, em especial, que garanta a diversidade linguística e a inclusão cultural.

## **2 INTELIGÊNCIA ARTIFICIAL: CONTEXTUALIZAÇÃO HISTÓRICA E FUNDAMENTOS DOS GRANDES MODELOS DE LINGUAGEM**

Mesmo com a efervescência atual em torno da temática "inteligência artificial", o desenvolvimento desta área de investigação se dá há mais de sete décadas. O pontapé inicial surgiu por volta da década de 1950, quando o matemático Alan Turing se propôs a responder à pergunta norteadora para o avanço subsequente do campo: "Podem as máquinas pensar?" (TURING, 1950). Para solucionar esta questão epistemológica, Turing criou uma metodologia experimental inicialmente chamada de "jogo da imitação". O experimento consistia em uma dinâmica envolvendo três participantes - um homem, uma mulher e um juiz (mulher ou homem), todos isolados em ambientes distintos e comunicando-se através de mensagens datilografadas. O objetivo principal desta primeira fase era avaliar a capacidade de persuasão dos participantes em convencer o juiz sobre suas respectivas identidades de gênero. Nesta etapa, era necessário conduzir o juiz ao erro, se fazendo passar por uma mulher e um homem, respectivamente.

Em seguida, Turing introduziu uma nova fase ao experimento, substituindo um dos participantes humanos por uma máquina, o que posteriormente viria a ser conhecido como "Teste de Turing". Nesta nova fase, o desafio consistia em um diálogo entre um ser humano e uma máquina, cabendo ao juiz a tarefa de discernir qual dos participantes era humano e qual era uma máquina. Embora posteriormente o experimento tenha sido contestado como um método realmente válido para medir a inteligência de uma máquina, o "Teste de Turing" se consolidou como um procedimento eficaz para avaliar a capacidade das máquinas em reproduzir comportamentos humanos.

A partir do "Teste de Turing", o campo de ciências da computação experimentou uma rápida expansão conceitual e metodológica. Um momento importante desse crescimento ocorreu durante o evento "Conferência de Dartmouth" em 1956, quando John McCarthy apresentou o termo "Inteligência Artificial" que hoje é amplamente aceito e utilizado.

Os desdobramentos após à conferência de Dartmouth intensificaram ainda mais as investigações no campo da IA, que até então era apenas um campo em ascensão, e resultou em um desenvolvimento acelerado, abrindo espaço para múltiplas linhas de pesquisa e a diversificação das abordagens metodológicas. Décadas mais tarde, houve um novo momento decisivo para a consolidação teórica do campo da inteligência artificial. Em 1995, Stuart Russell e Peter Norvig publicaram o livro *Inteligência Artificial: Uma Abordagem Moderna* (RUSSELL; NORVIG, 1995), que representa um divisor de águas para a legitimação teórica do campo da IA. O livro fornece um conteúdo abrangente sobre conceitos e técnicas, e foi considerado o determinante para posicionar de fato a área da inteligência artificial como disciplina perante a comunidade científica internacional.

Por sua trajetória de longa data, o arcabouço do campo da inteligência artificial é extenso e continua com uma crescente complexidade, visto que abrange e, ao mesmo tempo, se interliga com diversas outras áreas de conhecimento, como bem descrito pela professora e pesquisadora brasileira Dora Kaufman, em seu artigo *Inteligência Artificial: Repensando a mediação*:

A IA propicia a simbiose entre o humano e a máquina ao acoplar sistemas inteligentes artificiais ao corpo humano [...], e a interação entre o homem e a máquina como duas entidades distintas conectadas (homem-aplicativos, homem-algoritmos de IA). Tema de pesquisa em diversas áreas - computação, linguística, filosofia, matemática, neurociência, entre outras -, a diversidade de subcampos e atividades, pesquisas e experimentações, dificulta descrever o "estado-da-arte" atual da IA. Os estágios de desenvolvimento bem como as expectativas variam entre os campos e suas aplicações, que incluem os veículos autônomos, reconhecimento de voz, games, robótica, tradução de linguagem natural, diagnósticos médicos, assim por diante. Atualmente, os sistemas inteligentes permeiam praticamente todas as áreas de conhecimento (KAUFMAN, 2020).

## 2.1 SUBÁREAS DO CAMPO DE INTELIGÊNCIA ARTIFICIAL

Atualmente, o campo da inteligência artificial abrange um amplo espectro de subáreas especializadas, cada uma com suas ramificações, metodologias e aplicações específicas. Dentre as principais subáreas, destaca-se o processamento de linguagem natural (*Natural Language Processing*), o aprendizado de máquina (*Machine Learning*) e o aprendizado profundo (*Deep Learning*).

Figura 1 - Subáreas de Inteligência Artificial



Fonte: INTELIGÊNCIA ARTIFICIAL E CULTURA: perspectivas para a diversidade cultural na era digital, 2022, p. 104).

O entendimento da organização complexa e da profundidade das subáreas da inteligência artificial é fundamental para entender sua multidisciplinaridade. Seguindo uma sequência lógica, a IA é a grande área ou o "guarda-chuva", no qual a aprendizagem de máquina (*Machine Learning*) é uma das suas principais ramificações. O aprendizado de máquina, por sua vez, abrange o aprendizado profundo (*Deep Learning*), que se fundamenta principalmente no estudo e desenvolvimento de redes neurais artificiais (*Neural Networks*), que são arquiteturas algorítmicas complexas e com múltiplas camadas de processamento. O desenvolvimento das redes neurais é inspirado no funcionamento do cérebro humano e no processamento hierárquico das informações.

É a partir dessas diversas subdivisões que o campo da inteligência artificial se configura como uma área de estudo dinâmica e heterogênea, caracterizada pela interligação de diferentes disciplinas, em que cada uma dessas ramificações, ou subdivisões, mantém conexões indispensáveis entre si.

## 2.2 OS GRANDES MODELOS DE LINGUAGEM (*LLMs*)

Compreender a definição e o papel dos grandes modelos de linguagem (*LLMs - Large Language Models*) é essencial para posicioná-los dentro do contexto da inteligência artificial. Em uma definição básica, a IBM (*s.d*) afirma que os *LLMs* são projetados para entender e gerar texto como um humano, além de outras formas de conteúdo, com base na vasta quantidade de dados utilizados para treiná-los. Aprofundando para um recorte específico, é possível compreender os grandes modelos de linguagem como subconjuntos de algoritmos de aprendizado profundo (*Deep Learning*) (IBM, *s.d*).

Os *LLMs* utilizam complexas arquiteturas de redes neurais artificiais para gerar, processar e compreender textos, buscando simular, em certa medida, o funcionamento do cérebro humano. As redes neurais, que são a arquitetura de sustentação dos grandes modelos de linguagem, são compostas por múltiplas camadas interconectadas de unidades computacionais, também conhecidas como nós ou neurônios artificiais,

Redes neurais artificiais procuram imitar alguns aspectos da organização neuronal observada em organismos biológicos, sendo compostas de camadas que se interconectam, cada camada contendo um conjunto de pequenas unidades de

computação simples chamadas de neurônios artificiais (COSTA; COZMAN, 2024, p 138 – 139).

Essa estrutura hierárquica é organizada em três partes principais: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Portanto, o fluxo de dados percorre por essa estrutura de forma sequencial, ingressando na camada de entrada, propagando-se através das camadas ocultas e na sequência atingindo a camada de saída. É durante o processamento nas camadas ocultas que ocorrem os ajustes dos pesos<sup>1</sup> e parâmetros<sup>2</sup> da rede, visando o refinamento progressivo da informação e resultando valores de saídas cada vez mais alinhadas com os objetivos pré-treinados do modelo (AWS, *s.d*).

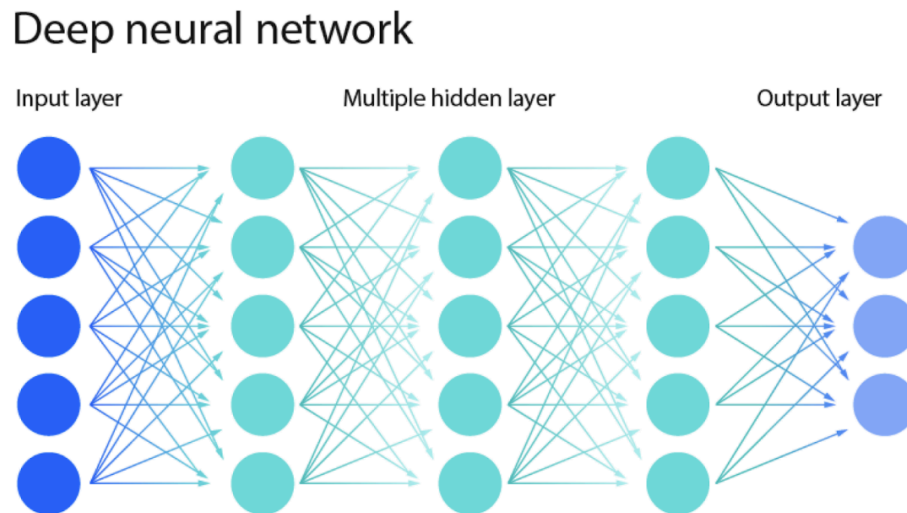
Existem diferentes modelos de redes neurais e com diversos objetivos. No caso específico dos grandes modelos de linguagem, a arquitetura de rede neural mais comum é a *Feed Forward*. Essa configuração é caracterizada por um fluxo de informações unidirecional (DATABRICKS, *s.d*), ou seja, os dados percorrem sequencialmente da camada de entrada, atravessando as camadas ocultas, até atingirem a camada de saída, sem conexão recorrente ou retroalimentação entre as camadas, sendo o principal papel das camadas ocultas realizar transformações e extrair representações cada vez mais complexas a partir dos dados de entrada.

---

<sup>1</sup> Pesos são valores que determinam a força da conexão entre neurônios, para decidir o quanto cada informação é importante e são ajustados durante o treinamento da rede neural.

<sup>2</sup> Parâmetros são valores ajustáveis (pesos e vieses) em um modelo que são aprendidos durante o treinamento com dados.

Figura 2 - Redes Neurais Múltiplas Camadas



Fonte: O que é uma rede neural?. IBM, s.d.<sup>3</sup>

O processo de aprendizado das redes neurais artificiais ocorre através da retropropagação (*backpropagation*), em que os parâmetros internos são ajustados para minimizar a divergência entre as saídas geradas e as respostas esperadas. A retropropagação permite que as camadas ocultas da rede extraiam por conta própria as características e representações mais relevantes dos dados de entrada, gerando, assim, uma saída otimizada. É relevante compreender a importância dos pesos e parâmetros para o desempenho dos *LLMs* e a sua influência nas respostas fornecidas, visto que esses elementos numéricos são responsáveis pela codificação do conhecimento adquirido durante a etapa de treinamento da rede. No artigo Viés no Aprendizado de Máquina em Sistemas de Inteligência Artificial: a diversidade de origens e os caminhos de mitigação, Cozman e Kaufman (2022) explicam a importância dos pesos e parâmetros:

---

<sup>3</sup> IBM. Rede neural profunda, 2024. Disponível em: <https://www.ibm.com/br-pt/topics/neural-networks#:~:text=As%20redes%20neurais%20feedforward%2C%20ou%20uma%20camada%20de%20sa%C3%ADda>. Acesso em: 25 out. 2024.

Em redes neurais profundas, os parâmetros aprendidos a partir de dados são chamados de weights (pesos); após a fase de treinamento (ou aprendizado), esses pesos compõem o algoritmo e passam a ser fixos. No caso de uma imagem, em que os pixels são os dados de entrada, a saída do sistema reflete a soma das multiplicações de pesos pelos pixels de entrada. Cada camada processa o que supõe-se serem conceitos mais abstratos do que da camada anterior, gerando o nível de abstração requerido pela saída. Por exemplo, a saída pode ser dog vs cat, e a entrada pode ser a imagem (conjunto de pixels); cada camada mais “profunda” (mais próximo da saída) tem valores representando conceitos mais abstratos que ajudam, eventualmente, a concluir se é gato ou cachorro. A questão da interpretabilidade (ou opacidade, ou não explicabilidade) decorre do desconhecimento do que as camadas realmente representam (KAUFMAN; COZMAN, 2022, p.197).

## TOKENIZAÇÃO E EMBEDDING

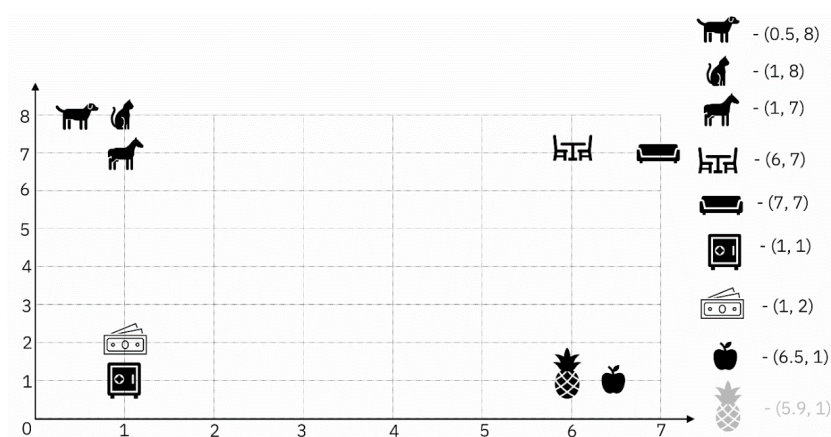
A operacionalização dos grandes modelos de linguagem está baseada em uma arquitetura de processamento dividida por etapas. A sua organização compreende em três etapas: Pré-processamento – fase inicial onde ocorre a preparação e estruturação dos dados de entrada; processamento central - estágio onde ocorre a aplicação e suas operações computacionais específicas e por fim, o pós-processamento – etapa final dedicada ao refinamento e adequação dos resultados gerados.

Esta estrutura organizada por etapas contribui para a confiabilidade do processamento de linguagem em larga escala, sendo um esquema indispensável para a engenharia dos sistemas de IA. Sendo assim, vale dar destaque para alguns processos que ocorrem na etapa de pré-processamento, visto que é nessa etapa que os dados de treinamento são coletados e passam por uma filtragem e limpeza. Algumas das etapas de pré-processamento são, por exemplo, a normalização, tokenização e remoção de *stop words*. A tokenização do conteúdo é uma operação indispensável, uma vez que transformam conteúdo textual em unidades numéricas. Como os algoritmos de aprendizado de máquina são fundamentalmente cálculos matemáticos que processam entradas binárias (0 e 1), a tokenização permite que dados textuais sejam interpretados.

A etapa de tokenização, é sobretudo uma prática do campo do Processamento de Linguagem Natural (PLN) e existem mais de uma abordagem para transformar as sentenças em tokens: segmentação por caracteres individuais, palavras completas, volume de tokens e unidades morfológicas (radicais), sendo esta, a separação amplamente aceita atualmente.

Embora a tokenização seja uma etapa imprescindível para segmentar o texto em unidades manipuláveis computacionalmente, os tokens isolados são estáticos, ou seja, unidimensional, logo, é necessário que exista uma representação multidimensional para que semanticamente o conteúdo faça sentido. Para que estes elementos linguísticos possam ser compreendidos e processados pelos algoritmos, é necessário que eles sejam transformados em representações vetoriais multidimensionais, processo este conhecido como *embedding*.

Figura 3 – Etapa de *embedding*



Fonte: Token e Embedding: conceitos da IA e LLMs. **Brains.Dev**, 2024. <sup>4</sup>

É na etapa de *embedding*, que ocorre durante o estágio de processamento computacional, que os números deixam de ser tratados de forma isolada e passam a ser analisados a partir de seus contextos de uso. Assim, cada token é então codificado em um espaço multidimensional, de modo que suas características semânticas e sintáticas sejam capturadas de maneira consistente, não só refletindo as relações que a palavra possui com outras palavras dentro fluxo textual, mas também permitindo que o modelo compreenda nuances de significado e associações linguísticas.

Em resumo, o processo de tokenização na etapa de pré-processamento divide o texto em tokens, unidades menores que representam palavras ou partes delas. Em seguida, a etapa

<sup>4</sup> Disponível em: <<https://brains.dev/2024/token-e-embedding-conceitos-da-ia-e-llms/>>. Acesso em: 26 de out. de 2024.

de embedding converte esses tokens em vetores numéricos, permitindo que os algoritmos de aprendizado profundo reconheçam suas características semânticas e sintáticas no contexto do texto.

### 2.3 A ARQUITETURA TRANSFORMER

Até o início da década de 2010, as principais abordagens em Processamento de Linguagem Natural utilizavam redes neurais recorrentes (*RNNs*) ou redes neurais convolucionais (*CNNs*). As redes neurais recorrentes eram em particular bem sucedidas no processamento de dados sequenciais, visto que conseguiam armazenar certa "memória" ao longo da sequência processada, em um modelo *seq2seq*.<sup>5</sup> Já as convolucionais, por sua vez, se destacavam no processamento de dados espaciais, por exemplo, imagens, considerando que aplicavam filtros específicos para esse tipo de estrutura e dado.

Contudo, em 2017, pesquisadores da Google introduziram um novo elemento no campo do *PLN*: os modelos *Transformers*. Em vez de processarem as sentenças de forma sequencial, como as *RNNs*, os *Transformers* analisam os dados de uma única vez e em seguida aplicam um filtro conhecido como mecanismo de atenção. Esse mecanismo de atenção permite que o modelo atribua diferentes pesos de importância a cada palavra de entrada, em relação as outras. Dessa forma, isso significa que o modelo consegue compreender o contexto gerado, identificando palavras ou elementos que são mais relevantes dentro do contexto.

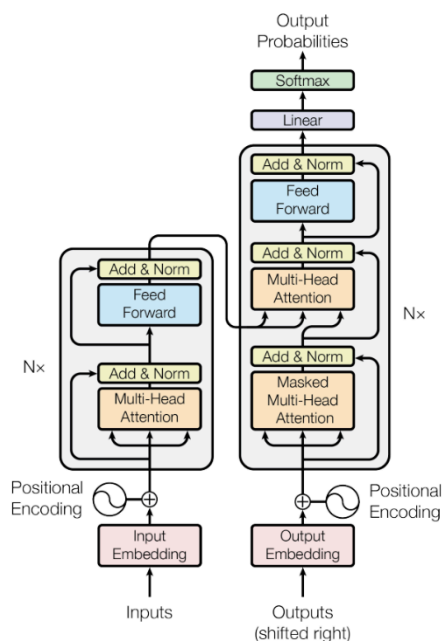
A capacidade de reconhecer relações complexas entre as palavras, processando todo o contexto de forma simultânea é o que difere os modelos *Transformers* dos demais modelos. Vaswani et al. (2017) explicam por que o mecanismo de atenção representa uma abordagem inovadora para o campo da inteligência artificial:

Os modelos de transdução de sequência dominante são baseados em redes neurais recorrentes ou convolucionais complexas que incluem um codificador e um decodificador. Os modelos de melhor desempenho também conectam o codificador e o decodificador por meio de um mecanismo de atenção. Propomos uma nova arquitetura de rede simples, o Transformer, baseada somente em mecanismos de atenção, dispensando recorrência e convoluções inteiramente (VASWANI, A. et al., 2017).

---

<sup>5</sup> Seq2seq é um modelo neural que converte uma sequência de entrada em outra sequência de saída, como na tradução de textos entre idiomas.

Figura 4 – A arquitetura de um modelo *Transformer*



Fonte: VASWANI, A. et al., 2017.

Além dos mecanismos de atenção, a arquitetura do modelo *Transformer* é formada por dois componentes básicos: o codificador (*encoder*) e o decodificador (*decoder*). O codificador processa o texto de entrada, transformando os dados em representações matemáticas que registram seu significado e contexto. Já o decodificador, a partir das representações numéricas produzidas pelo codificador, gera o texto de saída. Ao longo da arquitetura *Transformer*, há múltiplas camadas de codificadores e decodificadores, nas quais a informação é transmitida de um estágio a outro, considerando os mecanismos de atenção e filtros adicionais, até chegar à camada de saída. Essa organização permite que o modelo distribua atribuições e processe informações complexas de forma mais eficiente, resultando em respostas pertinentes ao contexto.

A diferença dos modelos *Transformers* para os demais acontece devido a sua capacidade de processar o contexto de forma ampla, ao invés de interpretar os tokens ou palavras de forma isolada, os Transformers utilizam o mecanismo de autoatenção

(*self-attention*) que permite compreender o significado de cada componente linguístico levando em conta todo o contexto à sua volta. Conseqüentemente, o mecanismo de autoatenção possibilita que os modelos considerem relações de longo alcance em um texto, identificando padrões que vão além da sequencialidade, assimilando o contexto global de uma entrada e as conexões entre ideias e/ou palavras que podem estar distantes entre si. Compreender o contexto amplo de uma entrada implica em respostas mais coerentes, fluídas e com maior precisão semântica, que se aproxima cada vez mais da habilidade humana de entender e gerar linguagem.

Ainda dentro da arquitetura *Transformer* e a sua capacidade de considerar um contexto amplo em uma sequência de dados, a janela de contexto dos *LLMs* é um parâmetro chave, pois limita a extensão do texto, medida em tokens ou unidades lexicais, que o modelo pode processar simultaneamente durante operações de interpretação de dados. A relevância deste parâmetro está intrinsecamente associada ao seu impacto na capacidade do modelo de gerar respostas coerentes e relevantes.

A amplitude da janela de contexto se refere ao volume de informação que o modelo pode processar, resultando na melhora da compreensão e na geração de textos, especialmente em interações mais longas. Ou seja, quanto maior a janela de contexto, mais informações o modelo pode processar de uma única vez, o que, por sua vez, ajuda a melhorar a compreensão do contexto. A capacidade de processar grandes volumes de textos é uma habilidade fundamental de ferramentas de inteligência artificial, visto que, contribui para que os modelos de linguagem façam conexões semânticas mais claras a partir de textos longos e conseqüentemente, gerem respostas mais alinhadas e compatíveis com base no prompt do usuário.<sup>6</sup>

### **3 MACROECONOMIA: OS IMPACTOS GEOPOLÍTICOS DA INTELIGÊNCIA ARTIFICIAL**

Os Grandes Modelos de Linguagem (*LLMs*) representam uma evolução no âmbito da inteligência artificial, transformando de maneira expressiva a dinâmica de interação entre

---

<sup>6</sup> Prompt é o comando dado pelo usuário a um *chatbot*, que usa um modelo *Transformer* para processá-lo.

humanos e máquinas. Os atuais *chatbots*<sup>7</sup> atuam como assistentes dotados e capacitados a partir de um imenso repertório de conhecimento<sup>8</sup>, sendo capazes de processar e gerar diversas modalidades de conteúdo. Sua versatilidade permite a execução autônoma de diferentes tipos de tarefas que normalmente demandavam a interação humana, como atendimento ao cliente, análises de dados, sumarização de conteúdo e até oferecendo sugestões personalizadas baseadas em padrões e comportamentos individuais. A discussão a respeito dessas tecnologias emergentes é um assunto de destaque global, não apenas pela alta aderência por parte dos usuários, mas também pelo impacto nas dinâmicas do mercado de trabalho, à medida que as empresas incorporam ferramentas baseadas em IA para otimizar processos e diminuir os custos.

Este cenário demanda um olhar sensível sob os impactos socioeconômicos, em especial, o movimento da força de trabalho no sul global, onde as consequências dessas mudanças tecnológicas podem ter um impacto maior nas estruturas sociais e econômicas. No contexto atual, essas tecnologias representam uma quebra em relação aos sistemas computacionais tradicionais, tendo em vista que até então esses sistemas eram limitados e possuíam especificidades de programação. As grandes empresas que se viam amarradas a processos que demandavam exclusivamente a intervenção humana, agora aderem ferramentas de IA para desempenhar diversas funções, sobretudo aquelas que não necessitam de ações intelectuais ou criativas, devido à versatilidade e capacidade de adaptação dos grandes modelos de linguagem em diferentes necessidades.

Porém, apesar do uso massivo de aplicações de IA por parte das empresas, novas habilidades de interação humano-máquina são concomitantemente requeridas para que o uso desses sistemas se torne de fato efetivo. Considerando as funções básicas de uma aplicação em IA, por exemplo, multifuncionalidade para executar tarefas e o acesso amplo à informação, o manuseio desses sistemas requer novas competências, como literacia digital e um pensamento crítico apurado.

---

<sup>7</sup> *Chatbot* é a interface que conecta o usuário final ao sistema, por exemplo, Gemini, ChatGPT, Copilot entre outros.

<sup>8</sup> O treinamento é feito a partir de um *dataset*, que engloba conteúdos disponíveis na internet, rótulos, metadados entre outros.

### 3.1 ÍNDICE DE PRONTIDÃO PARA IA (AIPI) - FUNDO MONETÁRIO INTERNACIONAL

O Índice de Prontidão para IA (AIPI), desenvolvido pelo FMI, tem como foco principal analisar o grau de preparação de diferentes países para a aderir sistemas baseados em inteligência artificial, visando captar de maneira macro as perspectivas para o fortalecimento tecnológico e eficiência dos países ao redor do mundo. O AIPI é relevante para a discussão, pois fornece dados significativos para compreender como os países estão investindo e se adaptando às tecnologias emergentes. Ao emparelhar diversos países, é possível observar as discrepâncias nos investimentos entre economias desenvolvidas e emergentes, permitindo identificar brechas de desenvolvimento e avanços no segmento da inteligência artificial. Os dados para compor o índice foram coletados e compilados por 8 instituições internacionais<sup>9</sup>, dessa forma, o índice AIPI serve como um estudo norteador para políticas públicas e pesquisas científicas, no qual a intenção primordial é identificar oportunidades de melhorias e desenvolvimento entre países, afastando-se de uma pesquisa classificatória entre os países.

A pesquisa foi estruturada a partir de um conjunto de indicadores macroestruturais, divididos em: 1. Preparação Fundamental para IA e 2. Segunda Etapa de Preparação para IA. Estas duas divisões contemplam a infraestrutura digital, o capital humano e as políticas de mercado de trabalho, a inovação e a integração econômica, além de regulamentação e ética dos países.

Quadro 1 - Indicadores Macroestruturais AIPI *index*

<b>Preparação Fundamental para IA (<i>Foundational AI Preparedness</i>)</b>	<b>Segunda Etapa de Preparação para IA (<i>Second-Generation AI Preparedness</i>)</b>
Infraestrutura Digital	Inovação e Integração Econômica
Capital Humano e Políticas de Mercado de Trabalho	Regulamentação e Ética

Fonte: Elaborado pelo autor.

<sup>9</sup> Fraser Institute, International Labor Organization, International Telecommunication Union, United Nations, United Nations Conference on Trade and Development, Universal Postal Union, World Bank, and World Economic Forum.

A composição do Índice de Prontidão para IA (AIPI) é a média simples das quatro principais dimensões: infraestrutura digital, capital humano, inovação tecnológica e estruturas legais, na qual cada dimensão, por sua vez, é calculada a partir da média simples de um conjunto de sub indicadores. Os dados dos sub indicadores foram extraídos a partir de três tipos de fontes: pesquisas oficiais, pesquisas com dados concretos e pesquisas de percepções. O índice utiliza a escala de 0 a 1, na qual os valores mais altos indicam cenários mais favoráveis para a adoção de sistemas de IA. A pesquisa realizada em 2023 abrangeu um total de 174 países, permitindo uma visão global e, ao mesmo tempo, um panorama dos desafios e discrepâncias entre economias desenvolvidas e economias emergentes.

A análise a seguir concentra-se exclusivamente no desempenho e nas classificações dos países lusófonos que estão incluídos<sup>10</sup> no AIPI, sendo, portanto, os seguintes países: Angola, Brasil, Cabo Verde, Guiné-Bissau, Portugal, Moçambique, São Tomé e Príncipe e Timor-Leste. A realização de uma análise comparativa dos países de língua portuguesa na adoção e implementação de tecnologias de inteligência artificial se mostra relevante para reconhecer padrões e identificar semelhanças e diferenças dentro desse grupo linguístico, considerando suas particularidades geográficas, econômicas e culturais. Ainda, permite destacar pontos comuns e divergentes na preparação e aplicação de sistemas de IA, além de explorar oportunidades de cooperação e intercâmbio de conhecimento dentro da comunidade lusófona.

### 3.2 A COMUNIDADE LUSÓFONA E A PRONTIDÃO PARA IA

Para mensurar quais são os reais impactos produtivos e econômicos da inteligência artificial, bem como compreender quais são estratégias necessárias para que os países adotem essas ferramentas em larga escala, o Fundo Monetário Internacional (FMI) desenvolveu o Índice de Prontidão para IA (AIPI)<sup>11</sup>. Este instrumento de análise visa avaliar o grau de integração dos países para aplicação de sistemas baseados em Inteligência Artificial, ainda que de maneira sutil, além de identificar padrões distintivos entre economias desenvolvidas e

---

<sup>10</sup> O país Guiné Equatorial (GNQ), que também adota a Língua Portuguesa como oficial, não foi incluído no estudo, visto que não foi contemplado no índice de preparação do FMI.

<sup>11</sup> Disponível em: <<https://www.imf.org/external/datamapper/datasets/AIPI>>. Acesso em: 27 out. 2024.

emergentes no processo de adoção dessas novas tecnologias. A relevância do AIPI se dá pelo estágio de escalonamento mundial de ferramentas de aprendizado de máquina, aprendizado profundo e redes neurais, servindo como balizador tanto para políticas públicas quanto para estratégias de negócios em transformação digital.

O Índice de Prontidão para Inteligência Artificial (AIPI) desempenha um papel relevante ao revelar a correlação entre o preparo tecnológico dos países e sua capacidade de promover a diversidade, incluindo a diversidade linguística. Países que estão mais bem colocados no índice, como Estados Unidos, Singapura, Reino Unido e países pertencentes da União Europeia, geralmente possuem maior infraestrutura tecnológica, investimentos em pesquisa e políticas inclusivas, permitindo um desenvolvimento mais abrangente de sistemas que considerem as particularidades culturais e linguísticas. Um exemplo de como os investimentos direcionados para tecnologia e pesquisa com foco em IA podem impactar a diversidade linguística é o projeto EuroLingua-GPT, financiado pela União Europeia e desenvolvido pelo centro nacional de IA aplicada na Suécia (AI SWEDEN) e a associação alemã Fraunhofer-Gesellschaft, que possui como objetivo principal treinar um novo grande modelo de linguagem para todas as línguas oficiais da União Europeia.

Essa prontidão tecnológica facilita a criação de modelos de IA mais diversificados e adaptados, enquanto nações com menor índice enfrentam desafios para garantir que suas línguas e culturas sejam adequadamente representadas e respeitadas, evidenciando um desequilíbrio global que pode acentuar a desvalorização de uma determinada língua em relação a outras línguas no contexto digital. Portanto, as inovações no campo da IA têm desencadeado transformações em cascata em diversas áreas, especialmente na comunicação e na linguagem, através da disseminação e produção de informações no âmbito digital, personalização de conteúdos para os usuários e também na automação na produção de textos.

Essas mudanças geram impactos significativos em comunidades linguísticas, em particular no que se refere à preservação da diversidade linguística. Como um bom ponto de partida, é importante atentar para o índice geral dos países lusófonos, o qual representa uma média dos quatro macros indicadores principais da pesquisa realizada pelo FMI. No índice geral de prontidão para a IA, os países lusófonos são classificados conforme apresentado na Tabela 1, com destaque para Portugal e Brasil, respectivamente. Já no gráfico 1, é possível visualizar um panorama dos países lusófonos em relação à média do próprio grupo. Essa

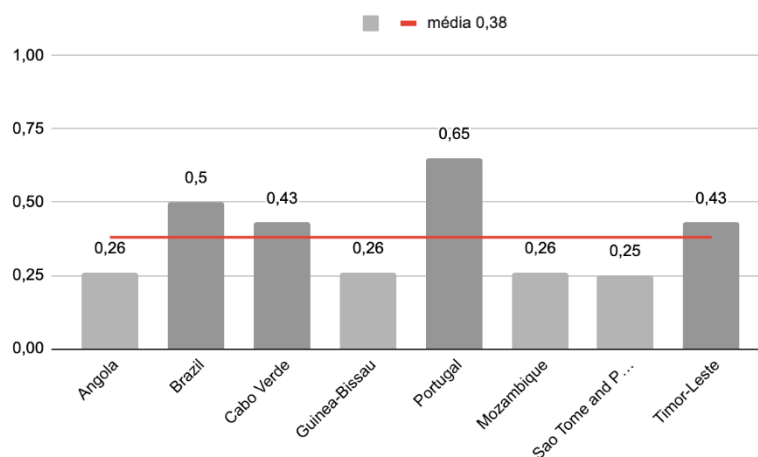
comparação possui como objetivo identificar países que possuem um ecossistema mais avançado para a implementação de sistemas de inteligência artificial, bem como países com potencial para avanços no setor.

Tabela 1 - Índice AIPI dos países lusófonos

País	ISO3	AIPI Index
Angola	AGO	0,26
Brasil	BRA	0,50
Cabo Verde	CPV	0,43
Guiné-Bissau	GNB	0,26
Portugal	PRT	0,65
Moçambique	MOZ	0,26
São Tomé e Príncipe	STP	0,25
Timor-Leste	TLS	0,43

Fonte: Elaborado pelo autor

Gráfico 1 - Média de grupo países lusófonos



Fonte: Elaborado pelo autor

Ao observar o Gráfico 1, é possível notar algumas disparidades entre os países da comunidade de língua portuguesa. Despontam países como Portugal, Brasil, Cabo Verde e Timor-Leste, os quais apresentam uma vantagem competitiva em relação aos demais países do grupo. Em contrapartida, Angola, Guiné-Bissau e Moçambique demonstram um potencial considerável para desenvolvimento e avanços no setor. Observa-se, ainda, um fracionamento claro dos países relação à média do grupo: enquanto metade dos países analisados se encontra acima dessa média, a outra metade apresenta índices inferiores, com São Tomé e Príncipe aproximadamente 34% abaixo da média do grupo, e Angola, Guiné-Bissau e Moçambique cerca de 31% abaixo. Essa disparidade reflete as heterogeneidades socioeconômicas, infraestruturais e de capital humano existentes entre os países lusófonos e compreender esses fatores, é determinante para o desenvolvimento de estratégias equilibradas e benéficas para promover a inclusão digital e o fortalecimento do ecossistema de inovação.

É importante contextualizar que a média global de todos os participantes do índice AIPI é de 0,47. No entanto, utilizar este valor como parâmetro comparativo não é o objetivo principal. Isso porque a média global inclui países criadores de tecnologias e os países líderes em inovação, pertencentes ao grupo de economias desenvolvidas, cuja realidade e capacidades diferem substancialmente das demais nações, especialmente daquelas em desenvolvimento. Desse modo, uma comparação mais significativa deve considerar principalmente os padrões e disparidades observados entre os países lusófonos, que compartilham características socioeconômicas, infraestruturais e de capital humano mais próximas.

Ao enquadrar a análise no âmbito dos países lusófonos, é relevante posicioná-los em relação ao grupo de economias emergentes, no qual totalizam 78 nações<sup>12</sup> conforme o Fundo Monetário Internacional. Esta abordagem fornece uma base de comparação mais adequada e permite uma avaliação mais precisa do progresso relativo desses países em inteligência artificial. Nesse sentido, o grupo de economias emergentes consta como 0,46 - ou seja, 17%

---

<sup>12</sup> Na pesquisa promovida pelo FMI os países com economias emergentes são: Albânia, Argélia, Angola, Argentina, Armênia, Azerbaijão, Bahamas, Bahrein, Barbados, Belarus, Belize, Bolívia, Bósnia e Herzegovina, Botsuana, Brasil, Brunei Darussalam, Bulgária, Cabo Verde, Chile, China, República Popular da, Colômbia, Costa Rica, República Dominicana, Equador, Egito, El Salvador, Eswatini, Fiji, Gabão, Geórgia, Guatemala, Guiana, Hungria, Índia, Indonésia, Irã, Iraque, Jamaica, Jordânia, Cazaquistão, Kuwait, Líbano, Líbia, Malásia, Maldivas, Maurício, México, Mongólia, Montenegro, Marrocos, Namíbia, Macedônia do Norte, Omã, Paquistão, Panamá, Paraguai, Peru, Filipinas, Polônia, Catar, Romênia, Federação Russa, Santa Lúcia, São Vicente e Granadinas, Arábia Saudita, Sérvia, Seychelles, África do Sul, Sri Lanka, Suriname, Síria, Tailândia, Trinidad e Tobago, Tunísia, Turkiye, República da, Ucrânia, Emirados Árabes Unidos, Uruguai, Venezuela.

acima da média observada para o conjunto de países lusófonos. Essa diferença entre os dois grupos aponta que, embora os países de língua portuguesa apresentem avanços consideráveis em determinadas áreas, ainda há espaço para desenvolvimento e melhoria no setor de tecnologia baseada em IA, quando comparados ao desempenho médio das demais economias emergentes.

Agora, ao observar os dados sob uma ótica diferente, a predominância de uma língua no contexto global pode exercer implicações diretas no desenvolvimento tecnológico. Segundo o ranking elaborado pelo *Ethnologue*, as cinco línguas mais faladas no mundo são respectivamente: Inglês (1,5 B), Chinês Mandarim (1,1 B), Hindi (608,8 M), Espanhol (559,5 M) e Árabe Padrão (332,5 M). A língua portuguesa ocupa a oitava posição nessa lista, com aproximadamente 263,8 milhões de falantes.

Tabela 2 - Ranking das 10 línguas mais faladas no mundo.

Posição	Língua	Número de falantes
1	Inglês	1,5 bilhão
2	Chinês Mandarim	1,1 bilhão
3	Hindi	608,8 milhões
4	Espanhol	559,5 milhões
5	Árabe Padrão	332,5 milhões
6	Francês	311,6 milhões
7	Bengali	278,2 milhões
8	Português	263,8 milhões

Fonte: Ethnologue 200. **Ethnologue**, 2024<sup>13</sup>.

Dessa forma, o índice AIPI do Fundo Monetário Internacional é um importante indicador para situar a comunidade lusófona no cenário tecnológico mundial, visto que a língua portuguesa é a oitava língua mais falada do mundo. Ao avaliar o progresso, os desafios e as lacunas no nível de prontidão para IA, o índice AIPI permite posicionar os países

<sup>13</sup> Disponível em: <<https://www.ethnologue.com/insights/ethnologue200/>>. Acesso em: 27 out. 2024.

lusófonos em relação ao avanço tecnológico global. Outro fator determinante, é que o índice fornece dados e percepções que ajudam a direcionar investimentos, políticas públicas e colaborações estratégicas em IA para a comunidade lusófona, fornecendo um ponto de referência, dado o desempenho de países líderes em investimento, para a construção de um ecossistema tecnológico que respeite diversidade linguística e cultural da língua portuguesa.

#### **4 GOVERNANÇA E LÍNGUA: A PRESENÇA DO PORTUGUÊS NA ERA DIGITAL**

Para que um grande modelo de linguagem tenha um bom desempenho em uma língua destino, é necessário que existam múltiplas camadas computacionais até o retorno ao usuário final. Nessa arquitetura computacional complexa, poucas empresas possuem capacidade técnica necessária de infraestrutura para que os LLMs alcancem um resultado satisfatório. Logo, as grandes corporações ocupam o topo da pirâmide tecnológica, sendo detentoras dos códigos-fonte e mantenedoras de todo o ecossistema tecnológico global, dentre elas Alphabet, OpenAI, Microsoft, entre outras. É nesse cenário de dominação tecnológica que surgem movimentos, ainda que críticos, que apontam para uma nova fase do capitalismo, denominada tecno-feudalismo. O conceito de tecno-feudalismo descreve uma transformação nas estruturas de domínio do poder, em que a tradicional dominação industrial cede lugar a uma dominação psicocomportamental, que também pode ser compreendida como uma dinâmica de controle algorítmico, conforme mencionado por Marta Peirano no livro *O Inimigo Conhece o Sistema* (2022).

É na interação entre homem e sistemas digitais que os algoritmos desempenham um elo, direcionando conteúdos e influenciando comportamentos de acordo com interesses comerciais e ideológicos específicos, que, como resultado, acaba moldando padrões de pensamento e ações no convívio social. Segundo a entrevista realizada pelo jornal *El País* (VEGA, 2023) com o escritor Yanis Varoufakis, o mundo vivencia uma nova ordem econômica, em que o valor econômico não é mais extraído pela produção de bens materiais, mas sim pelo controle do comportamento das pessoas através de tecnologias digitais. Como resultado, é possível presumir que o poder econômico está nas mãos das grandes corporações de tecnologia.

Essa nova configuração de poder exerce influência sobre a diversidade linguística e cultural em um âmbito global, uma vez que moldam inclusive as dinâmicas de preservação e expansão cultural, conforme interesses pré-determinados. As *bigtechs*, ao controlarem as infraestruturas básicas para o funcionamento de um *LLMs*, por exemplo, condicionam, conforme seus interesses, quais línguas e culturas recebem prioridade na esteira de desenvolvimento. A partir disso, é possível balizar discussões sobre equidade digital, democratização do conhecimento e preservação da diversidade linguística, uma vez que o desenvolvimento tecnológico atual tende a privilegiar certos idiomas e perspectivas culturais em detrimento de outros.

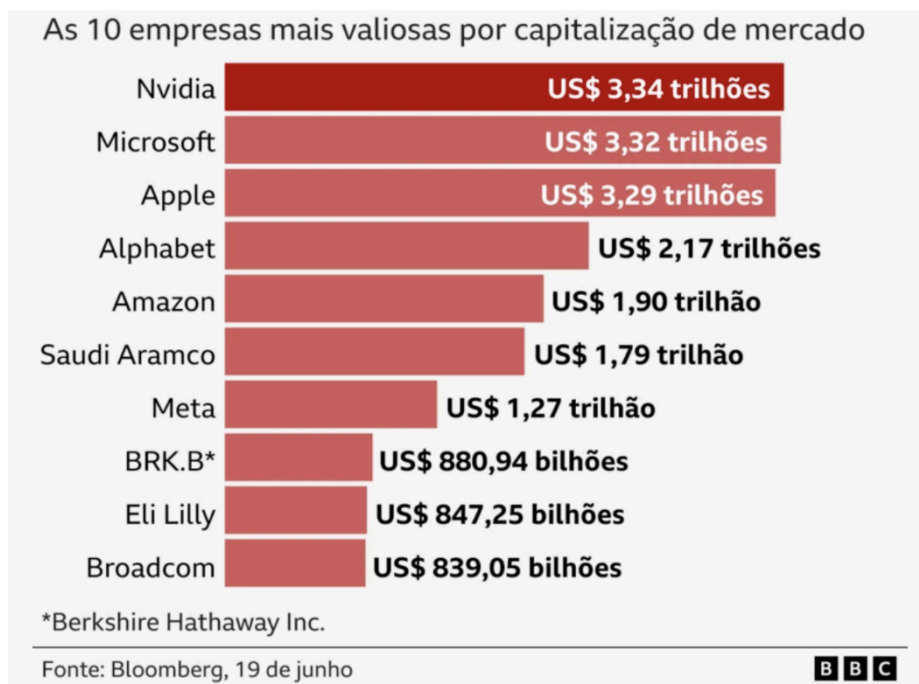
#### 4.1 DESAFIOS DE INFRAESTRUTURA NO BRASIL

A profundidade da temática vai além do desempenho do produto final (*chatbot*) em uma língua destino, é estratégico analisar a infraestrutura subjacente às tecnologias de inteligência artificial e os recursos necessários para seu pleno funcionamento. A implementação e disponibilização desses modelos computacionais aos usuários finais requer uma infraestrutura tecnológica de alta complexidade, fundamentada principalmente em *datacenters* (centros de processamento de dados). Essas instalações (*datacenters*) são o alicerce do processamento computacional, abrigando sistemas de servidores robustos responsáveis pelo armazenamento, processamento e gerenciamento de grande volume de dados. A arquitetura desses centros de processamento de dados demanda uma infraestrutura física especializada e um conjunto de outros componentes que incluem, por exemplo, sistemas de refrigeração, eficiência energética e redes de alta velocidade, para assim, garantir um desempenho adequado desses sistemas. Todos os componentes que sustentam um *datacenter* suscitam debates sobre os mais diversos desafios e que extrapolam a esfera tecnológica, visto que incluem questões relacionadas à crise climática e à necessidade de fontes de energia renováveis, em virtude do elevado consumo energético.

Já no contexto de infraestrutura computacional necessária para o processamento de modelos de IA, outro ponto de destaque são os *hardwares*, em especial as unidades de processamento gráfico (GPUs). As GPUs são componentes de alto desempenho

computacional, sendo que, atualmente, os modelos mais avançados do mercado são produzidos pela empresa NVIDIA.

Figura 5 - Valorização comercial da NVIDIA nos últimos 4 anos



Fonte: A empresa que ultrapassou Apple e Microsoft e se tornou a mais valiosa do mundo. **BBC**, 2024.

Esses dispositivos de processamento se destacam pela alta capacidade técnica, porém enfrentam barreiras de acesso devido aos elevados custos e à disponibilidade limitada. A aquisição dessas tecnologias geralmente exige negociações específicas com os fabricantes, o que restringe o acesso e afeta a democratização da inteligência artificial, especialmente em contextos com recursos financeiros limitados e desafios comerciais internacionais. No caso do Brasil, além do difícil acesso à matéria-prima para treinamento de modelo de linguagem, essa infraestrutura de base também enfrenta outros obstáculos, que incluem elementos como: serviços de computação em nuvem e alta conectividade de internet.

Existem diversos *datacenters* dedicados à inteligência artificial ao redor do mundo, controlados pelas maiores corporações de tecnologia. No contexto brasileiro, embora existam *datacenters* de grande porte, ainda há uma carência de investimentos específicos voltados para o desenvolvimento dessa infraestrutura primária. Além disso, outro fator que não

favorece o cenário brasileiro em IA são os trâmites fiscais: o arcabouço fiscal que regula este setor no país ainda é uma barreira tanto para o avanço interno quanto para a aproximação de empresas estrangeiras interessadas em investir no país (RAMOS, 2023). Por outro ponto de vista, o Brasil se beneficia de uma posição estratégica, devido à abundância de fontes de energia renovável, o que é um ponto de virada significativo, considerando o alto consumo de energia necessário para manter essas instalações.

## 4.2 GOVERNANÇA DA INFORMAÇÃO

Em um segundo nível da hierarquia tecnológica, destaque-se a necessidade de conjuntos de dados qualificados para o treinamento de grandes modelos de linguagem, sendo indispensáveis para o desenvolvimento desses sistemas. Embora os *LLMs* operem com volumes exorbitantes de dados, esses sistemas não se limitam a replicar os conteúdos de treinamento, mas sim aprendem os padrões subjacentes contidos nesse conjunto de dados. Como os grandes modelos de linguagem são, em essência, modelos matemáticos complexos, eles interpretam as informações e geram novos padrões com base nos dados assimilados. O volume de dados exigidos para o treinamento desses sistemas é imenso, estabelecendo uma correlação direta entre a quantidade de dados processados e o resultado devolvido por um *chatbot*.

A criação e o desenvolvimento de *datasets*<sup>14</sup> de alta qualidade representam um investimento significativo, dada a complexidade dos processos de refinamento e extração de conteúdo. Essa dependência de *datasets* extensos e qualificados é um fator determinante no processo de desenvolvimento de modelos de linguagem, evidenciando a necessidade de uma infraestrutura robusta para o processamento, coleta, curadoria e manutenção desses diversos repositórios de informação.

A predominância econômica das *bigtechs* se manifesta aqui de maneira clara, através da aquisição de dados legalmente, ilicitamente ou através da formalização de acordos comerciais com grandes grupos jornalísticos, para usufruir dos seus conteúdos para treinamento de seus modelos. Portanto, nesse caso, as empresas que possuem um grande

---

<sup>14</sup> *Dataset* é um conjunto estruturado de dados que inclui informações como conteúdos, rótulos e metadados usados para treinar modelos de IA.

poder de negociação levam vantagem competitiva dentro do setor e garantem acesso privilegiado a corporas textuais exclusivos para o treinamento de seus sistemas.

O efeito dessa busca incessante por dados faz ainda mais sentido, visto que há um escopo abrangente de línguas no mundo que não possuem um grande volume de dados qualificados para treinamento dos modelos, e como efeito, línguas com abundância de fontes para treinamento, como é caso da língua inglesa, possuem desempenho superior. Portanto, ferramentas de IA são, na sua maioria, modelos de linguagem *english-centric* (KEW, T.; SCHOTTMANN, F.; SENNRICH, R., 2023).

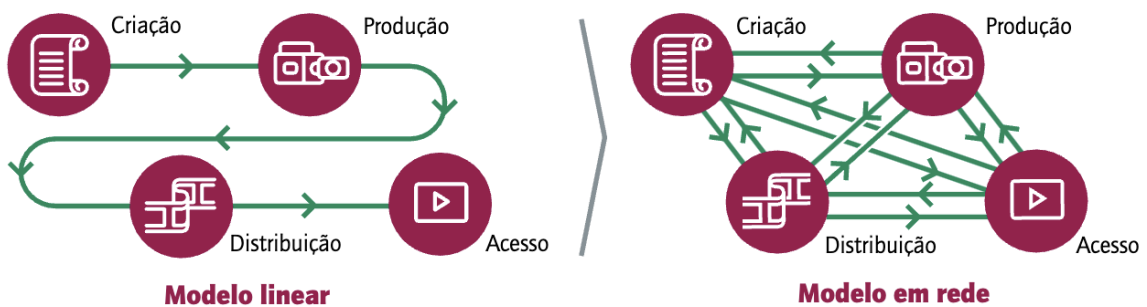
É nesse enquadramento de hegemonia linguística que o conceito de cadeia de valor cultural, proposto na convenção da diversidade em 2005 pela UNESCO, se faz relevante, já que tenta compreender a forma de consumo digital e suas implicações. No modelo analógico de distribuição cultural, a atribuição de papéis seguia uma lógica linear ao agregar valor a um produto ou serviço. Nesse modelo de distribuição, cada ator era responsável por atribuir valor de forma sequencial: criação, produção, distribuição, acesso e participação.

Contudo, o que ocorre atualmente é que a nova cadeia criativa é, portanto, uma rede interconectada (UNESCO, 2005). O modelo digital permite que as etapas ocorram simultaneamente, ou seja, as etapas de criação, produção, distribuição, disseminação, participação e fruição de produtos e serviços ocorrem ao mesmo tempo.

Portanto, segundo o relatório global da convenção de 2005 da UNESCO, as tecnologias digitais mudaram a forma de consumo digital, através da descentralização de processos. O que era sequencial agora passou a ser em rede, refletindo assim o ambiente colaborativo dos espaços digitais. Nesse sentido, o uso de sistemas de IA também engloba essa rede de consumo digital.

Figura 6 - Cadeia de valor cultural

## AS TECNOLOGIAS DIGITAIS TRANSFORMARAM A CADEIA DE VALOR CULTURAL



Fonte: Relatório Global da Conferência de 2005, UNESCO.

Essa nova configuração impõe adversidades em termos de gerenciamento e regulação dos fluxos de informação, considerando especialmente as assimetrias socioeconômicas e tecnológicas existentes entre o norte e o sul global. A fragilidade de uma governança de dados acarreta consequências negativas para a produção, circulação e acesso aos bens e serviços culturais nas regiões menos desenvolvidas tecnologicamente, aprofundando ainda mais as disparidades globais. Consequentemente, modelos de linguagem que possuem como base a língua inglesa podem afetar a diversidade linguística e cultural de outras línguas no âmbito digital.

Deste modo, os dados possuem um valor estratégico na economia digital atualmente. Este raciocínio se torna ainda mais pertinente ao considerar as complexas questões éticas, legais e socioeconômicas envolvidas na exploração e utilização desses recursos informacionais. Um exemplo de aplicação ética envolvida é o caso de uma grande corporação que foi multada e imposta a restrições operacionais no Brasil devido ao uso não autorizado de dados pessoais, incluindo informações sensíveis de menores. Este episódio ilustra os desafios éticos e jurídicos relacionados à gestão e ao uso de dados no desenvolvimento de sistemas de inteligência artificial, evidenciando a necessidade de mecanismos para a governança da informação.

### 4.3 TREINAMENTO DE GRANDES MODELOS DE LINGUAGEM

É conhecido que para treinar um grande modelo de linguagem é necessário um conjunto vasto de dados. Para dimensionar o tamanho, estima-se que o GPT-3 em sua configuração mais básica utilizou aproximadamente 300 bilhões de tokens em seu processo de treinamento, enquanto modelos mais recentes, como o Llama 3 da empresa Meta, processou volumes na ordem de 15 trilhões de tokens.<sup>15</sup> De acordo com documentos da Google, para a ferramenta de IA da Google, Gemini, 100 tokens equivalem de 60 a 80 palavras em inglês<sup>16</sup>. O aumento do consumo de dados demonstrado a partir de 2020, ano de lançamento do GPT-3, evidencia a necessidade de um volume exponencial para o treinamento efetivo de grandes modelos de linguagem. Portanto, dados qualificados para treinamento de modelos são um combustível principal por trás de um rendimento satisfatório de um *chatbot*. Atento à dessa demanda, em uma entrevista para a ONU em 2023, Heber Maia, diretor de Tecnologia da Informação do Ministério do Trabalho do Brasil, alerta sobre a carência de dados em língua portuguesa:

Hoje nós estamos perdendo profissionais, especialistas extremamente competentes, que trabalham com inteligência artificial, mas quando vão desenvolver suas pesquisas, eles precisam trabalhar com conjuntos de dados em língua inglesa. Não temos, por exemplo, conjuntos de dados suficientes para desenvolvimento de determinadas pesquisas de inteligência artificial porque nós não nos articulamos para isso. Estamos cientes e convencidos de que precisamos unificar os países lusófonos para que essas pesquisas em inteligência artificial sejam realizadas na nossa língua (MAIA, H. 2023).

A composição das bases de dados utilizadas no treinamento desses modelos é constituída predominantemente de conteúdos disponíveis na internet. Conforme documentado no artigo *Language Models are Few-Shot Learners* (BROWN, T. B. et al., 2017), a distribuição das fontes de dados apresenta uma concentração volumosa em repositórios específicos: aproximadamente 60% provêm do Common Crawl, uma organização sem fins lucrativos dedicada ao rastreamento e disponibilização de dados públicos da internet, enquanto 22% originam-se do WebText2. O percentual remanescente é contemplado a partir de fontes diversificadas, incluindo acervos bibliográficos e conteúdos enciclopédicos provenientes da Wikipedia.

---

<sup>15</sup> Disponível em: <

<https://about.fb.com/br/news/2024/04/apresentando-meta-llama-3-o-grande-modelo-de-linguagem-de-codigo-aberto-mais-capaz-ate-hoje/>>. Acesso em: 28 out. 2024.

<sup>16</sup> Disponível em: < <https://ai.google.dev/gemini-api/docs/tokens?hl=pt-br&lang=python> >. Acesso em: 28 out. 2024.

Tabela 3 - Distribuição

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Fonte: BROWN, T. B. et al. **Language Models are Few-Shot Learners**, 2020.

#### 4.4 A PRESENÇA DA LÍNGUA PORTUGUESA NA INTERNET

A análise da estrutura de treinamento dos modelos de linguagem demonstra uma dependência significativa de dados disponíveis na internet para o treinamento de seus algoritmos. Neste contexto, torna-se estratégico examinar a distribuição linguística dos conteúdos na internet através de métricas específicas que mensuram a representatividade das diferentes línguas no espaço digital global.

Atualmente, a organização sem fins lucrativos OBDILCI (Observatório da Diversidade Linguística e Cultural na Internet) possui um papel fundamental para o mapeamento e monitoramento das línguas no espaço digital, servindo como balizador para pesquisas na área. Um dos projetos do OBDILCI contempla o acompanhamento dos indicadores da presença da linguagem na internet e é a partir desses indicadores que é possível ter uma dimensão da distribuição linguística na internet de forma crítica e organizada.

Um dos indicadores significativos para o acompanhamento das línguas é o índice de ciberglobalização (OBDILCI, 2024). Este indicador avalia quantitativamente a presença e relevância de uma língua no ambiente digital. Porém, este parâmetro é estratégico ao ser analisado em conjunto com dois outros indicadores: a porcentagem de conteúdo disponível e a presença digital da língua. O primeiro mensura o volume total de informações online em determinada língua, enquanto o segundo estabelece uma comparação entre a quantidade de conteúdo web disponível em uma língua e a participação global de seus falantes.



Tabela 4 - Indicadores da Presença de Línguas na Internet

Língua (ISO)	Línguas	% de Conteúdo Disponível	Presença Digital
eng	Inglês	20,42%	1,45
zho	Chinês	18,88%	1,30
spa	Espanhol	7,70%	1,48
hin	Hindi	3,82%	0,67
rus	Russo	3,73%	1,57
ara	Árabe	3,65%	0,89
fra	Francês	3,41%	1,18
por	Português	3,09%	1,25
jpn	Japonês	2,20%	1,91
deu	Alemão	2,15%	1,72

5

Fonte: Indicadores da Presença de Línguas na Internet. **Observatório da Diversidade Linguística e Cultural na Internet, 2024.**

A partir da tabela 4, é possível visualizar a distribuição linguística em contexto digital, no qual a língua inglesa possui uma proporção de conteúdo consideravelmente maior que a língua portuguesa. Por outro lado, a presença digital da língua portuguesa se mostra competitiva ao comparar com línguas que possuem quase o dobro de conteúdo disponível, como o espanhol.

Porém, ao parear o índice de ciberglobalização entre a língua inglesa e a língua portuguesa é possível notar uma força digital superior da língua inglesa, conforme demonstrado na tabela 5:

Tabela 5 - Índice de ciberglobalização

ISO	Língua	CGI	CGI%
eng	Inglês	1.99	14.13%
por	Português	0.20	1.39%

Fonte: Indicadores da Presença de Línguas na Internet. **Observatório da Diversidade Linguística e Cultural na Internet**, 2024.<sup>17</sup>

Embora o inglês tenha uma posição dominante em termos de conteúdo disponível na internet, outras línguas também mantêm uma presença relevante no ambiente digital. Esta distribuição, mesmo que desigual, mostra que existe um volume considerável de conteúdo em diferentes idiomas e que refletem a diversidade linguística presente no mundo virtual. Esta variedade de línguas no ciberespaço, apesar de suas diferenças em quantidade de conteúdo, é necessária para a representatividade dos idiomas na internet.

Porém, a representatividade digital de línguas não parece estar refletida no treinamento de modelos de linguagem, dado que boa parte das ferramentas de IA são treinadas majoritariamente em inglês, não atendendo adequadamente a demanda global por uma diversidade linguística e cultural. Essa realidade levanta discussões importantes sobre a influência geopolítica que envolve o uso da linguagem, evidenciando como estruturas de poder e colonização ainda afetam as configurações linguísticas do cenário global atual.

## 5.1 LÍNGUA E PODER

O linguista Louis-Jean Calvet, em suas investigações sobre as relações entre língua e poder, analisou os mecanismos pelos quais as dinâmicas de poder linguístico se estabelecem e se perpetuam, denunciando como certas línguas exercem hegemonia sobre outras. Para Calvet

---

<sup>17</sup> Disponível em: <<https://www.obdilci.org/projects/main/>>. Acesso em: 30 out. 2024.

(2002), as línguas estão inseridas em um processo de imposição linguística, no qual uma língua se sobressai sobre outras em contextos de poder político, econômico e cultural, configurando uma forma de dominação linguística. Essa perspectiva permite entender as dinâmicas glotopolíticas, nas quais as decisões em torno do uso de uma língua influenciam a sobrevivência ou extinção de outras, com as línguas dominantes ameaçando ou exterminando as línguas minoritárias.

Portanto, é necessário que exista uma gestão linguística, que segundo Calvet (2007), podem ocorrer *in vivo* e *in vitro*. A gestão *in vivo* é quando as modificações na língua ocorrem cotidianamente durante a interação dos falantes, sem uma imposição externa. Já a gestão *in vitro* se dá por meio de intervenções planejadas na língua através dos governos. Sendo assim, a gestão *in vivo* e *in vitro* tendem a ser conflituosas devido a sua natureza, logo, o planejamento linguístico possui uma atribuição importante para que exista uma aproximação entre a gestão *in vivo* e *in vitro*:

Os instrumentos de planejamento linguístico aparecem, portanto, como uma tentativa de adaptação e de utilização *in vitro* de fenômenos que sempre se manifestaram *in vivo*. E a política linguística vê-se então diante, ao mesmo tempo, dos problemas de coerência entre os objetivos do poder e as soluções intuitivas que são frequentemente postas em prática pelo povo, bem como do problema de certo controle democrático, a fim de não deixar os "decisores" fazerem o que bem entendem (CALVET, 2007, p.71).

O planejamento linguístico é, em síntese, a implementação prática das políticas linguísticas, que podem ocorrer através de instrumentos como planejamento linguístico interno e planejamento linguístico externo. O primeiro ocupa-se de questões estruturais da língua, já o segundo, organiza as questões sociopolíticas da língua. Ou seja, o planejamento linguístico constitui *in vitro* uma espécie de réplica dos fenômenos produzidos continuamente *in vivo* (CALVET, 2007, p.85).

Desse modo, cabe aqui mencionar o esforço constante da indústria das línguas (CALVET, 2007) em posicionar-se para evitar um domínio tecnológico. Um exemplo foram os investimentos da França para garantir a presença da língua francesa em produtos de informática, com o objetivo de barrar os empréstimos linguísticos do inglês. Como resultado, a França criou um vocabulário em francês para ferramentas digitais, afastando o domínio linguístico do inglês. O conceito de indústria das línguas mencionado por Calvet pode ser facilmente aplicado atualmente:

No início dos anos 1980, surgiu a expressão "indústrias da língua" para designar o conjunto das novas tecnologias de informação, um cruzamento de informática, inteligência artificial, ciências cognitivas e lingüística. Trata-se, ou deveria tratar-se, da produção de objetos (dicionários eletrônicos, corretores ortográficos, softwares de processamento de textos, de tradução automática, bases de dados, bancos de conhecimentos, etc.) e produtos linguísticos (neologia, terminologia...) no quadro de uma pesquisa de ponta de caráter multidisciplinar (CALVET, 2007, p.100).

Ainda dentro dessa temática de indústria das línguas, alguns países começaram a perceber os desafios sociais que os grandes modelos de linguagem suscitam: o predomínio cultural por meio da língua. Do ponto de vista linguístico, muitas línguas enfrentam dois problemas principais. Primeiro, não possuem dados suficientes para treinar adequadamente os sistemas de IA. Além disso, as manifestações orais - como narrativas, causos e tradições faladas - têm pouca representatividade nesses treinamentos, o que prejudica especialmente as línguas com forte tradição oral, como as indígenas e africanas.

Do ponto de vista cultural, o panorama também é preocupante, porque a língua vai muito além da comunicação: ela carrega a identidade, os valores e a visão de mundo de um povo. Segundo a hipótese de Sapir-Whorf a língua que uma pessoa fala impacta sua percepção de mundo. A partir dessa concepção surge o conceito de relativismo linguístico (versão fraca da hipótese), que aponta como as estruturas linguísticas podem influenciar o pensamento de uma falante. Portanto, o relativismo linguístico propõe que diferentes línguas moldam distintas formas de compreender e interpretar a realidade. Esta teoria argumenta que as categorias gramaticais, vocabulário e estruturas específicas de uma língua condicionam, de alguma maneira, o modo como seus falantes processam diferentes vivências. Essa proposta teórica foi revisitada no livro *Textos Base de Linguagem*, escrito por MARCONDES (2009):

Segundo a hipótese Sapir-Whorf, a língua de uma determinada comunidade organiza sua cultura, sua visão de mundo, pois uma comunidade vê e compreende a realidade que a cerca através das categorias gramaticais e semânticas de sua língua. Há portanto uma interdependência entre linguagem e cultura. Um povo vê a realidade através das categorias de sua língua, mas sua língua se constitui com base em sua forma de vida (MARCONDES, 2009, p.64).

Portanto, ao conectar a ideia de que uma língua carrega a cultura e a visão mundo de determinada comunidade linguística, conseqüentemente, no momento em que os modelos de inteligência artificial não conseguem representar adequadamente uma língua, eles também falham em capturar a diversidade cultural das comunidades que a falam. Isso pode resultar em

uma homogeneização das expressões culturais e um reforço da dominação cultural do norte global.

Apesar de todas as implicações, as grandes corporações deixam claro que seus modelos de linguagem são treinados quase exclusivamente em inglês, como a META, que no lançamento do modelo Llama-3 afirma:

Para treinar o melhor modelo de linguagem, a curadoria de um conjunto de dados de treinamento grande e de alta qualidade é fundamental. Em linha com os nossos princípios de design, investimos pesadamente em dados de pré-treinamento. O Llama 3 é pré-treinado em mais de 15T de tokens, todos coletados de fontes disponíveis publicamente. Nosso conjunto de dados de treinamento é sete vezes maior que o usado para o Llama 2 e inclui quatro vezes mais código. Para estarmos preparados para os próximos casos de uso multilíngue, mais de 5% do conjunto de dados de pré-treinamento do Llama 3 consistem em dados de alta qualidade em idiomas diferentes do inglês, abrangendo mais de 30 idiomas. Contudo, não esperamos o mesmo nível de desempenho nestas línguas do que em inglês. (FACEBOOK COMPANY, 2024).

Quando os *LLMs* são predominantemente treinados em determinada língua, sem uma atenção às demais línguas, há o risco de reforçar uma hegemonia linguística. No caso do português brasileiro, isso acarreta o apagamento de algumas formas de uso, por exemplo, os regionalismos e particularidades linguísticas, enfraquecendo a presença da multiplicidade linguística na esfera digital. Portanto, para evitar esse cenário, é necessário haver um planejamento linguístico durante a concepção dessas tecnologias, de modo que os impactos e prejuízos linguísticos sejam considerados. Sendo assim, o planejamento linguístico é um possível caminho para garantir uma representação justa e equitativa para as diferentes línguas, promovendo o fortalecimento e aumentando o alcance no contexto digital.

## **6 DIVERSIDADE LINGUÍSTICA**

As línguas sub-representadas em *datasets* de treinamento possuem visíveis desvantagens, visto que o desempenho tende a ser inferior e, como consequência prática, é possível observar algoritmos adaptados quase que exclusivamente para um contexto monolíngue. No GitHub, que é uma plataforma de hospedagem de código mundialmente conhecida, alguns dados de treinamento do modelo GPT-3 da OpenAI foram disponibilizados,

entre eles, a distribuição de línguas do *dataset*, revelando a predominância do inglês sobre as demais línguas.

Tabela 10 - Distribuição de línguas nos dados de treinamento GPT-3

Posição	Idioma	Quantidade de Documentos	Percentual do Total de Documentos
1	Inglês (en)	235.987.420	93.6882%
2	Alemão (de)	3.014.597	1.19682%
3	Francês (fr)	2.568.341	1.01965%
4	Português (pt)	1.608.428	0.63856%
5	Italiano (it)	1.456.350	0.57818%
6	Espanhol (es)	1.284.045	0.50978%
7	Holandês (nl)	934.788	0.37112%
8	Polonês (pl)	632.959	0.25129%
9	Japonês (ja)	619.582	0.24598%

Fonte: Distribuição das línguas nos dados de treino para GPT-3. Github, s.d.<sup>18</sup>

Porém, ir no sentido contrário do monolinguismo é compreender que a diversidade linguística é um ativo vital para a noção de identidade, no qual a língua se configura como um elemento indissociável da cultura e da herança imaterial de um povo. Nesse sentido, (RUÍZ, 1984) afirma que as línguas são recursos linguísticos valiosos, e que não devem ser tratadas apenas como meio de comunicação, mas sim como ativos que enriquecem a sociedade e promovem o desenvolvimento, a criatividade e a inclusão.

Ainda dentro desse contexto, o Brasil criou o Inventário Nacional da Diversidade Linguística (INDL), cujos objetivos principais são: reconhecer as línguas como referências

---

<sup>18</sup> Disponível em:

<[https://github.com/openai/gpt-3/blob/master/dataset\\_statistics/languages\\_by\\_document\\_count.csv](https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_document_count.csv)>. Acesso em: 30 out. 2024.

culturais brasileiras, valorizando o plurilinguismo em suas múltiplas manifestações; apoiar os processos sociais e políticos voltados à promoção das línguas e de suas comunidades de falantes; fomentar pesquisas e documentação acerca da diversidade linguística nacional; e, por fim, gerenciar um banco de conhecimentos sobre a pluralidade linguística, visando reconhecer línguas específicas como patrimônio imaterial.

Essa iniciativa valida o entendimento de que a preservação e o fortalecimento da diversidade linguística são essenciais para o reconhecimento da identidade e da riqueza cultural do Brasil.

Língua é uma entidade abstrata, variedade é sua manifestação concreta, heterogênea e dinâmica. Do ponto de vista dialetológico ou sociolinguístico, nenhum indivíduo fala uma língua (o Português, o Espanhol), mas sim variedades de língua(s). Toda língua tem múltiplas variedades, que são comumente conhecidas por uma gama de termos como “sotaques”, “jargões”, “gírias”, “falares”, “patoás”, “dialetos”, entre outros. Embora existam definições técnicas para esses termos, eles são muitas vezes empregados para se referir a “sub-línguas”, sejam elas variedades não dominantes, ou variedades que não possuem escrita e/ou tradição literária, ou mesmo que não tenham um respaldo institucional do Estado, adquirindo, muitas vezes, uma perspectiva pejorativa. A dificuldade de compreensão desses fenômenos complexos advém, em grande parte da falta de dados sobre diversidade linguística, mas também decorre da multiplicidade de perspectivas, em muitos casos contraditórios, sobre o que deve ser reconhecido como língua e como variedade de uma língua (INDL, 2016).

Em síntese, reconhecer a diversidade linguística e direcionar políticas que considerem a pluralidade da língua é um elemento central e que, ao mesmo tempo, promove o fortalecimento da identidade local e favorece a equidade no acesso a recursos e tecnologias globais.

No artigo *Análise dos Usos de Inteligência Artificial e suas Implicações para a Diversidade Cultural no Brasil* (PIAZZON, L. et al., 2022) inserido no caderno *Inteligência Artificial e Cultura (2022)* organizado pelo Núcleo de Informação e Coordenação do Ponto BR (NIC), os autores elaboram sobre a acessibilidade e diversidade linguística frente às aplicações de IA apontando para que:

[...] há preocupação quanto ao possível impacto cultural decorrente do uso dessas tecnologias, uma vez que nem sempre contemplam dialetos locais e variações linguísticas (UNESCO, 2022). O estudo identificou experiências que retratam essa questão, por exemplo, ao serem evitados regionalismos no treinamento de aplicações ou mesmo ao serem identificadas limitações maiores no funcionamento de algumas ferramentas em português do que em inglês (PIAZZON, L. et al, 2022, p.158).

Diante disso, sistemas baseados em IA podem não ser capazes de capturar as nuances de uma língua, o que resulta na homogeneização cultural ou na falha em compreender as variações linguísticas próprias de um idioma. A longo prazo, essa limitação pode resultar no apagamento de uma língua, o que, por sua vez, implica na perda de uma cultura no âmbito digital. Embora o tema seja relevante para diversas línguas, é particularmente pertinente para o português do Brasil devido à sua ampla variedade linguística.

A preocupação com a diversidade linguística foi abordada na “Recomendação sobre a Ética da Inteligência Artificial”, publicada em 2022 pela UNESCO, que destaca a necessidade de garantir acessibilidade a todos os grupos linguísticos, promovendo a inclusão das comunidades locais. Além disso, a cartilha de recomendação sublinha a importância da preservação das línguas:

Os atores de IA devem promover a justiça social e salvaguardar a equidade e a não discriminação de qualquer tipo, em conformidade com o direito internacional. Isso implica uma abordagem inclusiva para garantir que os benefícios das tecnologias de IA estejam disponíveis e sejam acessíveis a todos, levando em consideração as necessidades específicas de diferentes grupos etários, sistemas culturais, grupos linguísticos, pessoas com deficiência, meninas e mulheres, e pessoas desfavorecidas, marginalizadas e vulneráveis ou em situações de vulnerabilidade (UNESCO, 2022).

Além disso, sobre os deveres e responsabilidades dos Estados-membros, a cartilha de recomendação aponta:

Os Estados-membros devem trabalhar para promover o acesso inclusivo para todos, incluindo as comunidades locais, aos sistemas de IA com conteúdo e serviços relevantes em âmbito local, respeitando o multilinguismo e a diversidade cultural. Os Estados-membros devem também combater as exclusões digitais e garantir o acesso inclusivo e a participação no desenvolvimento da IA. [...] (UNESCO, 2022).

Já na área de política, no tópico cultura da cartilha de recomendação sobre ética em IA proposto pela UNESCO, há uma orientação sobre o tratamento das línguas:

Os Estados-membros são encorajados a examinar e abordar o impacto cultural dos sistemas de IA, especialmente aplicativos de processamento de linguagem natural (Natural Language Processing – NLP), como tradução automática e assistentes de voz, em relação às nuances da linguagem e expressão humanas. Essas avaliações devem fornecer informações para o projeto e a implementação de estratégias que maximizem os benefícios desses sistemas, suprimindo lacunas culturais e aumentando a compreensão humana, além de abordar implicações negativas, como a redução do uso, o que poderia levar ao desaparecimento de línguas ameaçadas, dialetos locais e variações tonais e culturais associadas à linguagem e à expressão humanas (UNESCO, 2022).

## 6.1

### EXEMPLOS DE PROBLEMAS DE ALINHAMENTO LINGUÍSTICO E CULTURAL - PORTUGUÊS BRASILEIRO

Considerando a importância do alinhamento linguístico das aplicações de IA, surgem exemplos de interações falhas entre falantes de português (Brasil) e sistemas de IA, que claramente demonstram a falta de refinamento para a língua portuguesa e a ausência de compreensão de suas especificidades. Um exemplo recente é o modelo Llama 3.2, da Meta, lançado no Brasil no segundo semestre de 2024. Durante a interação com o modelo, é possível perceber que a intenção do usuário não é completamente compreendida, resultando na geração de imagens confusas, textos agramaticais e até mesmo violações de segurança. Este episódio destaca a limitação do modelo em se adaptar à diversidade linguística/cultural brasileira.

Figura 7 - Prompt: Gere a imagem da Xuxa comendo um pé de moleque



Fonte: Gerado pelo autor

Figura 8 - Prompt: Gere uma imagem de uma pessoa “aperreada” no nordeste do Brasil



Fonte: Gerado pelo autor

Figura 9 - Prompt: Crie a imagem do homem aranha plantando bananeira



Fonte: Folha de SP, 2024.

A análise de diferentes interações com o *chatbot* da Meta revelam limitações significativas no processamento de expressões idiomáticas e regionalismos brasileiros. Na figura 7, ao receber o comando "gere a Xuxa comendo pé de moleque", o modelo interpretou

a expressão 'pé de moleque' em seu sentido literal [como o pé de uma criança], em vez de reconhecer o tradicional doce brasileiro feito de amendoim. Isso resultou em uma imagem descontextualizada que não correspondia à intenção inicial do usuário.

Na figura 8, durante uma interação com o modelo Llama 3.2, foi solicitada a geração de uma imagem de alguém 'aperreado' - termo comum no português brasileiro, especialmente no Nordeste, usado para descrever alguém frustrado ou chateado. Ao contrário do significado real da expressão, o modelo gerou a imagem de uma pessoa sorridente, deixando evidente sua falta de domínio sobre regionalismos.

Já na figura 9, ao retornar à solicitação "Crie a imagem do homem aranha plantando bananeira", o modelo falhou em reconhecer que 'plantar bananeira' se refere a um movimento acrobático característico da capoeira, interpretando a expressão literalmente como o ato de cultivar uma bananeira [planta]. Estes exemplos ilustram como os modelos de IA ainda apresentam dificuldades na compreensão de expressões idiomáticas, regionalismos e elementos culturais específicos do português brasileiro, o que pode limitar sua eficácia em contextos que exigem entendimento mais profundo da cultura e da língua local.

Embora os exemplos de geração de imagens sejam mais evidentes e facilmente perceptíveis, as implicações mais profundas em relação à diversidade linguística merecem atenção. Alguns desses modelos de IA podem estar contribuindo para o apagamento de línguas e variedades linguísticas. Segundo um estudo conduzido por pesquisadores da AWS<sup>19</sup>, atualmente 57% do conteúdo disponível na internet é gerado por aplicações de IA, sendo que a tendência é que esse número aumente até 2026.

Ainda dentro dessa temática, de acordo com um estudo publicado na revista Nature, quando um modelo de linguagem é treinado com dados gerados por outras aplicações de IA, ou seja, a partir de dados sintéticos, acarreta um colapso do modelo e conseqüentemente, a queda de desempenho. Esse resultado gera a falta de personalização do conteúdo, com um impacto significativo no enfraquecimento de línguas minoritárias e na perda de diversidade cultural, resultando em uma forma de apagamento linguístico e cultural.

Como forma de mitigar esses problemas, técnicas como o *Fine-Tuning* e Aprendizagem por Reforço a partir do Feedback Humano (*RLHF*) têm sido aplicadas para ajustar modelos de IA a contextos específicos. O *Fine-Tuning* permite que os modelos sejam

---

<sup>19</sup> Disponível em: <<https://arxiv.org/pdf/2401.05749>>. Acesso em: 02 nov. 2024

refinados com dados representativos de uma língua ou comunidade específica, aumentando a precisão em contextos locais. Já o *RLHF* utiliza o feedback humano, em que falantes nativos de uma língua específica avaliam as respostas geradas pelo *chatbot*, oferecendo reforços positivos ou negativos aos modelos de linguagem.

Esses processos ajudam a orientação do modelo para que gerem respostas cultural e linguisticamente apropriadas para determinado idioma, ajustando o desempenho do modelo com base nas expectativas e nuances da língua.

Porém, recursos técnicos, quando observados de forma isolada, não são suficientes para estancar as adversidades tecnológicas atuais. É essencial adotar políticas e incentivos integrados que considerem o ecossistema de IA na totalidade, além de um olhar atento para as línguas e a diversidade cultural. A partir dessa perspectiva, a diretora da Organização dos Estados Ibero-Americanos, Ana Paula Laborinho, destaca a importância de uma estratégia de IA voltada para a língua portuguesa. Essa estratégia, segundo Laborinho, deve incluir uma cooperação ampliada, abrangendo outras línguas, além de garantir a coerência das políticas implementadas. Laborinho enfatiza que:

Está sobejamente estudada a relação entre línguas e geopolítica, bem como os benefícios de partilhar uma língua global como é o caso do português. Mas, atualmente, uma língua global não se define apenas pelo número de falantes, o número de continentes em que é língua oficial ou os recursos naturais desses países. Todos estes indicadores são muito favoráveis ao posicionamento internacional do português, mas os desafios neste mundo contemporâneo implicam um idioma tecnologicamente avançado, virado para a inovação e com uma produção científica relevante [...] Nos últimos anos, tem havido um trabalho colaborativo em torno da reflexão e ação sobre o português no espaço digital. É urgente dispormos de indicadores e investigações que contribuam para o seu posicionamento neste domínio (LABORINHO, 2024).

A união dos países pertencente a comunidade lusófona pode ser uma potencial via para a promoção do acesso igualitário às línguas por meio das tecnologias digitais, desde que, o objetivo comum dessa cooperação entre os países falantes da língua portuguesa esteja centrada em prol de minimizar os impactos linguísticos dessas aplicações de IA.

## 7 CONCLUSÃO

A preservação linguística no domínio virtual representa uma abordagem potente e eficaz para assegurar que a cultura de um povo se mantenha ativa no espaço digital, assim como serve de instrumento para proteger línguas ameaçadas de extinção. Segundo András Kornai há um declínio das línguas no ambiente digital, causado pela ausência de recursos e representações adequadas, o que resulta na extinção dessas línguas nas interações tecnológicas modernas, o que é chamado pelo autor de *digital language death* (KORNAI, 2013).

A internet e as tecnologias digitais oferecem plataformas que facilitam a documentação, o armazenamento e a divulgação de materiais linguísticos, permitindo que comunidades falantes de línguas minoritárias compartilhem e revitalizem seu patrimônio cultural. Iniciativas como repositórios digitais, redes sociais e aplicativos de mensagens têm sido utilizadas para criar "praças virtuais" onde essas línguas ganham visibilidade e uso cotidiano, especialmente entre as gerações mais jovens. O desenvolvimento de aplicações de inteligência artificial tem potencializado a tradução e a acessibilidade de idiomas, promovendo um senso de pertencimento e identidade cultural entre grupos linguísticos. Nesse ponto de vista, o ambiente digital serve como um meio de preservação, mas também como um espaço de interação e como produto, fortalece a diversidade linguística global.

Pensando em um ecossistema digital, o incentivo à construção de *datasets* em português é fundamental para a preservação e valorização da língua, especialmente em um contexto de crescente digitalização e globalização.

A criação de bases de dados expressivas e diversificadas em português permite o avanço e desenvolvimento das novas tecnologias, garantindo que grandes modelos de linguagem estejam alinhados culturalmente em relação à comunidade de falantes de língua portuguesa. Como consequência, o fomento de base de dados também serve como instrumento para a documentação e promoção das múltiplas variantes regionais e manifestações culturais da língua portuguesa. Portanto, promover a inclusão de dados de treinamento que reflitam inclusive o português falado em diferentes contextos, como dialetos, jargões e expressões culturais, pode ser um caminho possível para evitar a perda de identidade

linguística e garantia que os modelos possam compreender e gerar textos que sejam relevantes para a comunidade de língua portuguesa.

Nesse contexto, iniciativas voltadas à documentação e digitalização de dialetos e expressões idiomáticas regionais ganham destaque nesse processo, ao incorporarem uma visão plural aos dados de treinamento, contribuem, de certo modo, para a permanência da diversidade do português em suas diferentes manifestações culturais e sociais, evitando assim um apagamento digital. Essa movimentação poderia caminhar em conjunto com a retomada de uma das frentes apagadas do INDL: a inclusão das variedades regionais do português brasileiro no inventário nacional da diversidade linguística. Hoje, o INDL comporta apenas as diferentes línguas, não contemplando a variedade dialetológica no Brasil. O mapeamento das diferentes variantes linguísticas no Brasil fomentaria, portanto, o Atlas Linguístico do Brasil, trazendo uma visão completa do cenário linguístico no território brasileiro.

A disponibilização de dados linguísticos em formato aberto representa outro pilar nesse processo: esta prática democratiza o acesso a recursos essenciais para pesquisadores, educadores e desenvolvedores, possibilitando a criação de tecnologias que respeitam as especificidades do português. Em paralelo, o estabelecimento de políticas públicas de incentivo à pesquisa voltadas para a preservação e evolução do idioma no contexto digital torna-se indispensável para fortalecer a posição do português na rede tecnológica global, especialmente no campo da inteligência artificial e do processamento de linguagem natural. O desenvolvimento de sistemas de monitoramento das transformações linguísticas completa esse ecossistema, permitindo o acompanhamento contínuo das mudanças na língua e a atualização constante das bases de dados. Esse processo dinâmico de documentação, análise e monitoramento contribui para a inclusão digital e o fortalecimento da identidade cultural, assegurando a vitalidade e a relevância do português brasileiro no contexto digital global.

Porém, esses incentivos não devem estar isolados geopoliticamente. Em julho de 2024, o governo brasileiro lançou o Plano Brasileiro de Inteligência Artificial (PBIA), visando traçar estratégias para posicionar o país como um dos protagonistas no contexto tecnológico mundial. O plano destaca a importância de fomentar a pesquisa e a inovação em IA, promovendo iniciativas que sustentem o Brasil para competir em pé de igualdade com outras nações no setor da inteligência artificial. Incentivos como esses são necessários não apenas

para impulsionar o desenvolvimento científico nacional, mas também para consolidar o Brasil como um país competitivo e inovador no panorama internacional de IA.

Por fim, cabe pensar em uma política linguística para grandes modelos de linguagem, já que processam e manipulam a língua portuguesa, com o propósito de garantir que esses sistemas sejam efetivos e representativos da diversidade linguística e cultural do português brasileiro. Essa política deve incluir diretrizes claras sobre a coleta, o tratamento e a utilização de dados em português, assegurando que as variações regionais e sociais da língua sejam respeitadas e integradas aos modelos. A implementação de uma política linguística abrangente também deve considerar questões éticas relacionadas ao uso de dados, como a proteção da privacidade dos falantes e a validação das contribuições dos grupos linguísticos.

Como resultado, um planejamento linguístico estratégico e direcionado para *LLMs*, serve como balizador do desenvolvimento tecnológico ao afastar a homogeneização da língua e garantir uma equidade digital para os falantes de português. Dessa forma, um planejamento linguístico coerente vai ao encontro da preservação linguística e valorização da língua portuguesa em sua totalidade, promovendo um ambiente digital mais inclusivo e representativo.

## REFERÊNCIAS

- AI PREPAREDNESS INDEX (AIPI). IMF, 2024. Disponível em: <https://www.imf.org/external/datamapper/datasets/AIPI>. Acesso em: 14 ago. 2024.
- Kornai, A. **Digital Language Death**, 2013. PLoS ONE 8(10): e77056. <https://doi.org/10.1371/journal.pone.0077056>
- AWS. **O que é o aprendizado profundo em IA?**, *s.d.* Disponível em: <https://aws.amazon.com/pt/what-is/deep-learning/>. Acesso em: 5 out. 2024.
- AWS. **O que é uma rede neural?**, *s.d.* Disponível em: <https://aws.amazon.com/pt/what-is/neural-network/>. Acesso em: 11 out. 2024.
- CAZZANIGA, F. et al. **Gen-AI: Artificial intelligence and the future of work**. IMF Staff Discussion Note, SDN2024/001, Washington, DC: International Monetary Fund, 2024.
- BRAINS DEV. **Token e embedding**: conceitos da IA e LLMs, 2024. Acesso em: 26 de out. de 2024.
- BROWN, T. B. et al. **Language Models are Few-Shot Learners**, 2020. Disponível em: <https://arxiv.org/abs/2005.14165>. Acesso em: 20 out. 2024.
- CALVET, L.-J. **As políticas linguísticas**. São Paulo: Parábola, 2007.
- CALVET, L.-J. **Sociolinguística: uma introdução crítica**. São Paulo: Parábola, 2002.
- COMITÊ GESTOR DA INTERNET NO BRASIL. **Inteligência Artificial e Cultura: perspectivas para a diversidade cultural na era digital**. São Paulo, 2022. Disponível em: [https://cetic.br/media/docs/publicacoes/7/20220928131646/estudos\\_setoriais-inteligencia\\_artificial\\_e\\_cultura.pdf](https://cetic.br/media/docs/publicacoes/7/20220928131646/estudos_setoriais-inteligencia_artificial_e_cultura.pdf). Acesso em: 22 nov. 2024.
- DATABRICKS. **Rede neural artificial**, *s.d.* Disponível em: <https://www.databricks.com/br/glossary/artificial-neural-network>. Acesso em: 12 out. 2024.

EUROLINGUA-GPT. Disponível em: <<https://www.ai.se/sv/projekt/eurolingua-gpt>>. Acesso em: 14 out. 2024.

ETHNOLOGUE 200, 2024. Disponível em: <https://www.ethnologue.com/insights/ethnologue200/>. Acesso em: 13 out. 2024.

FACEBOOK COMPANY. **Apresentando Meta Llama 3**: o grande modelo de linguagem de código aberto mais capaz até hoje, 2024. Disponível em: <https://about.fb.com/br/news/2024/04/apresentando-meta-llama-3-o-grande-modelo-de-lingua-gem-de-codigo-aberto-mais-capaz-ate-hoje/>. Acesso em: 21 nov. 2024.

IBM. **Rede neural profunda**, 2024. Disponível em: <https://www.ibm.com/br-pt/topics/neural-networks#:~:text=As%20redes%20neurais%20feedforward%2C%20ou.e%20uma%20camada%20de%20sa%C3%ADda>. Acesso em: 25 out. 2024.

OBDILCI. **Indicators for the presence of languages in the internet**, 2024. Disponível em: <https://www.obdilci.org/projects/main/>. Acesso em: 21 nov. 2024.

INSTITUTO DO PATRIMÔNIO HISTÓRICO E ARTÍSTICO NACIONAL (IPHAN). **Guia de pesquisa e documentação para o INDL**: patrimônio cultural e diversidade linguística. Brasília, 2014.

KAUFMAN, D.; COZMAN, F. G. **Viés no aprendizado de máquina em sistemas de inteligência artificial**: a diversidade de origens e os caminhos de mitigação. *Revista USP*, v. 135, 2022.

KEW, T.; SCHOTTMANN, F.; SENNRICH, R. **Turning English-centric LLMs Into Polyglots**: How Much Multilinguality Is Needed?, 2023.

LABORINHO, A. P. **Língua Portuguesa e os desafios da Inteligência Artificial**. [S. l.], 2024. Disponível em: <https://www.dn.pt/autor/1931030047/ana-paula-laborinho/>. Acesso em: 22 nov. 2024.

MAIA, H. [S. l.], 2023. Disponível em: <https://news.un.org/pt/story/2023/09/1820267>. Acesso em: 10 nov. 2024.

MARCONDES, D. **Textos Básicos de Linguagem**. Rio de Janeiro: Zahar, 2009.

UNESCO. **Recomendação sobre a Ética da Inteligência Artificial**. Paris, 2022. Disponível em: [https://unesdoc.unesco.org/ark:/48223/pf0000381137\\_por](https://unesdoc.unesco.org/ark:/48223/pf0000381137_por). Acesso em: 10 nov. 2024.

RAMOS, D. B. Reforma tributária: vai gerar a saída de empresas de tecnologia do Brasil?  
Disponível em:

<<https://mitsloanreview.com.br/reforma-tributaria-vai-gerar-a-saida-de-empresas-de-tecnologia-do-brasil/>>. Acesso em: 5 dez. 2024.

RUIZ, R. **Orientations in language planning**. [S. l.], 1984.

THOMPSON, B. et al. **A Shocking Amount of the Web is Machine Translated**: Insights from Multi-Way Parallelism. [S. l.], 2024. Disponível em: <https://arxiv.org/pdf/2401.05749>. Acesso em: 13 nov. 2024.

UNESCO. **Recomendação sobre a Ética da Inteligência Artificial**. Paris: UNESCO, 2022. Disponível em: [https://unesdoc.unesco.org/ark:/48223/pf0000381137\\_por](https://unesdoc.unesco.org/ark:/48223/pf0000381137_por). Acesso em: 10 nov. 2024.

UNESCO. **Repensar as políticas culturais**: criatividade para o desenvolvimento, Relatório global da Convenção de 2005. Paris, 2024. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000266025>. Acesso em: 10 nov. 2024.

VASWANI, A. et al. **Attention is all you need**. [S. l.], 2017. Disponível em: <https://arxiv.org/abs/1706.03762>. Acesso em: 12 out. 2024.