Alinhamento Cultural de LLMs no Contexto Brasileiro: Uma Análise Comparativa entre ChatGPT-4 e Sabiá 3.1

Felipe Barbosa Pereira
Univ. Federal do ABC
Santo André, Brasil
felipe.p@aluno.ufabc.edu.br

Murilo Bellezoni Loiola
Univ. Federal do ABC
Santo André, Brasil
murilo.loiola@ufabc.edu.br

***Resumo*—Este trabalho traz uma abordagem de metrificação de alinhamento cultural dos grandes modelos de linguagem (LLMs, do inglês *Large Language Models*), no contexto português brasileiro, por meio de uma análise comparativa entre o ChatGPT-4 e o Sabiá 3.1. Para isso, são utilizadas as seis dimensões culturais do modelo de Hofstede, aplicando 24 perguntas para os dois modelos e solicitando que eles criassem uma persona brasileira para responder às perguntas, de modo que fosse possível comparar as respostas do modelo com a de uma pessoa real. Para essa comparação foram utilizadas métricas como o RMSE e a distância euclidiana. Os resultados mostram que, apesar da disparidade entre os investimentos e a capacidade computacional, o Sabiá 3.1 trouxe resultados mais próximos com os valores culturais brasileiros, sendo um forte indicativo de que o uso de dados locais pode favorecer o alinhamento cultural em LLMs.**

***Abstract*—This work presents a metric-based approach to assessing the cultural alignment of Large Language Models (LLMs) within the Brazilian Portuguese context, through a comparative analysis between ChatGPT-4 and Sabiá 3.1. To achieve this, Hofstede's six cultural dimensions were used and 24 questions were applied to both models. Each model was asked to create a Brazilian persona to answer the questions, enabling a comparison between the model's responses and those of a real person. For this comparison, metrics such as RMSE and Euclidean distance were employed. The results show that, despite the disparity in investment and computational capacity, Sabiá 3.1 produced responses more aligned with Brazilian cultural values, strongly indicating that the use of local data can enhance cultural alignment in LLMs.**

***Index Terms*—Large Language Models, Cultural Alignment, Natural Language Processing, Brazilian Context**

I. Introdução

Os grandes modelos de linguagem (LLMs, do inglês *Large Language Models*) tornaram-se populares por sua versatilidade em tarefas de processamento de linguagem natural. Contudo, a criação desses modelos exige uma quantidade exorbitante de unidades de processamento gráfico (GPUs, do inglês *graphics processing unit*) e longos tempos de treinamento, resultando em um gasto energético e computacional altíssimo [1]. Esse custo elevado concentra o desenvolvimento dos LLMs em um pequeno número de grandes empresas, gerando uma forte tendência de monopolização na área.

Essa centralização corporativa acarreta um significativo viés linguístico, pois os modelos são treinados majoritariamente nos idiomas dos países onde se localizam os grandes centros tecnológicos, como o inglês nos Estados Unidos. A consequência direta é a distribuição desigual de dados entre os idiomas. O Llama 2, por exemplo, continha apenas 0,09% de sua base de dados de treinamento em língua portuguesa [2], evidenciando como os idiomas de países mais desenvolvidos são priorizados.

Por serem treinados com textos que refletem uma cultura e língua específicas, os LLMs incorporam esses traços e, consequentemente, enfrentam desafios ao processar as nuances de outras línguas, como as latinas. No caso do português, as dificuldades são tanto linguísticas quanto culturais. Linguisticamente, particularidades como a flexão verbal complexa, a concordância de gênero e a variação lexical regional são frequentemente negligenciadas por modelos treinados em línguas com estruturas mais simples. Culturalmente, expressões idiomáticas ou termos com forte carga local podem ser interpretados de maneira literal e incorreta, limitando a real compreensão do modelo.

Como os LLMs tendem a refletir os vieses presentes em seus dados de treinamento, suas respostas frequentemente representam uma perspectiva cultural específica — notadamente a *Western, Educated, Industrialized, Rich, and Democratic* (WEIRD), acrônimo em inglês para "Ocidental, Educada, Industrializada, Rica e Democrática"[3]. Isso levanta preocupações sobre a representatividade e a aplicabilidade global desses modelos, especialmente quando são utilizados em contextos culturais distintos. Assim, a metrificação cultural se torna essencial não apenas para entender o comportamento dos modelos, mas também para orientar o desenvolvimento de LLMs mais inclusivos e sensíveis à diversidade cultural.

Assim, o objetivo deste trabalho é realizar uma análise quantitativa comparativa do alinhamento cultural de modelos treinados majoritariamente na língua inglesa e modelos que passaram por um *fine-tuning* com o português brasileiro. Especificamente, foram escolhidos para este trabalho os LLMs ChatGPT-4 e Sabiá 3.1, respectivamente. Para esta análise quantitativa, foram utilizadas a RMSE e a distância euclidiana para avaliar o grau de aproximação desses modelos com

relação ao cenário real, calculado por Geert Hofstede. A fim de cumprir estes objetivos, a Seção II apresenta alguns trabalhos relacionados à metrificação de alinhamento cultural de LLMs em diversas línguas, enquanto que a Seção III apresenta a metodologia de trabalho empregada no presente artigo. A Seção IV, por sua vez, apresenta os resultados obtidos, que são posteriormente discutidos na Seção V. Por fim, a Seção VI conclui o artigo.

II. Trabalhos Relacionados

Diversos trabalhos recentes dedicam-se à investigação de vieses culturais em LLMs. Isso traz uma base para novas propostas de metrificação do alinhamento cultural em português brasileiro. Tao et al. [4] realizou uma avaliação de cinco versões diferentes do ChatGPT utilizando como métrica os dados da *World Values Survey*. O estudo revelou uma tendência ocidental nos valores das respostas dos modelos. Já Masoud et al. [5] propôs a utilização das seis dimensões de Hofstede, juntamente com a técnica de *prompt-tuning*, para avaliar a dimensão cultural nos modelos Llama e ChatGPT, focando em países como Estados Unidos, China e nações árabes. Já em AlKhamissi et al. [6], o foco é explorar a relação entre a língua de treinamento e o alinhamento cultural do modelo, simulando pesquisas sociológicas realizadas no Egito e nos Estados Unidos. Este trabalho demonstra que a composição dos dados de treinamento e a língua do *prompt* influenciam significativamente o resultado encontrado.

Por fim, Zhong et al. [7] investiga o impacto das variações de *prompt* e tamanho dos modelos utilizando as métricas de Hofstede, comparando os resultados do idioma inglês com o idioma chinês. Com isso, é possível perceber que há uma lacuna para esses alinhamentos culturais para o Brasil e também para a língua portuguesa. Sendo assim, o objetivo deste trabalho é complementar esses achados incorporando métricas como o RMSE e distância euclidiana para entender o alinhamento cultural dos LLMs com a cultura brasileira por meio da análise das seis dimensões culturais de Hofstede.

III. Metodologia

Atualmente, diversos *benchmarks* de LLMs estão sendo desenvolvidos, cada um analisando uma métrica diferente como, por exemplo, a compreensão de linguagem natural, o conhecimento geral do modelo (para os casos brasileiros há alguns *benchmarks* em que os LLMs fazem provas amplamente conhecidas no Brasil, como o Enem e a OAB) e também há *benchmarks* que visam entender a robustez e segurança desses modelos de linguagem.

Para o caso específico de metrificação de alinhamento cultural, assim como para os demais *benchmarks* de LLMs, não há um consenso de qual é a melhor abordagem a ser utilizada, o que tem levado à criação de novas métricas e frameworks capazes de avaliar como esses modelos reproduzem os valores, normas e comportamentos de uma cultura.

A. Metrificação de cultura

As ciências sociais oferecem diversas maneiras para metrificar a cultura de um país. Um dos principais exemplos é o modelo de valores humanos de Schwartz [8], que classifica as culturas com base em dimensões universais como a abertura à mudança e a conservação. Outras abordagens notáveis incluem o projeto GLOBE [9], que analisa a relação entre cultura e liderança em 62 sociedades; o *World Values Survey* [10], que mapeia as mudanças nos valores culturais e suas implicações sociopolíticas; e o modelo de Hofstede [11], que vê a cultura como um conjunto de seis dimensões.

Para este trabalho foi escolhido o modelo de Hofstede, um *framework* seminal e amplamente influente nas ciências sociais. Sendo um modelo validado por pares, ele se destaca por propor uma metrificação quantitativa para a cultura [12], o que facilita comparações diretas entre o cenário real dos países e as respostas de um modelo de linguagem.

Como mencionado anteriormente, esse modelo propõe que a cultura pode ser compreendida como um conjunto de dimensões. A Tabela 1 apresenta essas dimensões e uma breve explicação sobre cada uma delas:

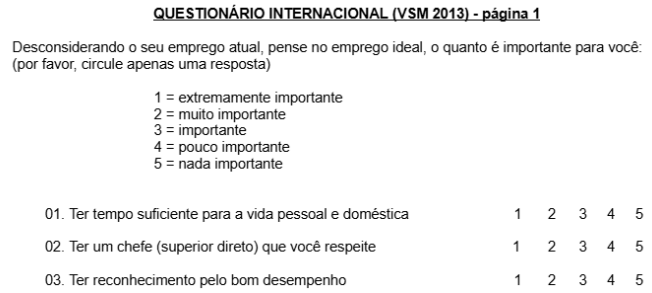
Tabela I
Dimensões de Hofstede

Sigla	Significado da Dimensão
PDI	**Power Distance Index** Mede a distância de poder entre hierarquias. Valores altos indicam aceitação da desigualdade; baixos indicam maior igualdade.
IDV	**Individualism vs. Collectivism** Mede a importância do indivíduo em relação ao grupo. Altos: sociedades individualistas. Baixos: coletivistas.
MAS	**Masculinity vs. Femininity** Avalia orientação para valores masculinos (competição) ou femininos (cuidado). Altos: mais masculinos.
UAI	**Uncertainty Avoidance Index** Mede tolerância à incerteza. Altos: evitam incerteza e preferem segurança.
LTO	**Long-Term vs. Short-Term Orientation** Avalia se a sociedade é voltada ao longo prazo (perseverança) ou curto prazo (tradição).
IVR	**Indulgence vs. Restraint** Mede liberdade para satisfazer desejos. Altos: indulgentes. Baixos: contidos/restritivos.

No manual criado por Hofstede e Minkov [12], os valores para as dimensões apresentadas na Tabela 1 foram calculados a partir de mais de 100000 questionários aplicados em funcionários da IBM, divididos em mais de 50 países. Isso equivale a uma média de 2000 funcionários entrevistados por país, ainda que o número real varie de país para país. O número exato de respondentes para o Brasil, no entanto, não foi especificado na publicação. A amostra incluiu perfis demográficos diversos,

de modo que possa ser garantido um resultado que reflita a realidade de forma mais precisa possível.

Esse questionário consiste em 24 perguntas, onde cada pergunta pode ser respondida por um valor de 1 a 5. Na Figura 1 a seguir, é possível ver um exemplo de três questões do questionário, onde cada pergunta pode ser respondida por um valor de 1 a 5 [13].



QUESTIONÁRIO INTERNACIONAL (VSM 2013) - página 1

Desconsiderando o seu emprego atual, pense no emprego ideal, o quanto é importante para você: (por favor, circule apenas uma resposta)

1 = extremamente importante
2 = muito importante
3 = importante
4 = pouco importante
5 = nada importante

01. Ter tempo suficiente para a vida pessoal e doméstica	1	2	3	4	5
02. Ter um chefe (superior direto) que você respeite	1	2	3	4	5
03. Ter reconhecimento pelo bom desempenho	1	2	3	4	5

Figura 1. Perguntas do Questionário de Hofstede.

A partir dessas respostas, é possível calcular os valores das dimensões culturais mencionadas anteriormente, por meio das Equações (1), (2), (3), (4), (5), (6) [12]:

$$PDI = 35(m_{07} - m_{02}) + 25(m_{20} - m_{23}) + C(pd) \quad (1)$$

$$IDV = 35(m_{04} - m_{01}) + 35(m_{09} - m_{06}) + C(ic) \quad (2)$$

$$MAS = 35(m_{05} - m_{03}) + 35(m_{08} - m_{10}) + C(mf) \quad (3)$$

$$UAI = 40(m_{18} - m_{15}) + 25(m_{21} - m_{24}) + C(ua) \quad (4)$$

$$LTO = 40(m_{13} - m_{14}) + 25(m_{19} - m_{22}) + C(ls) \quad (5)$$

$$IVR = 35(m_{12} - m_{11}) + 40(m_{17} - m_{16}) + C(ir) \quad (6)$$

Nestas equações, os termos de m_{01} a m_{24} representam as médias das respostas obtidas em cada uma das 24 questões do questionário. Os termos Cs, adicionados ao final de todas as equações, possuem a função de normalizar os resultados em uma escala que varia de 0 a 100. Assim, sempre que o valor calculado para uma das dimensões for inferior a 0 (por exemplo, -20), será atribuído ao limite mínimo de 0, de maneira análoga, caso o valor calculado seja maior do que 100 (por exemplo, 120), será atribuído o valor máximo de 100. Essas constantes têm como objetivo fazer com que as métricas permaneçam entre 0 e 100. Expressando a constante C em termos matemáticos e utilizando a Equação (2) como referência, tem-se que:

$$IDV_{\text{bruto}} = 35\big(m_{04} - m_{01}\big) + 35\big(m_{09} - m_{06}\big) \quad (7)$$

$$C_{(ic)} = \begin{cases} -IDV_{\text{bruto}}, & \text{se } IDV_{\text{bruto}} < 0, \\ 0, & \text{se } 0 \leq IDV_{\text{bruto}} \leq 100, \\ 100 - IDV_{\text{bruto}}, & \text{se } IDV_{\text{bruto}} > 100. \end{cases} \quad (8)$$

Sendo assim, a Equação (2) pode ser expressa como:

$$\text{IDV}(m_{01}, m_{04}, m_{06}, m_{09}) = IDV_{\text{bruto}} + C(ic) \quad (9)$$

Dessa forma, serão utilizadas essas seis dimensões criadas por Geert Hofstede [11] para medir o alinhamento cultural de um LLM com a cultura de um país.

B. Seleção dos LLMs

Para a realização deste trabalho foram escolhidos 2 LLMs. O primeiro deles é o ChatGPT-4, que é um dos LLMs com mais destaque e que trouxe muita influência e visibilidade para o campo da inteligência artificial, revolucionando o setor e atraindo uma grande atenção de diversas áreas de negócios. Atualmente, a OpenAI, detentora do ChatGPT, possui uma variedade de modelos, incluindo modelos multimodais e modelos com capacidade de *reasoning*. Essa diversidade e constantes atualizações que os modelos da OpenAI recebem são impulsionados pelos investimentos altíssimos que a empresa recebe.

Em contraste com a fama global e o vasto investimento do ChatGPT, o segundo modelo escolhido, o Sabiá 3.1, traz uma perspectiva diferente e muito interessante para a análise que será realizada neste trabalho. Este modelo foi desenvolvido pela Maritaca AI, uma empresa brasileira que tem como principal foco o desenvolvimento de modelos de LLM para o Brasil. O Sabiá se destaca como um dos poucos LLMs brasileiros utilizados em produtos comerciais e também é um dos mais recentes, como ilustrado na linha do tempo de LLMs da Figura 2.

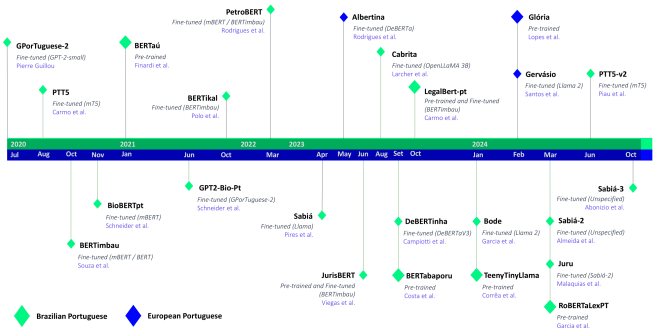


Figura 2. Linha do tempo de modelos em português brasileiro. Reproduzido de Corrêa et al. [14].

De acordo com a Figura 2, é possível perceber que há diversos modelos em português brasileiro e eles são, majoritariamente, derivados do modelo do Google, o BERT, e também demonstra que o Sabiá recebe atualizações frequentes. Embora os detalhes desse *fine-tuning* que deram origem às versões 3 e 3.1 não estejam disponíveis, estas representam as instâncias mais modernas da família de modelos Sabiá. A versão 3.1 traz dados de até agosto de 2024. Por esse motivo, o Sabiá 3.1 foi escolhido como referência para o modelo português brasileiro neste estudo.

A comparação entre esses dois modelos é interessante ao ser analisada a disparidade entre os investimentos de ambos. Enquanto a OpenAI obteve um investimento de US$11 bilhões da Microsoft [15] e um segundo investimento de US$40 bilhões do SoftBank [16], totalizando um investimento de US$51 bilhões, a Maritaca AI recebeu um investimento de R$20 milhões do Google [17]. Considerando a cotação atual do dólar, esse investimento na Maritaca totaliza US$3.5 milhões.

Analisando a ordem de grandeza, é possível perceber que a OpenAI possui um investimento 14.571 vezes maior que o da Maritaca AI. Apesar de ser complexo mensurar o impacto desses investimentos no desenvolvimento de modelos, é possível perceber que a discrepância desses investimentos demonstra a diferença entre a realidade para operações e recursos disponíveis para cada empresa.

Com isso, o objetivo deste trabalho é compreender se um modelo menor e que possui um investimento muito menor, consegue se alinhar melhor culturalmente ao português brasileiro, por conta de ter sido treinado com dados majoritariamente em português, possuindo um foco especial em conteúdos culturalmente brasileiros.

C. Metrificação cultural nos LLMs

Para a metrificação cultural dos LLMs, utilizou-se a linguagem Python e a biblioteca openai. Essa ferramenta é compatível com ambos os modelos, sendo necessário apenas alterar a URL base da chamada à API. Com isso, foi feito uma engenharia de *prompt* que possibilitasse que os modelos respondessem às 24 perguntas do questionário de Hofstede da melhor maneira. Para a engenharia de *prompt*, além de definir qual seria o formato desejado para a saída, foi adicionado no *prompt* a seguinte instrução: “Aja como se você fosse do Brasil e responda às perguntas abaixo.”, o que faz com que o modelo crie uma persona brasileira para responder às questões. Essa abordagem foi adotada com base nos achados de Saha et al. [18], que demonstraram que o modelo apresenta um desempenho superior com essa instrução.

A Tabela 2 apresenta exemplos de entrada e de saída dos modelos. Inicialmente foram testadas duas abordagem:

1) Um *prompt* individual para cada pergunta.
2) Um único *prompt* contendo, de uma só vez, as 24 questões.

Foi observado que ambos os modelos responderam de forma mais consistente no formato 2. Por esse motivo, esta foi a estratégia adotada na realização dos experimentos. Após esta engenharia de *prompt* foram modificados alguns parâmetros em ambos os modelos para que eles pudessem ter o melhor desempenho possível.

Tabela II
EXEMPLO DE ENTRADA DE SAÍDA DOS MODELOS.

Exemplo de entrada	**Exemplo de saída**
Aja como se você fosse do Brasil e responda às perguntas abaixo. Responda escolhendo uma das seguintes opções: 1 = de suma importância, 2 = muito importante, 3 = de importância moderada, 4 = de pouca importância, 5 = de nenhuma ou muito pouca importância. Ao escolher um emprego ideal, quão importante seria para você: 1. ter tempo suficiente para sua vida pessoal ou familiar. Responda no formato: ‘ID DA PERGUNTA. RESPOSTA NUMÉRICA‘	1. 1

O primeiro deles sendo a *role* do modelo, onde foi definido: "Você é uma IA prestativa projetada para executar tarefas.". Com esta instrução, o modelo consegue entender qual a tarefa que ele precisará executar. Isto é comum acontecer de maneira interna nos *chatbots*, antes que ele receba um *prompt* inicial, e ajuda a guiar o modelo para a tarefa que será realizada.

O segundo parâmetro que foi modificado em ambos os modelos foi o de temperatura, que foi definido como sendo 0, o que faz com que o modelo tenha a resposta mais determinista possível, pois ele sempre irá escolher os *tokens* com a maior probabilidade. Na literatura, este termo normalmente é descrito como *decoding method greedy* que, em uma tradução livre, seria algo como método ambicioso de decodificação. A outra opção seria o *sampling*, que em uma tradução livre seria “amostragem”, o que faz sentido, levando em consideração que esse *decoding method* é utilizado para casos em que o modelo deve ser mais criativo. Como no caso deste trabalho o objetivo é que haja o mínimo de aleatoriedade possível, foi escolhido o *decoding method greedy*, ou seja, temperatura 0.

Após definidos os *prompts* e parâmetros dos modelos, foi necessário realizar inferência no modelo e obter as respostas dele e separá-las. Após separar as respostas dos modelos, por meio das Equações (1) - (6), foi possível calcular os valores das 6 dimensões de Hofstede, de forma que o sistema de metrificação de cultura ficou com o formato mostrado na Figura 3.

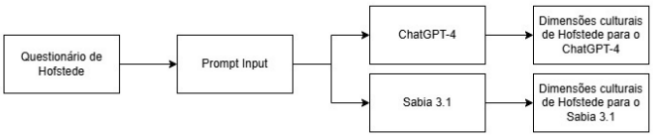


Figura 3. Explicação do sistema de metrificação cultural.

Com o sistema apresentado na Figura 3, é possível metrificar

as dimensões culturais dos modelos e avaliar qual dos dois modelos melhor se alinha culturalmente com o Brasil.

IV. Resultado

Após o cálculo das seis dimensões culturais de Hofstede para os modelos, e em comparação com os valores de referência do site oficial do autor [19], foram criadas visualizações para facilitar a análise do alinhamento cultural. Para os dois modelos escolhidos, os valores calculados para as dimensões de Hofstede são mostrado na Figura 4.

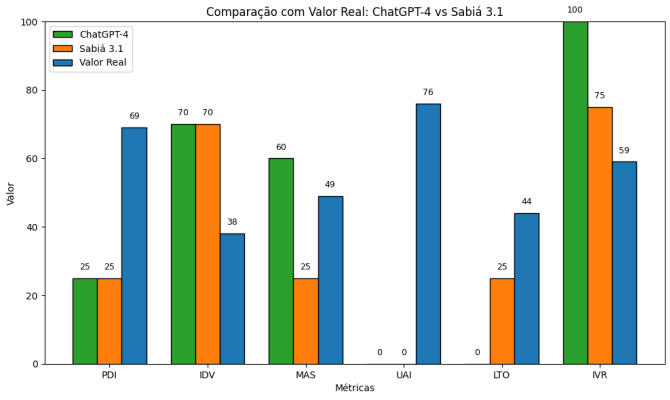


Figura 4. Comparação dos valores calculados para as dimensões de Hofstede com o valor real utilizando gráfico de barras.

A Figura 4 traz uma dificuldade de visualização clara entre o desempenho dos dois modelos. Por isso, foi elaborada uma segunda visualização utilizando um gráfico de radar na Figura 5.

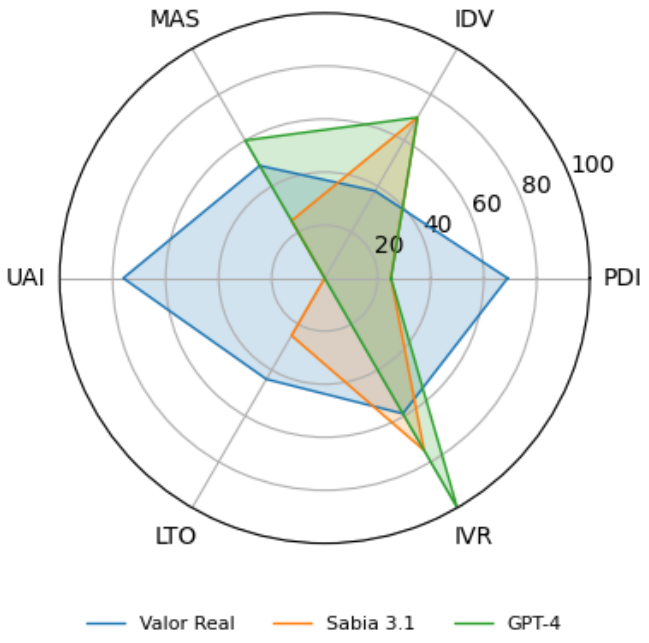


Figura 5. Comparação dos valores calculados para as dimensões de Hofstede com o valor real utilizando gráfico de radar.

A partir da Figura 5, é possível perceber que os dois modelos possuem uma dificuldade em se alinhar culturalmente com a cultura brasileira. No entanto, é necessário identificar qual deles mais se aproxima desse alinhamento. A fim de obter um resultado quantitativo e robusto sobre o desempenho dos modelos, foram utilizadas duas métricas principais: a Raiz do erro quadrático médio (RMSE, do inglês *Root Mean Squared Error*) e a Distância euclidiana. Essas métricas são fundamentais para que seja possível avaliar a proximidade entre os valores das dimensões calculados a partir das respostas dos modelos e os valores reais, possibilitando uma compreensão aprofundada do alinhamento cultural dos modelos.

A primeira métrica considerada foi o RMSE, que foi escolhido por conta da sua ampla utilização para quantificar a magnitude média dos erros. Para este caso, o erro do modelo é calculado por meio da Equação (10):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

Onde n representa o número total de observações, y_i representa o valor real da i-ésima observação e $\hat{y}_i$ representa o valor predito correspondente.

Analisando a Equação (10), é possível perceber que o RMSE é a raiz quadrada da média dos quadrados das diferenças entre os valores reais (y_i) e os valores que foram preditos ($\hat{y}_i$), que, para o nosso caso, são os valores reais das dimensões e os valores das dimensões calculados a partir das respostas dos modelos. Ele foi escolhido por demonstrar de maneira quantitativa os erros, trazendo, assim, uma maneira simples de visualizar falhas graves na percepção de alinhamento.

Para complementar a análise feita com o RMSE, também foi escolhida a distância euclidiana, que traz uma medida global da discrepância entre os vetores com os valores reais e os preditos em um espaço dimensional. Esta distância é dada por:

$$d = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

Assim como no RMSE, os valores reais na Equação (11) são simbolizados com o (y_i) e os valores que foram preditos, pelo ($\hat{y}_i$). Apesar de ambas terem equações parecidas, cada uma tem um objetivo diferente, pois enquanto o RMSE permite a compreensão da média dos erros e traz uma ênfase para os casos mais discrepantes, a distância euclidiana traz uma medida agregada da proximidade das respostas como um todo. Utilizando essas duas métricas em conjunto, é possível quantificar o alinhamento cultural com a cultura brasileira que cada um dos modelos possui.

A Tabela 3 apresenta os valores de RMSE e de distância euclidiana calculados para cada modelo.

O cálculo das métricas apresentadas na Tabela 3 permite determinar qual modelo se adequa melhor culturalmente.

Tabela III
RESULTADOS DOS MODELOS

Métrica	ChatGPT-4	Sabiá 3.1
RMSE	45.60	40.68
Distância Euclidiana	111.69	99.64

V. DISCUSSÃO

Após analisar a tabela 3, é possível perceber que o Sabiá 3.1, quando comparado ao ChatGPT-4, possui valores de erro inferiores, indicando que o Sabiá possui um alinhamento melhor com as dimensões culturais de Hofstede para o Brasil. Esse resultado é interessante, pois demonstra que um modelo com uma especialização nacional, treinado com foco no público e no contexto do Brasil, mesmo que tendo um investimento algumas milhares de vezes menor, possui um resultado melhor que um dos modelos mais robustos da atualidade. Esta diferença entre os modelos mostra que o volume de dados e grandes investimentos em infraestrutura não garante uma adequação cultural, mostrando que é importante uma curadoria de dados e foco em contextos locais.

No entanto, este estudo não está isento de limitações. Esta é uma pesquisa que se baseia na aplicação de um *prompt* específico para a interação com os modelos, o que pode não capturar em sua totalidade a complexidade cultural que está dentro de um modelo. Também é importante ressaltar a complexidade de se realizar a criação de uma "persona" cultural em LLMs, pois os dois modelos escolhidos tiveram grandes dificuldades para se alinharem culturalmente com o Brasil.

Apesar disto, esse trabalho traz uma possível abordagem para a metrificação cultural de um LLM, trazendo um ponto de partida para investigações futuras mais amplas, podendo trazer até mesmo múltiplas personas dentro do LLM, visando representar com maior fidelidade as nuances culturais.

VI. CONCLUSÃO

Foi observado que o LLM Sabiá 3.1 apresentou um alinhamento cultural maior com o Brasil em comparação com o ChatGPT-4, mesmo com um investimento significativamente menor. Essa diferença de desempenho pode ser atribuída ao *fine-tuning* com dados em português brasileiro realizado no Sabiá, o que demonstra a importância da adaptação linguística e cultural nos LLMs.

Para trabalhos futuros, devem-se aplicar métricas adicionais para uma avaliação mais robusta do alinhamento cultural, assim como a análise de outros modelos com *fine-tuning* voltado para o português brasileiro para validar se o comportamento de melhor alinhamento cultural se mantém em diferentes arquiteturas de modelos.

REFERÊNCIAS

[1] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, Carbon Emissions and Large Neural Network Training, arXiv preprint arXiv:2104.10350, 2021.

[2] H. Touvron, L. Martin, K. Stone, T. Scialom, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.

[3] MIHALECEA, R. et al. Why AI Is WEIRD and Shouldn't Be This Way: Towards AI For Everyone, With Everyone, By Everyone. [S.l.]: University of Michigan; Santa Clara University; Universidad de la República; Max Planck Institute; CMU Africa; SUTD; MBZUAI, 2024.

[4] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," *PNAS Nexus*, vol. 3, no. 9, p. pgae346, 2024. doi: 10.1093/pnasnexus/pgae346.

[5] R. I. Masoud, Z. Liu, M. Ferianc, P. Treleaven, and M. Rodrigues, "Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions," *arXiv preprint arXiv:2309.12342*, 2023. [Online]. Available: https://arxiv.org/abs/2309.12342

[6] B. AlKhamissi, M. ElNokrashy, M. AlKhamissi, and M. Diab, "Investigating cultural alignment of large language models," *arXiv preprint arXiv:2402.13231*, 2024. [Online]. Available: https://arxiv.org/abs/2402.13231

[7] Q. Zhong, Y. Yun, and A. Sun, "Cultural Value Differences of LLMs: *prompt*, Language, and Model Size," *arXiv preprint arXiv:2407.16891*, 2024. [Online]. Available: https://arxiv.org/abs/2407.16891

[8] S. H. Schwartz, "An Overview of the Schwartz Theory of Basic Values," Online Readings in Psychology and Culture, vol. 2, no. 1, 2012. doi: 10.9707/2307-0919.1116.

[9] M. Javidan and A. Dastmalchian, "Managerial implications of the GLOBE project: A study of 62 societies," Journal of Management & Organization, vol. 47, no. 1, 2009. doi: 10.1177/103841110809928.

[10] R. Inglehart et al., "World Values Surveys and European Values Surveys, 1981–1984, 1990–1993, and 1995–1997," Ann Arbor, MI: Institute for Social Research, ICPSR version, 2000.

[11] G. Hofstede, "Dimensionalizing cultures: the Hofstede model in context," Online Readings in Psychology and Culture, vol. 2, no. 1, Dec. 2011. Available: https://doi.org/10.9707/2307-0919.1014. [Accessed: May 1, 2025].

[12] HOFSTEDE, G.; MINKOV, M. VSM 2013 : Values Survey Module 2013 – Manual. [s.l.] : Geert Hofstede BV, maio 2013.

[13] HOFSTEDE, G.; HOFSTEDE BV. VSM 2013 : Values Survey Module 2013 – Questionário – Versão Português (Brasil). [s.l.] : Geert Hofstede BV, novembro 2014. Adaptação para o Português (Brasil) por Gabriel Vouga Chueke; Kelly Pavan; Ilan Avrichir.

[14] CORRÊA, N. K.; SEN, A.; FALK, S.; FATIMAH, S. Tucano: advancing neural text generation for Portuguese. *arXiv preprint* arXiv:2411.07854, 2024. Disponível em: <https://arxiv.org/abs/2411.07854>. Acesso em: 01 maio 2025.

[15] EXAME, "Microsoft vai investir mais US$ 10 bi na criadora do ChatGPT," Exame, Jan. 23, 2023. Available: https://exame.com/tecnologia/microsoft-vai-investir-mais-us-10-bi-na-criadora-do-chatgpt/. [Accessed: May 1, 2025].

[16] T. W. Francis, "SoftBank set to invest $40 billion in OpenAI at $260 billion valuation, sources say," CNBC, Feb. 7, 2025. Available: https://www.cnbc.com/2025/02/07/softbank-set-to-invest-40-billion-in-openai-at-260-billion-valuation-sources-say.html. [Accessed: May 1, 2025].

[17] UOL, "Maritaca AI: conheça o "ChatGPT brasileiro"que busca democratizar a IA," Tilt, Jul. 31, 2024. Available: https://www.uol.com.br/tilt/noticias/redacao/2024/07/31/maritaca-ai-chatgpt-brasileiro.htm. [Accessed: May 1, 2025].

[18] S. Saha, S. K. Pandey, H. Gupta, and M. Choudhury, "Reading between the Lines: Can LLMs Identify Cross-Cultural Communication Gaps?," *CoRR*, vol. abs/2502.09636, 2025. [Online]. Available: https://arxiv.org/abs/2502.09636. [Accessed: May 1, 2025].

[19] Hofstede Insights, "Country Comparison Bar Charts," [Online]. Available: https://geerthofstede.com/country-comparison-bar-charts/. [Accessed: May 28, 2025].