

Cesar: Um modelo de recomendação conversacional em português brasileiro

Matheus Rodrigues^{1,*,+} and Willian Oliveira^{1,+}

¹Universidade Federal de Pernambuco, Centro de Informática - CIn, Recife, 50740-560, Brazil

*mrsf@cin.ufpe.br

+these authors contributed equally to this work

ABSTRACT

This paper introduces the Cesar model, a novel language model for conversational recommendation. In a landscape where personalized recommendations play a crucial role in widely-used platforms, such as social media and streaming services, addressing the challenge of efficiently capturing user-specific desires remains paramount. Cesar tackles this gap by employing advanced natural language processing and transfer learning techniques, adapting to the context of movie recommendations in Brazilian Portuguese. Unlike traditional approaches relying on implicit user features like likes and watch time, Cesar can understand and respond to more specific queries, such as preferences for genre, duration, and specific cast members. Furthermore, the model addresses the "Cold Start" challenge by leveraging prior knowledge to provide meaningful recommendations from the user's initial interactions. Developed within the accessible Google Colab environment, Cesar's design facilitates future adaptations for high availability environments. Results indicate that Cesar presents a promising approach to enhance user experience in conversational recommendation systems, offering improved refinement, dynamic interaction, and explainability in generating recommendation lists. This work significantly contributes to expanding the conversational recommendation theme to new languages, advancing the understanding and application of language models in specific contexts, thereby enriching user-platform interactions.

Introdução

Recomendações personalizadas desempenham um papel central em aplicações amplamente utilizadas, como mídias sociais, plataformas de vídeos e comércio online. Do ponto de vista do usuário, essas recomendações são frequentemente apresentadas em listas dinâmicas, ajustando-se à medida que o usuário interage com a plataforma. Em serviços como Netflix ou YouTube, as recomendações evoluem com base em interações do usuário, como tempo de retenção, curtidas e comentários.

Abordagens tradicionais de recomendação tendem a analisar características implícitas do usuário, como preferências expressas por meio de tempo de retenção, curtidas, comentários ou compartilhamento. No entanto, essas características não capturam facilmente desejos específicos do usuário, como a preferência por "assistir a um filme de ação com até 1h30 de duração estrelado por Samuel Jackson". Por outro lado, características explícitas, como preenchimento de formulários de cadastro, podem ser limitadas para gerar informações relevantes. Além disso, o desafio do "Cold Start" pode impactar as previsões nos primeiros momentos de interação do usuário.

Os recentes avanços dentro da área de processamento de linguagem natural conseguiram expor as diversas aplicações possíveis para os large language models (LLM). Os LLM's estão dispostos através de modelos pré-treinados e podem ser utilizados em aplicações de domínio específico tanto diretamente quanto através de técnicas de transfer learning para gerar adaptação do modelo para um contexto específico. No domínio específico de conversação, as aplicações de recomendação tem ser tornado populares devido a possibilidade de lidar com conhecimento prévio do modelos para lidar com problemáticas como extração de características, interação dinâmica do usuário para refinamento de recomendações e explicabilidade sobre lista de recomendações, tópicos que normalmente não são abordados pelos sistemas tradicionais e, quando são, não conseguem agregar valor ao usuário a nível em nível linguagem natural. Casos de usos de sistemas de recomendação conversacionais incluem desde uma interações do tipo "restaurantes parecidos com o restaurante do Ratatouille" até "explique porque essa recomendação é relevante para mim".

Este presente trabalho propõe o modelo Cesar, um modelo de linguagem para recomendação conversacional. Com a proposta de expandir a temática re recomendação conversacional para novas linguas, este trabalho faz uso de técnicas de transfer learning para adaptar o modelo base para português brasileiro em um domínio textual de recomendação de filmes. Além disso, todo o desenvolvimento foi limitado ao ambiente disponível no Google Colab, possibilitando que adaptações futuras de recomendação possam ser executadas em um ambiente de alta disponibilidade.

Metodologia

Este projeto adota uma abordagem metodológica abrangente e sistemática, com o objetivo de desenvolver um sistema de recomendação de filmes com especial ênfase no idioma português brasileiro. A metodologia de desenvolvimento do projeto foi baseada no CRISP-DM, integrando técnicas avançadas de análise de dados e aprendizado de máquina profundo. Os códigos utilizados para o desenvolvimento deste trabalho podem ser acessados em repositório do [GitHub](#)ⁱ.

Para os objetivos desse projeto, foi priorizada a adaptação e a utilização de um dataset específico, o ReDial, que foi traduzido e utilizado para treinar e avaliar nosso modelo de recomendação.

Além disso, foram realizadas experimentações com alguns modelos de linguagem de estado da arte, adaptados para funcionar em ambientes com recursos de GPU limitados. A avaliação de desempenho dos modelos foi uma etapa crucial, utilizando um conjunto de dados de teste específico fracionado do ReDial original e métricas especialmente desenhadas para avaliar sistemas de recomendação.

Nas subseções abaixo serão abordados os seguintes tópicos: Processo de seleção, adaptação e processamento do conjunto de dados; Avaliação das LLMs candidatas; Explicação breve sobre a arquitetura do modelo selecionado; Descrição do processo de treinamento; e Descrição do processo de avaliação desenvolvido.

Dataset

No âmbito do desenvolvimento de sistemas avançados de recomendação de filmes, a escolha do conjunto de dados adequado é fundamental. Entre as opções disponíveis, o conjunto de dados ReDial, que se refere à 'Diálogos de Recomendação', se destaca. Esta coleção é composta por diálogos anotados que são focados na recomendação de filmes entre usuários, fruto de uma colaboração entre instituições de renome, como Polytechnique Montréal, MILA – Quebec AI Institute, Microsoft Research Montréal, HEC Montréal e Element AI.

Embora existam outros datasets relevantes, como o OpenDialkg e o doc2dial, que frequentemente aparecem em estudos de estado da arte, a escolha do ReDial para nosso projeto foi uma decisão estratégica. Devido a limitações de tempo e orçamento, o projeto priorizou o ReDial com base em seu número de citações entre os papers avaliados na nossa exploração inicial. Esta decisão destaca a relevância do ReDial na comunidade acadêmica e sua aplicabilidade ao nosso objetivo de desenvolver um sistema de recomendação de filmes que também é capaz de interações espontâneas e humanizadas.

Uma consideração importante em nosso projeto foi o idioma original do conjunto de dados ReDial, que é em inglês. Dado o foco do nosso sistema de recomendação no idioma Português Brasileiro, foi necessário realizar um extenso processo de tradução e adaptação linguística. Este processo envolveu a tradução de mais de 10 mil conversas completas, mantendo a integridade da estrutura do dataset.

Com base em trabalhos como MMarco¹ e T5², a tradução deste conjunto de dados fez uso da API gratuita do modelo de Linguagem MariTalk mantida pela empresa brasileira Maritaca AI, com o seguinte prompt:

```
Tarefa: Traduza o texto abaixo de forma natural para o Português do Brasil:
```

```
{MENSAGEM_ORIGINAL}
```

```
Lembrete: Não é necessário explicar a tradução, apenas traduza o texto de forma natural.
```

```
Tradução:
```

Conforme trabalhos anteriores, é esperado que a tradução automática gere ruídos. Entretanto, mesmo em processamento contínuo, a tradução levou aproximadamente 400 horas de processamento e seria impeditivo realizar, para este projeto, uma tradução humana. O dataset final deste processo, e que será utilizado no treinamento do nosso modelo, está disponível no [HuggingFace](#)ⁱⁱ.

Modelos de Linguagem

Os modelos de linguagem pré-treinados para fins generativos estão constantemente passando por otimizações para redução do número de parâmetros possibilitando maior número de experimentos e replicações para casos específicos. Técnicas de otimização de LLMs incluem PEFTⁱⁱⁱ, LoRa³ e Lima⁴, que possibilitam que os recursos necessários para treinamento dos modelos sejam reduzidos.

ⁱ<https://github.com/willianfco/GPT-projeto-crs-PT-BR>

ⁱⁱhttps://huggingface.co/datasets/matheusrdgsf/re_dial_ptbr

ⁱⁱⁱ<https://huggingface.co/blog/peft>

Para a seleção de modelos deste trabalho, foram avaliados os modelos dispostos no hub [HuggingFace^{iv}](https://huggingface.co/models) em versões quantizadas e refinadas para conversação, facilitando assim o processo de obtenção dos pesos e carregamento do modelo em ambientes virtuais e com disponibilidade reduzida de GPU. O segundo critério de escolha foi a colocação dos modelos no Leaderboard da Huggingface para benchmarks de atividades conversacionais.

Por fim, entre os modelos restantes, foi realizada uma rodada de inferências com os modelos que apresentavam pesos inferiores à 10gb, que é o máximo possível para conseguir realizar o processo de treinamento em A100 40GB, reservando espaço para batch de 64 amostras e otimizador Adam.

O modelo selecionado foi o [Zephyr 7B Beta quantizado em 8 Bits^v](https://huggingface.co/TheBloke/zephyr-7B-beta-GPTQ/tree/gptq-8bit-actorder_True), derivado do **Mistral 7B** que usa *Sliding Window Attention* (SWA)⁵ e *Grouped-query attention* (GQA)⁶, conseguindo lidar melhor com sequências longas mantendo um tempo reduzido de inferência. Este modelo ocupa aproximadamente 8.17 GB em memória e tem a seguinte configuração:

Fator	Valor
Hidden Size	4096
Attention Heads	32
Hidden Layers	32
Key Value Heads	8
Vocabulário	32000

Table 1. Especificações do modelo selecionado

Popularmente referenciado na categoria de modelos de linguagens compactos, **Zephyr 7B Beta** é uma evolução no campo dos modelos de linguagem, destacando-se por sua eficiência e alinhamento com as intenções do usuário. Utilizando técnicas como *Distilled Supervised Fine-Tuning* (dSFT) e *Distilled Direct Preference Optimization* (dDPO), o modelo não só melhora a precisão em tarefas específicas, mas também responde de maneira mais natural e alinhada a prompts complexos. Este modelo estabeleceu um novo padrão em benchmarks de chat, superando até modelos de 70 bilhões de parâmetros e sem a necessidade de anotações humanas, o que o torna uma escolha ideal para aplicações práticas em processamento de linguagem natural com baixa disponibilidade de hardware⁷.

Finetune

Para realização do refinamento do modelo base, foi utilizado o Trainer, API da Huggingface que possibilita encapsular o treinamento utilizando utilizando Parameter-Efficient Fine-Tuning (PEFT) e Low-Rank Adaptation (LoRa). Além do encapsulamento, a integração com o Hub da Huggingface permite utilizar algumas funções utilitárias para publicação do modelo e parâmetros utilizados no treinamento.

A tarefa de treinamento é a modelagem de linguagem causal. Para calcular a função de custo neste treinamento, é necessário computar a diferença entre a distribuição de probabilidade prevista pelo modelos para a próxima palavra na sequência e a distribuição de probabilidade real. A função de custo utilizada é a Entropia Cruzada. O objetivo do treinamento é a minimização da Loss. Como otimizador, foi utilizado o AdamW [referência AdamW].

Para um determinado passo de tempo t , a Cross-Entropy Loss \mathcal{L}_t pode ser definida como:

$$\mathcal{L}_t = - \sum_{i=1}^V y_{t,i} \log(p_{t,i}) \quad (1)$$

Aqui, V é o tamanho do vocabulário, $y_{t,i}$ é 1 se a próxima palavra for a palavra i e 0 caso contrário, e $p_{t,i}$ é a probabilidade prevista pelo modelo para a próxima palavra ser a palavra i .

A Loss total \mathcal{L} para uma sequência de comprimento T é a média das perdas em todos os passos de tempo:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^T \mathcal{L}_t \quad (2)$$

No contexto de um modelo de linguagem causal, onde X é a sequência de palavras de entrada e Y é a sequência de palavras de saída, o processo de "chunking" é incorporado para melhorar a eficiência do modelo na captura de relações causais em segmentos menores de texto. O "chunking" envolve dividir a sequência de entrada X em pedaços (ou "chunks") significativos,

^{iv}<https://huggingface.co/models>

^vhttps://huggingface.co/TheBloke/zephyr-7B-beta-GPTQ/tree/gptq-8bit-actorder_True

como frases, para que o modelo possa focar em relações causais locais. A predição da próxima palavra \hat{Y}_{t+1} dado o contexto até o tempo t em um determinado "chunk" pode ser formulada como:

$$P(\hat{Y}_{t+1}|X_{\text{chunk}}, \hat{Y}_1, \dots, \hat{Y}_t) = \text{softmax}(f_{\text{LM}}(X_{\text{chunk}}, \hat{Y}_1, \dots, \hat{Y}_t)) \quad (3)$$

Como trata-se do refinamento de um modelo, isto é, existe um processamento de pré-treino que já realizou ajustes nos pesos do modelo, comumente se utiliza o mesmo prompt ou template de treinamento para gerar as amostras de treinamento. Para este trabalho, foi utilizada a mesma estrutura de prompt, com alteração do texto base. Para a amostragem de refinamento, cada conversa do conjunto de treino foi convertida em uma estrutura separada entre instrução, usuário e assistente. Para todo caso, o primeiro id da conversação é considerado o usuário e outro o assistente.

```
<|system|> Você é um chatbot para indicação de filmes.  
Responda de maneira educada sugestões de filmes para os usuários.  
  
<|user|> Olá, como você está? Estou procurando recomendações de filmes.  
  
<|assistant|> Eu estou indo bem. Qual tipo de filmes você gosta?  
  
<|user|> Eu gosto de animações como The Triplets of Belleville (2003)  
e Waking Life (2001).  
  
<|user|> Eu também gosto de Mary and Max (2009).  
  
<|user|> Qualquer coisa artística  
  
<|user|> Você tem alguma recomendação animada que seja um pouco mais  
dramática, como por exemplo A Scanner Darkly (2006)?  
  
<|user|> Eu gosto de comédias, mas prefiro filmes com um  
pouco mais de profundidade.  
  
<|assistant|> Isso é um desafio, mas eu me lembrarei de algo.  
  
<|assistant|> Final Fantasy: The Spirits Within (2001) pode ser uma boa.  
  
<|user|> Ah, parece legal!  
  
<|assistant|> É animado, de ficção científica e tem ação.  
  
<|user|> Legal  
  
<|user|> Cuide-se, abraços!  
  
<|assistant|> Adeus
```

É importante notar que dada a natureza conversacional dos dados, nem sempre as interações ocorrem de maneira alternada entre usuário e assistente. Além disso, durante a etapa de treinamento, as amostras são dispostas através de uma máscara de atenção causal, fazendo com que uma mesma amostra de treinamento contenha passos intermediários de predição. Com a utilização do LoRA³, a etapa de ajuste de pesos é realizada apenas em duas matrizes low-rank, com um número consideravelmente menor de pesos em relação ao modelo original.

O detalhamento completo dos parâmetros podem ser encontrados no card do modelo. Os parâmetros ajustados para o finetuning estão dispostos abaixo, os demais permanecem com valores padrões do modelo base.

Métricas de Avaliação

Como o processo de treinamento desenvolvido neste projeto visa melhorar a capacidade do modelo conversacional tanto no quesito Recomendação de Filmes, quanto no idioma Português do Brasil, optamos pelas seguinte métricas para avaliar os modelos: % Respostas Contendo Filmes, Hits, % Respostas em Português.

Parâmetros	Valor
Bits	8
Learning Rate	5e-07
Epochs	5
Batch Size	64
Max Sequence Length	512

Table 2. Especificações do modelo selecionado

A métrica **% Respostas Contendo Filmes** busca avaliar se a resposta gerada após um contexto em que, de fato, se esperava sugestão de pelo menos 3 filmes, apresentou algum filme no formato esperado (todas as palavras capitalizadas e o ano de lançamento ao final) como os exemplos a seguir:

- A Nightmare on Elm Street 3: Dream Warriors (1987);
- "Billy Madison" (1995);
- It (2017).

A métrica **Hits** busca avaliar se o modelo sugeriu algum dos filmes esperados, isto é, filmes que realmente foram sugeridos na resposta esperada do modelo. Já a métrica **% Respostas em Português** busca avaliar se a quantidade de respostas fornecidas no idioma alvo do treinamento é superior após ajuste fino dos modelos.

Além dessas métricas, também realizamos algumas avaliações qualitativas referentes à geração de respostas aos mesmos Prompts em cada modelo: *"Recomende um filme de ação de 2h com ator Samuel Jackson"*, *"Indique um filme para crianças que gostam de carros e ação"*, *"Qual filme indicado para idosos com preferência de duração curta?"*, *"Qual o melhor filme do mundo?"*

O objetivo dessas perguntas é comparar como os modelos performam semanticamente diante dos mesmos prompts e também avaliar se a geração dos textos condiz com a informação real, por exemplo duração mencionada dos filmes, atores presentes e sinópses mencionadas.

Resultados

Nesta seção, será exposto os resultados obtidos após o processo de ajuste fino do modelo Cesar. Este segmento é essencial, pois reflete o desempenho do modelo em três métricas chave: a porcentagem de respostas contendo recomendações de filmes, a precisão dessas recomendações (Hits@1), e a proporção de respostas fornecidas em português.

Estes indicadores são fundamentais para avaliar o sucesso do treinamento, principalmente em relação à capacidade adquirida em recomendar filmes de forma relevante e de operar eficientemente no idioma Português do Brasil ao compararmos com os modelos base. A tabela a seguir apresenta um resumo dos resultados alcançados pelos modelos Zephyr-7B-beta, Zephyr-7B-gptq-8bits e o nosso, Cesar.

Modelo	%respostas contendo filmes	hits@1	%resposta em português
Zephyr-7B-beta	0.78	0.16	0.71
Zephyr-7B-gptq-8bits	0.78	0.15	0.70
Cesar	0.85	0.1	0.78

Table 3. Resultados das avaliações nos modelos

A análise da tabela revela aspectos interessantes sobre o desempenho adquirido após o ajuste fino do modelo Cesar. Observamos que o modelo se destaca significativamente na métrica de '% Respostas Contendo Filmes' e na proporção de respostas em português, indicando sua eficiência em recomendar filmes e em manter o idioma do treinamento. Por outro lado, o Zephyr-7B-beta mostra superioridade na precisão das recomendações de filmes (Hits@1), sugerindo que o modelo base não quantizado tem uma maior capacidade de alinhar suas sugestões com as expectativas predefinidas. Contudo, ao avaliarmos os filmes sugeridos pelo Cesar, mesmo não sendo exatamente os esperados, eles estão contidos no mesmo gênero cinematográfico e fazem total sentido as sugestões, o que nos indica que a baixa na métrica Hits@1 não é necessariamente um ponto negativo.

Table 4. Respostas para "Recomende um filme de ação de 2h com ator Samuel Jackson"

Modelo	Resposta	Tempo (s)
zephyr 7b beta	Um filme de ação recomendado com Duration de 2 horas e estrela de Samuel L. Jackson é "Avengers: Infinity War" (2018). Nele, Jackson desempenha o papel de Nick Fury, diretor do S.H.I.E.L.D., que reúne os Vingadores para enfrentar o mais poderoso vilão do Universo, o Titã maldito Thanos (interpretado por Josh Brolin). A ação é intensa e os efeitos especiais são espetaculares, garantindo uma experiência cinematográfica emocionante para os fãs de ação.	8,5
zephyr-7B-beta-GPTQ	Uma opção de filme de ação de 2 horas com o reconhecido ator Samuel L. Jackson seria "The Avengers" (2012). Neste filme, Jackson interpreta Nick Fury, o diretor do S.H.I.E.L.D., que reúne uma equipe de super-heróis para combater uma ameaça mundial. O filme contém uma variedade de ações intensas e dramáticas, o que o torna uma excelente escolha para aqueles que procuram um filme de ação com Samuel L. Jackson.	6,3
cesar	Uma recomendação de filme de ação de 2 horas com o proeminente ator Samuel L. Jackson é "The Avengers" (2012). Neste filme, Jackson interpreta o papel de Nick Fury, o diretor do S.H.I.E.L.D., um agência de segurança mundial que reúne um grupo de super-heróis para enfrentar a maior ameaça que a Terra já conheceu. Com um elenco estrelado por atores como Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, Scarlett Johansson e Jeremy Renner, "The Avengers" oferece uma mistura explosiva de ação, drama e humor.	7,8

Table 5. Respostas para "Indique um filme para crianças que gostam de carros e ação"

Modelo	Resposta	Tempo (s)
zephyr 7b beta	Um filme para crianças que gostam de carros e ação é "Cars 3" (2017). Neste filme, o velho amigo de Lightning McQueen, Doc Hudson, aparece em sonhos para ajudá-lo a superar um desafio desesperado em seu caminho para a vitória na grande corrida de Piston Cup. Juntos, eles aprenderão sobre o valor da perseverança e do espírito de competição, enquanto desafiam rivais novos e aprendem a importância de se adaptar ao mundo em constante mudança.	8
zephyr-7B-beta-GPTQ	Um filme que eu recomendo para crianças que gostam de carros e ação é "Cars 3" (2017). Neste filme, o velho e amado personagem de Lightning McQueen enfrenta novos desafios em sua carreira de corrida quando se encontra com um jovem e talentoso piloto chamado Jackson Storm. Juntos, eles se envolverão em uma aventura emocionante e cheia de ação, que inclui corridas em estradas de montanha e o desafio final da "Super Race". O filme é produzido pela Pixar e é considerado uma sequência divertida e ação-packed para os fãs da franquia "Cars".	8,2
cesar	Um filme indicado para crianças que gostam de carros e ação é "Cars 3" (2017). Neste filme, o personagem de Lightning McQueen enfrenta a dificuldade de se adaptar à velocidade e tecnologia das novas gerações de carros, enquanto tenta se classificar para a prestigiosa Cara Rali de Radiador Springs. Com uma mistura de ação, comédia e aventura, o filme é divertido e aprendível para os amantes de carros de qualquer idade.	6,3

Table 6. Respostas para "Qual filme indicado para idosos com preferência de duração curta?"

Modelo	Resposta	Tempo (s)
zephyr 7b beta	Um filme indicado para idosos com preferência de duração curta é "The Silent Patient" (2019), dirigido por Baillie Walsh. O filme dura aproximadamente 1 hora e 43 minutos e conta a história de uma mulher muda de falar que foi acusada de assassinar seu marido. O mistério de sua silêncio e o desejo de descobrir a verdade atrás de seu crime fazem de "The Silent Patient" um filme intrigante e emocionante, com um final surpreendente. Recomendado para aqueles que gostam de mistérios psicológicos e de duração curta.	9,1
zephyr-7B-beta-GPTQ	Uma sugestão de filme indicado para idosos com preferência de duração curta é "The Best Exotic Marigold Hotel" (2011). Este filme britânico-indiano com duração de aproximadamente 124 minutos apresenta um elenco talentoso de atores consagrados como Judi Dench, Bill Nighy e Maggie Smith, que se aposentaram e decidiram se mudar para a Índia para um lance mais barato. A história é divertida, com uma boa dose de comédia e drama, e a duração curta é adequada para aqueles que preferem filmes mais enxugados.	7,4
cesar	Um filme indicado para idosos com preferência de duração curta é "The Best Exotic Marigold Hotel" (2011), com duração de aproximadamente 124 minutos (2 horas e 4 minutos). Este filme conta com um elenco de estrelas consagradas do cinema britânico e indiano, como Judi Dench, Bill Nighy, Maggie Smith e Dev Patel, e tem um ritmo leve e uma trama agradável, com um tema central sobre o envelhecimento e a segunda oportunidade em vida.	6,5

Table 7. Respostas para "Qual o melhor filme do mundo?"

Modelo	Resposta	Tempo (s)
zephyr 7b beta	Ouça bem, por favor. Não há filme melhor do mundo, pois a percepção de beleza e qualidade cinematográfica é subjetiva e varia de pessoa para pessoa. Cada espectador tem sua preferência e gosto, portanto, não existe um filme que seja considerado melhor do mundo em termos universais. Entretanto, há filmes clássicos e aclamados pela crítica que podem ser recomendados como excelentes escolhas de entretenimento, tais como "Cidadão Kane" de Orson Welles, "Apocalypse Now" de Francis Ford Coppola, "2001: Uma Odisseia no Espaço" de Stanley Kubrick, "Casablanca" de Michael Curtiz, "Cinema Paradiso" de Giuseppe Tornatore, "Amélie" de Jean-Pierre Jeunet, entre outros. Espero que isso ajude você a escolher um filme que seja agradável para sua preferência pessoal.	13,9
zephyr-7B-beta-GPTQ	Eu não posso dizer que existe um melhor filme do mundo, pois a preferência pessoal e o gosto são sujeitos. Cada pessoa tem seus próprios critérios para escolher o filme mais favorável. Então, eu sugiro que você especifique o gênero, ano ou outros detalhes para que eu possa fornecer sugestões mais precisas.	4,5
cesar	Meu programa de aprendizado não permite opiniões pessoais, então eu não posso responder que algum filme é o melhor do mundo. A escolha de um filme como o melhor é uma questão pessoal e subjetiva. Eu sugiro que você faça uma pesquisa de filmes baseada em sua preferência de gênero, estilo, atores e outros fatores, e escolha o que melhor se adapta aos seus gostos.	5,3

Discussão

A análise dos resultados dos três modelos – Zephyr-7B-beta, Zephyr-7B-gptq-8bits e Cesar – indica um desempenho geral consistente, com nuances positivas decorrentes do ajuste fino. Notavelmente, o modelo Cesar se destacou na função de recomendar filmes, superando os demais modelos em 7 pontos percentuais na métrica de recomendações padronizadas e na produção de respostas em português. Este avanço sugere uma eficácia aprimorada do Cesar em entender e responder às necessidades específicas dos usuários brasileiros em relação à recomendação de filmes.

Além disso, a utilização de um classificador BERT para a avaliação automática do idioma das respostas mostrou-se uma ferramenta valiosa, garantindo precisão na análise linguística. Porém, ressalta-se a importância de uma verificação manual para complementar e validar esses resultados automatizados, especialmente considerando as nuances e contextos específicos em que o idioma pode ser empregado.

Para futuros aprimoramentos, a construção de um dataset maior e mais diversificado pode ser crucial. Isso não apenas reforçaria as capacidades do modelo em recomendar filmes de forma mais precisa, mas também potencializaria a compreensão e geração de respostas em português, abrangendo uma gama mais ampla de expressões e regionalismos. A ideia é que ao ampliar o escopo do dataset, o modelo poderia se tornar ainda mais adaptável e sensível às variações linguísticas e culturais presentes no Brasil, melhorando assim a experiência do usuário final.

Em conclusão, os resultados obtidos indicam um caminho promissor para o desenvolvimento de modelos de IA especializados em recomendações conversacionais e multilíngues. Com ajustes contínuos e refinamentos direcionados, é possível alcançar uma interação ainda mais eficiente e personalizada com os usuários.

References

1. Bonifacio, L. *et al.* mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897* (2021).
2. Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *The J. Mach. Learn. Res.* **21**, 5485–5551 (2020).
3. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models (2021). [2106.09685](#).
4. Zhou, C. *et al.* Lima: Less is more for alignment (2023). [2305.11206](#).
5. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).
6. Ainslie, J. *et al.* Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245* (2023).
7. Tunstall, L. *et al.* Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944* (2023).

Agradecimentos

Gostaríamos de expressar nossa gratidão à Maritaca AI, uma empresa nacional de Modelos de Linguagem de Grande Escala, pela disponibilização gratuita da API de sua LLM. Essa contribuição foi fundamental para a tradução do dataset em tempo hábil, possibilitando avanços significativos em nosso projeto.

Declaração de Contribuição dos Autores

M.R. e W.O. conceberam e conduziram os experimentos, analisaram os resultados e revisaram o manuscrito em conjunto. Todos os autores contribuíram de maneira igual e colaborativa em cada etapa do processo de pesquisa e desenvolvimento deste projeto.