Evaluation of Entry-Level Open-Source Large Language Models for Information Extraction from Digitized Documents

Francisco Clerton Almeida[1], Carlos Caminha[1,2]

[1] Programa de Pós Graduação em Informática Aplicada - PPGIA, Unifor, Brasil
clertonjradv@edu.unifor.br
[2] Universidade Federal do Ceará, UFC, Brasil
caminha@ufc.br

Abstract. The rise of Large Language Models (LLMs) has transformed the field of natural language processing (NLP), offering a wide range of proprietary and *open-source* models varying significantly in size and complexity, often measured by billions of parameters. While larger models excel in complex tasks like summarization and creative text generation, smaller models are suited for simpler tasks such as document classification and information extraction from unstructured data. This study evaluates *open-source* LLMs, specifically those with 7 to 14 billion parameters, in the task of extracting information from OCR texts of digitized documents. The effectiveness of OCR can be influenced by factors such as skewed images and blurred photos, resulting in unstructured text with various issues. The utility of these models is highlighted in Intelligent Process Automation (IPA), where software robots partially replace humans in validating and extracting information, enhancing efficiency and accuracy. The documents used in this research, provided by a state treasury department in Brazil, comprise personal verification documents. Results show that *open-source* entry-level models perform 18% lower than a cutting-edge proprietary model with trillions of parameters, making them viable free alternatives.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: Natural Language Processing, Large Language Models, Information Extraction.

1. INTRODUÇÃO

A ascensão dos Modelos de Linguagem de Larga Escala, em inglês *Large Language Models* (LLMs), tem revolucionado a área de processamento de linguagem natural (PLN), oferecendo uma vasta gama de modelos proprietários e de código aberto [Minaee et al. 2024]. Esses modelos variam significativamente em termos de tamanho e complexidade, comumente medidos pela quantidade de parâmetros, que frequentemente chegam à casa dos bilhões. Modelos maiores e de código fechado, como GPT, Claude, e Gemini, têm se destacado em tarefas complexas, como sumarização, tradução automática, e geração de texto criativo. Por outro lado, modelos menores são geralmente indicados para tarefas mais simples, como geração de conjuntos de dados sintéticos [Silva et al. 2024], classificação de documentos [Martins and Silva 2021] e extração de informações [Cardoso and Pereira 2020] em dados não estruturados.

Na categoria de modelos com aproximadamente 7 bilhões de parâmetros, encontramos várias famílias notáveis, incluindo Mistral, LLama, Starling, e Zephyr. Alguns desses modelos têm demonstrado desempenho comparável a modelos de código fechado muito maiores em diversas tarefas, oferecendo uma alternativa eficiente e economicamente viável.

A tarefa de extração de informação de documentos digitalizados envolve um passo preliminar crucial: a realização do OCR (Optical Character Recognition). A eficácia do OCR pode variar dependendo da qualidade do algoritmo utilizado e da digitalização do documento. Exemplos de desafios comuns incluem imagens inclinadas, fotos distantes ou borradas. Esses fatores podem resultar em texto não estruturado com problemas, como palavras omitidas, palavras adicionais, ou palavras agrupadas (por exemplo, “palavraa mais” ao invés de “palavra a mais”).

A utilidade de modelos de extração de informação em documentos digitalizados é evidente em tarefas

de Automação Inteligente de Processos [Chakraborti et al. 2020]. Nessas tarefas, os robôs de *software* assumem a responsabilidade de validar e extrair informações (sendo supervisionados por humanos), substituindo a necessidade de intervenção completamente manual das pessoas. Isso não só aumenta a eficiência e a precisão dos processos, mas também libera os recursos humanos para se concentrarem em atividades mais estratégicas e menos repetitivas, promovendo uma maior produtividade e redução de erros humanos. Muitas organizações, tanto privadas quanto públicas, precisam validar documentos submetidos a seus processos internos, e a qualidade de digitalização desses documentos muitas vezes dificulta a análise rápida e eficiente.

Este artigo foca na avaliação de LLMs de código aberto, especificamente aqueles com entre 7 e 14 bilhões de parâmetros, na tarefa de extração de informações a partir de textos OCR de imagens de baixa qualidade de documentos. O objetivo desta tarefa é, dado um documento OCR, extrair um JSON (JavaScript Object Notation) de informações relevantes e estruturadas. Os documentos utilizados nesta pesquisa foram fornecidos pela Secretaria da Fazenda do Estado do Ceará e incluem documentos pessoais comprobatórios. Ao todo, foram processados 326 documentos, de seis tipos diferentes, totalizando 678.294 *tokens*.

Os resultados mostram que os modelos de código aberto de entrada têm desempenho geral 19% inferior (para alguns tipos de documentos houve um empate técnico, com menos de 5% de diferença) a um modelo proprietário de ponta, com trilhões de parâmetros. Isso os posiciona como alternativas gratuitas relevantes para a execução desta tarefa. Esses resultados sugerem a possibilidade interessante de refinar um modelo de código aberto de entrada com previsões de um modelo maior, um processo conhecido como destilação de modelos. Tal abordagem poderia transformar o modelo de entrada em um especialista nessa tarefa, alcançando desempenho similar ao modelo de maior porte.

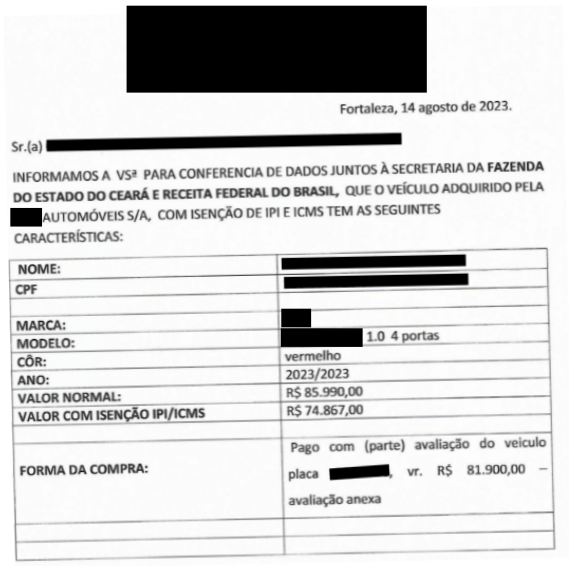
Este artigo está organizado da seguinte forma: na Seção 2, apresentamos uma revisão do estado da arte, discutindo as técnicas iniciais e os avanços recentes em extração de informações, com ênfase no uso de LLMs para essa finalidade. A Seção 3 descreve a metodologia adotada, incluindo detalhes sobre o conjunto de dados utilizado, os modelos de linguagem avaliados, o processo de criação dos *prompts* utilizando *few-shot learning*, o ambiente de execução dos experimentos e a métrica de avaliação empregada. Na Seção 4, apresentamos os resultados obtidos, com uma análise detalhada da acurácia dos modelos para diferentes tipos de documentos. A Seção 5 discute os achados principais, comparando o desempenho dos modelos de código aberto com o modelo proprietário GPT-4o, e explorando as implicações práticas e as possibilidades de refinamento dos modelos. Por fim, a Seção 6 conclui o artigo destacando as principais contribuições, limitações do estudo e sugestões para trabalhos futuros.

2. ESTADO DA ARTE

A extração de informação tem sido um campo essencial dentro da ciência de dados e Processamento de Linguagem Natural (PLN). Inicialmente, técnicas como OCR eram empregadas para converter textos em imagens digitalizadas para texto editável. Simultaneamente, expressões regulares (*regex*) eram amplamente utilizadas para identificar padrões em textos, permitindo a extração de dados estruturados a partir de documentos não estruturados. Um exemplo dessa aplicação é demonstrado no trabalho de Papadopoulos et al., onde eles utilizam *regex* para identificar e extrair informações específicas de grandes corpos científicos [Papadopoulos et al. 2020]. Avanços subsequentes trouxeram o uso de Reconhecimento de Entidades Nomeadas, que permitiu a identificação de nomes de pessoas, lugares e organizações em textos, melhorando a precisão da extração de informação [Weston et al. 2019].

A ascensão dos Modelos de Linguagem de Larga Escala revolucionou o campo da extração de informação. Na prática, LLM de diferentes tamanhos e arquiteturas têm sido aplicados com sucesso em tarefas variadas, como a tradução automática, resposta a perguntas e sumarização de textos. Especificamente para extração de informação, estudos têm demonstrado a eficácia dos LLMs. Por exemplo, Zhang et al. utilizam modelos de linguagem para extrair e localizar informações em documentos

a)



Fortaleza, 14 agosto de 2023.

Sr.(a) [redacted]

INFORMAMOS A VSª PARA CONFERENCIA DE DADOS JUNTOS À SECRETARIA DA **FAZENDA DO ESTADO DO CEARÁ E RECEITA FEDERAL DO BRASIL,** QUE O VEÍCULO ADQUIRIDO PELA [redacted] AUTOMÓVEIS S/A, COM ISENÇÃO DE IPI E ICMS TEM AS SEGUINTES CARACTERÍSTICAS:

NOME:	[redacted]
CPF	[redacted]
MARCA:	[redacted]
MODELO:	[redacted] 1.0 4 portas
CÔR:	vermelho
ANO:	2023/2023
VALOR NORMAL:	R$ 85.990,00
VALOR COM ISENÇÃO IPI/ICMS	R$ 74.867,00
FORMA DA COMPRA:	Pago com (parte) avaliação do veiculo placa [redacted], vr. R$ 81.900,00 – avaliação anexa

b)

********** ********** Fortaleza, 14 agosto de 2023 Sr.(a) **********
- CPF.********** INFORMAMOS A VS PARA CONFERENCIA DE DADOS JUNTOS A SECRETARIA DAFAZENDA DO ESTADO DO CEARA E RECEITA FEDERAL DO BRASIL, QUE O VEICULO ADQUIRIDO TELA ***** AUTOMOVEIS S/A, COM ISENCAO DE IPI E ICMS TEM AS SEGUINTES NOME: ********** CPF ******** MARCA: ********* MODELO: ******** 1.0 4 portas COR: vermelho ANO: 2023/2023 R$ 85.990,00 VALOR NORMAL: VALOR COM ISENCAO IPI/ICMS R$ 74.867,00 Pago com (parte) avaliacAo do veiculo FORMA DA COMPRA: placa **** vr. R$ 81.900,00 avaliacao anexa

c)

{
"Modelo do automovel": " ***** ***** 1.0 4 portas",
"Placa do automovel": " ***",
"Valor normal":"R$ 85.990,00",
"Valor com isencao":"R$ 74.867,00"
}

Figura 1: **Exemplo de documento digitalizado e as informações a serem extraídas.** Em a) é possível observar a imagem do documento, rotacionada e borrada o que dificulta a realização do OCR. Em b) é possível observar o texto detectado pelo algoritmo de OCR, observam-se falhas, como: detecção de caracteres errados; palavras faltantes; e palavras juntas. Em c) é possível observar o JSON que deseja-se extrair com o LLM. Informações pessoais foram omitidas das imagens para preservar a privacidade desses indivíduos.

complexos, enquanto outra pesquisa aplicou um LLM para a síntese de evidências a partir de grandes volumes de dados textuais, mostrando uma significativa melhoria na precisão e eficiência da extração de informações [Zhang et al. 2023; Gartlehner et al. 2023].

Para a extração de informação em documentos digitalizados, o uso de LLMs também tem mostrado avanços promissores. Estudos destacaram como o ajuste fino de LLMs pode melhorar a extração de informações estruturadas de textos científicos complexos, evidenciando o impacto dessas tecnologias na pesquisa e análise de dados [Townsend et al. 2023].

3. METODOLOGIA

3.1 Conjunto de Dados

Os dados utilizados neste estudo consistem em documentos digitalizados enviados pelo contribuinte à Secretaria da Fazenda do Estado do Ceará (SEFAZ-CE) para solicitações de isenção fiscal. Esses documentos passaram por um processo interno de extração de texto utilizando a tecnologia de OCR Tesseract 5.3.1.2. O resultado inclui textos OCR diretamente associados às imagens escaneadas correspondentes. A Figura 1 ilustra a imagem de um dos documentos utilizados nesta pesquisa (Figura 1 a) e o texto detectado automaticamente pela biblioteca de OCR (Figura 1 b).

Seis tipos de documentos foram utilizados, cada um com suas próprias informações e características específicas. A Tabela I ilustra os tipos de documentos, as informações a serem extraídas, quantidade de documentos (N), média de *tokens* por documento (μ_{tokens}) e somatório de *tokens* (Σ_{tokens}). Ao todo, foram processados 326 documentos, de seis tipos diferentes, totalizando 678.294 *tokens*. Os tipos de documentos utilizados nesta pesquisa foram: Comprovante de endereço; Laudo Médico; Requerimento de isenção de ICMS; Requerimento de isenção de IPI; Procuração; e Declaração de Concessionária. Os atributos a serem extraídos foram definidos por auditores da SEFAZ, com base na sua experiência na análise desses tipos de documentos, esses profissionais elencaram os atributos que são mais importantes

para validar os pedidos de isenção fiscal solicitados pelo contribuinte.

Documento	**Informações**	**N**	$\mu_{\textbf{tokens}}$	$\Sigma_{\textbf{tokens}}$
Comp. Endereço	Nome e Endereço	40	2334,45	93.378
Laudo Médico	Nome, CPF e CID	47	2374,19	111.537
Req. ICMS	Nome, Fundamento, Assunto e CPF	85	1750,35	150.530
Req. IPI	Nome, CPF, Data de Transmissão e Protocolo	80	2249,97	179.998
Procuração	Nome e CPF do Outorgante, Nome e CPF do Outorgado	46	1443,25	40.431
Concessionária	Modelo do automóvel, Placa, Valor e Valor com isenção	28	2226,52	102.420
		326	-	**678.294**

Tabela I: Informações a serem extraídas por tipo de documento e contagem de *tokens* dos documentos por tipo.

Foi construída uma coleção de referência a partir da rotulagem de todos os documentos utilizados nesta pesquisa. Auditores da SEFAZ participaram do processo, criando para cada documento um arquivo JSON com as informações que deveriam ser detectadas. Um exemplo dessa rotulagem pode ser observado na Figura 1 c, onde a partir do documento ilustrado na Figura 1 a, foi construído o JSON correspondente. Cada documento foi revisado manualmente para garantir a precisão das informações, resultando em uma coleção de referência para comparação com os resultados dos modelos.

3.2 LLMs utilizados

A seguir, são descritos os LLMs utilizados neste estudo:

—**CapybaraHermes-2.5-Mistral-7B**: É conhecido por seu desempenho equilibrado em tarefas de compreensão e geração de texto. A arquitetura Hermes foca na eficiência computacional, enquanto o Mistral é otimizado para tarefas de NLP complexas.

—**Firefly-Llama2-13B-v1.2**: Firefly é uma versão ajustada do Llama2 com 13 bilhões de parâmetros, projetada para melhor desempenho em tarefas de extração de informações específicas. A versão 1.2 inclui melhorias na precisão e velocidade de processamento, tornando-o adequado para aplicações que exigem respostas rápidas e precisas.

—**Laser-Dolphin-Mixtral-2x7B-DPO**: Uma combinação inovadora de dois modelos de 7 bilhões de parâmetros (Mixtral), utilizando a técnica Dolphin para aprimorar a precisão em tarefas de processamento de linguagem.

—**Llama-2-13B-Chat**: Uma versão específica do Llama-2 projetada para interação conversacional. Com 13 bilhões de parâmetros, é adaptada para gerar respostas mais naturais e contextualizadas, sendo particularmente eficaz em tarefas que requerem entendimento detalhado do contexto.

—**LLama-Pro-8B-Instruct**: Este modelo de 8 bilhões de parâmetros é ajustado para tarefas de instrução e extração de dados, proporcionando um equilíbrio entre desempenho e eficiência. É otimizado para seguir instruções específicas, o que o torna ideal para aplicações estruturadas de extração de informações.

—**Meta-Llama-3-8B-Instruct-hf**: Meta-Llama-3 é a terceira geração da família de LLMs Llama, com 8 bilhões de parâmetros e ajustes para tarefas de instrução.

—**Mistral-7B-Instruct-v0.2**: Uma versão otimizada do modelo Mistral de 7 bilhões de parâmetros, projetada para tarefas de instrução com a versão 0.2 incluindo refinamentos na precisão da extração. Este modelo é eficaz em tarefas que exigem compreensão detalhada e instruções claras.

—**Starling-LM-7B-beta**: O modelo Starling de 7 bilhões de parâmetros é uma versão beta, focada em experimentações para melhorar a capacidade de compreensão e geração de texto em tarefas específicas. Esta versão beta está em constante evolução, incorporando *feedback* de usuários para aprimorar sua eficácia.

Por uma questão de performance, todos os modelos foram utilizados em versões quantizadas em 4 bits utilizando o método AWQ (*Adaptive Weight Quantization*) [Zhang et al. 2022]. A quantização

Modelo	Link Hugging Face	Acesso
Capybara Hermes-2.5-Mistral	https://huggingface.co/TheBloke/CapybaraHermes-2.5-Mistral-7B-AWQ	20/01/2024
Firefly-Llama2	https://huggingface.co/TheBloke/Firefly-Llama2-13B-v1.2-AWQ	20/01/2024
Laser-Dolphin-Mixtral	https://huggingface.co/TheBloke/laser-dolphin-mixtral-2x7b-dpo-AWQ	20/01/2024
Llama-2-13B-Chat	https://huggingface.co/TheBloke/Llama-2-13B-chat-AWQ	20/01/2024
LLama-Pro-8B-Instruct	https://huggingface.co/TheBloke/LLaMA-Pro-8B-Instruct-AWQ	20/01/2024
Meta-Llama-3-8B-Instruct-hf	https://huggingface.co/solidrust/Meta-Llama-3-8B-Instruct-hf-AWQ	20/01/2024
Mistral-7B-Instruct	https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-AWQ	20/01/2024
Starling-LM	https://huggingface.co/solidrust/Starling-LM-7B-beta-AWQ	20/01/2024

Tabela II: Modelos utilizados e link de acesso no *Hugging Face*.

AWQ permite uma execução mais eficiente e rápida dos modelos, mantendo a precisão ao mesmo tempo em que reduz significativamente os requisitos de armazenamento e processamento. A Tabela II apresenta os links, dentro da plataforma *Hugging Face*, para acessar os modelos utilizados. Para efeito de comparação com um modelo topo de linha, foi considerada a utilização do GPT-4o. Deste modo, os resultados da extração com o GPT-4o serão comparados com os resultados dos LLMs de código aberto descritos nesta seção.

3.3 Prompt Elaborado utilizando Few-Shot Learning

O *prompt* desempenha um papel crucial na performance de LLMs em tarefas diversas. A qualidade, clareza e especificidade dos *prompts* influenciam diretamente a precisão e a eficiência dos modelos em realizar a tarefa especificada . A técnica de *few-shot learning* permite que os modelos de linguagem realizem novas tarefas com uma quantidade limitada de exemplos de treinamento. Este método é particularmente útil quando os dados rotulados são escassos, caros de serem obtidos ou tem quantidade de *tokens* superior ao tamanho contexto do LLM, nesse caso não sendo possível incluí-los no *prompt* [Hu et al. 2021].

Utilizamos *few-shot learning* para extrair informações específicas de documentos digitalizados utilizados neste estudo, fornecendo ao modelo dois exemplos de entrada e saída desejados. Esses exemplos foram removidos do conjunto de teste para garantir a validade dos resultados. Destacamos ainda que os exemplos passados foram específicos de cada tipo de documento.

Neste estudo, desenvolvemos um *prompt* detalhado para a extração de informações específicas de textos. A definição completa do *prompt* pode ser encontrada na Figura 2.

Os *prompts* foram projetados para garantir a precisão e a eficiência na extração de informações. A especificação clara de que o nome do requerente deve ser um nome de pessoa (e não de empresa) e que o endereço deve seguir imediatamente o nome ajuda o modelo a focar nas informações mais relevantes e evitar dados irrelevantes. Instruir o modelo a ignorar informações que não se encaixam nas categorias especificadas garante que a saída seja limpa e relevante, evitando ruídos. Exigir a apresentação dos dados em formato JSON com campos específicos ('Nome do Requerente' e 'Endereço do Requerente') assegura que a saída seja estruturada e consistente, facilitando seu uso posterior. A inclusão de uma etapa de verificação assegura que as informações extraídas estejam corretas, permitindo ajustes conforme necessário e aumentando a confiabilidade dos resultados.

Este processo foi aplicado para outros tipos de documentos, ajustando apenas os campos específicos a serem extraídos conforme a necessidade.

3.4 Ambiente de Execução dos Experimentos e Métrica de Avaliação

Os experimentos foram realizados no *Google Colab*, utilizando as bibliotecas *Transformers* (4.41.2), *Accelerate* (0.32.1) e *AutoAWQ* (0.2.5), em uma GPU A100 de 40 GB.

A performance dos modelos foi avaliada pela acurácia, considerando três níveis: documento individual, tipo de documento e total. A acurácia de um documento é a razão entre o número de atributos

Objetivo
Extrair duas informações específicas de um texto: nome do requerente e endereço do requerente. É crucial coletar apenas dados relevantes e precisos.
Instruções Detalhadas

(1) **Extração de Informações**:
—**Nome do Requerente**: Extraia nomes que claramente identifiquem pessoas, ignorando nomes de empresas.
—**Endereço do Requerente**: Identifique o endereço que aparece imediatamente após o nome do requerente.
(2) **Ignorar informações fora do escopo**:
—Ignore informações que não se encaixam nas categorias especificadas.
(3) **Formatação de Saída**:
—Apresente os dados em formato JSON, com campos para "Nome do Requerente" e "Endereço do Requerente". Use "Não Encontrado" para informações que você não encontrar.
(4) **Verificação e Conformidade**:
—Verifique se as informações extraídas estão corretas e ajustadas conforme necessário.

Exemplos de Entrada e Saída
Exemplo de entrada 1:

```
''{OCR_exemplo1}''
```

Exemplo de Saída 1:

```
{JSON_exemplo1}
```

Exemplo de entrada 2:

```
''{OCR_exemplo2}''
```

Exemplo de Saída 2:

```
{JSON_exemplo2}
```

Figura 2: **Definição do *prompt* utilizado para extração das informações.** A informações que aparecem entre chaves representam variáveis python que são substituídas por textos detectados por pacotes de OCR e os seu respectivos JSON de saída da coleção de referência.

corretamente extraídos (N_{corretos}) e o número total de atributos esperados ($N_{\text{esperados}}$):

$$\text{Acurácia}_{\text{doc}} = \frac{N_{\text{corretos}}}{N_{\text{esperados}}} \tag{1}$$

A acurácia para um tipo de documento é a média das acurácias dos documentos desse tipo. A acurácia total é a média das acurácias dos tipos de documentos:

$$\text{Acurácia}_{\text{total}} = \frac{1}{D} \sum_{i=1}^{D} \text{Acurácia}_i \tag{2}$$

Onde D é o número de tipos de documentos.

Atributos adicionais retornados pelo LLM não foram considerados na avaliação. Se o LLM não retornou um JSON válido, a acurácia do documento foi considerada zero.

4. RESULTADOS

Os resultados da avaliação são apresentados na Tabela III, que detalha a acurácia por tipo de documento e a média geral para cada modelo. Os resultados apresentados evidenciam o desempenho de diferentes modelos na tarefa de extração de informações de documentos digitalizados. Conforme esperado, o modelo GPT-4o apresentou resultados superiores a todos os outros modelos, alcançando uma acurácia total de $0,92$. Este desempenho excepcional pode ser atribuído ao fato de o GPT-4o ser um modelo de ponta, com trilhões de parâmetros, o que lhe confere uma capacidade de processamento e entendimento de texto significativamente maior.

Apesar da superioridade do GPT-4o, os modelos de código aberto demonstraram um desempenho notável, reduzindo a diferença para menos de 5% em alguns tipos de documentos. O modelo Laser-Dolphin-Mixtral-2x7B-DPO destacou-se como o melhor modelo de código aberto, obtendo uma acurácia total de $0,74$. Esse resultado representa uma diferença de apenas 18% em comparação ao

Modelo	Endereço	Laudo	ICMS	IPI	Proc.	Conc.	$\text{Acurácia}_{\text{total}}$
Capybara Hermes	0,53 (-0,20)	0,57 (-0,28)	0,94 (-0,06)	0,99 (-0,01)	0,58 (-0,38)	0,73 (-0,10)	0,72 (-0,20)
Firefly-Llama2	0,46 (-0,27)	0,18 (-0,67)	0,97 (-0,03)	0,99 (-0,01)	0,45 (-0,51)	0,66 (-0,17)	0,61 (-0,31)
Laser-Dolphin	0,57 (-0,16)	0,62 (-0,23)	0,97 (-0,03)	0,97 (-0,03)	0,60 (-0,36)	0,73 (-0,10)	0,74 (-0,18)
Llama-2-13B	0,34 (-0,39)	0,15 (-0,70)	0,94 (-0,06)	0,96 (-0,04)	0,39 (-0,57)	0,68 (-0,15)	0,57 (-0,35)
LLaMA-Pro	0,42 (-0,31)	0,31 (-0,54)	0,68 (-0,32)	0,91 (-0,09)	0,43 (-0,53)	0,59 (-0,24)	0,55 (-0,37)
Meta-Llama-3	0,56 (-0,17)	0,46 (-0,39)	0,81 (-0,19)	0,98 (-0,02)	0,70 (-0,26)	0,80 (-0,03)	0,71 (-0,21)
Mistral-7B	0,50 (-0,23)	0,47 (-0,38)	0,95 (-0,05)	0,94 (-0,06)	0,51 (-0,45)	0,77 (-0,06)	0,69 (-0,23)
Starling-LM	0,57 (-0,16)	0,49 (-0,36)	0,97 (-0,03)	0,97 (-0,03)	0,50 (-0,46)	0,75 (-0,08)	0,70 (-0,22)
GPT-4o	**0,73**	**0,85**	**1,00**	**1,00**	**0,96**	**0,83**	**0,92**

Tabela III: Acurácia por tipo de documento e total para cada modelo. Os valores entre parênteses representam a diferença em relação ao modelo de referência GPT-4o.

GPT-4o, mesmo possuindo apenas 14 bilhões de parâmetros, uma fração do tamanho do modelo da Open AI.

5. DISCUSSÃO

Os resultados desta pesquisa demonstram que, para a tarefa de extração de informações em documentos digitalizados, os modelos *open-source* de entrada apresentaram um desempenho comparável ao modelo GPT-4o em diversos tipos de documentos. Embora o GPT-4o tenha obtido a maior acurácia geral, os modelos de código aberto, como o Laser-Dolphin-Mixtral-2x7B-DPO, mostraram uma diferença de desempenho relativamente pequena. Este achado ressalta a viabilidade de utilizar LLMs de código aberto para tarefas de extração de informações em contextos onde os recursos financeiros ou de infraestrutura são limitados, proporcionando uma alternativa eficaz e economicamente viável.

Vale destacar também que os resultados de todos os LLMs poderiam ser aprimorados caso se utilizasse uma métrica de similaridade para comparar os atributos extraídos dos documentos com maior flexibilidade. Devido a possíveis erros introduzidos pelo OCR, como caracteres incorretos, é comum que o erro não seja originado do LLM. A aplicação de uma métrica de similaridade permitiria uma comparação mais robusta e realista dos valores extraídos, mitigando o impacto de pequenos erros e refletindo melhor as necessidades práticas em cenários de automação inteligente de processos. Em muitas aplicações reais, é necessário comparar valores extraídos para realizar correspondências, como verificar se um indivíduo é cônjuge de outro comparando o nome na identidade e na certidão de casamento, onde um caractere incorreto pode ser tolerado.

Por fim, os resultados apresentados neste artigo revelam que há um potencial para ajuste fino e destilação [Gu et al. 2024] de modelos que é significativo. Ajustar um modelo de código aberto utilizando as respostas do GPT-4o poderia especializar o modelo, fazendo com que ele "imitasse" o modelo maior na tarefa específica de extração de informação em documentos digitalizados.

6. CONCLUSÃO

Este estudo avaliou a eficácia de Modelos de Linguagem de Larga Escala de código aberto, com parâmetros variando entre 7 e 14 bilhões, na tarefa de extração de informações de textos OCR oriundos de documentos digitalizados de baixa qualidade. Os resultados indicam que, embora o modelo GPT-4o tenha obtido a melhor performance geral com uma acurácia de $0,92$, modelos de código aberto, como o Laser-Dolphin-Mixtral-2x7B-DPO, mostraram-se competitivos, alcançando uma acurácia de $0,74$. Este desempenho posiciona os modelos *open-source* como alternativas viáveis e econômicas para aplicações em que recursos financeiros ou de infraestrutura são limitados.

É importante destacar a questão da propriedade dos dados utilizados nos experimentos. Os documentos empregados neste estudo contêm informações sensíveis e confidenciais, fornecidas pela Secretaria da Fazenda do Estado do Ceará. Para mitigar riscos de vazamento de dados (data leakage) ao utilizar modelos proprietários como o GPT-4o, realizamos todos os experimentos utilizando a API da

OpenAI. Conforme os termos de uso da OpenAI, os dados enviados via API não são utilizados para treinamento dos modelos, diferentemente dos dados submetidos através da interface do ChatGPT. Assim, asseguramos que as informações sensíveis não seriam armazenadas ou utilizadas para outros fins, mantendo a confidencialidade e a integridade dos dados.

Como trabalhos futuros, propomos a destilação de um modelo de código aberto de entrada utilizando as respostas de um modelo proprietário de alta performance, como o GPT-4o. A destilação pode ser realizada em larga escala, fazendo uso de milhares de inferências (exemplos), diferente dos dados rotulados manualmente nesta pesquisa, que por motivos óbvios têm poucos exemplos. Esse processo, conhecido como aprendizado por destilação [Gu et al. 2024], tem o potencial de transformar um modelo menor em um especialista na tarefa específica de extração de informações de documentos digitalizados.

Por fim, reconhecemos algumas ameaças à validade deste estudo. Primeiramente, o conjunto de dados utilizado, embora realista, é limitado em tamanho e variedade, o que pode afetar a generalização dos resultados para outros contextos ou tipos de documentos. Além disso, a qualidade variável dos textos OCR, decorrente de documentos digitalizados de baixa qualidade, pode introduzir erros que influenciam o desempenho dos modelos, não necessariamente refletindo suas capacidades reais. Por fim, apesar das medidas tomadas para proteger os dados ao utilizar modelos proprietários, dependemos das garantias fornecidas pelos provedores de serviços, o que pode representar um risco residual.

REFERÊNCIAS

Cardoso, B. and Pereira, D. Evaluating an aspect extraction method for opinion mining in the portuguese language. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, pp. 137–144, 2020.

Chakraborti, T., Isahagian, V., Khalaf, R., Khazaeni, Y., Muthusamy, V., Rizk, Y., and Unuvar, M. From robotic process automation to intelligent process automation. In *International Conference on Business Process Management*. Springer, pp. 215–228, 2020.

Gartlehner, G., Kahwati, L., Hilscher, R., Thomas, I., Kugley, S., Crotty, K., Viswanathan, M., Nussbaumer-Streit, B., Booth, G., Erskine, N., Konet, A., and Chew, R. Data Extraction for Evidence Synthesis Using a Large Language Model: A Proof-of-Concept Study. *medRxiv*, October, 2023. [Online]. Disponível em: `https://doi.org/10.1101/2023.10.02.23296415`.

Gu, Y., Dong, L., Wei, F., and Huang, M. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Hu, E. J., Shi, L., Squadrato, R., Tay, Y., Ruder, S., and Raffel, C. Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2021. [Online]. Disponível em: `https://arxiv.org/abs/2106.09685`.

Martins, V. and Silva, C. Text classification in law area: a systematic review. In *Anais do IX Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, pp. 33–40, 2021.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Papadopoulos, D., Papadakis, N., and Litke, A. A Methodology for Open Information Extraction and Representation from Large Scientific Corpora: The CORD-19 Data Exploration Use Case. *Applied Sciences* 10 (16): 5630, 2020.

Silva, M. d. L. M., Mendonça, A. L. C., Neto, E. R. D., Chaves, I. C., Caminha, C., Brito, F. T., Farias, V. A. E., and Machado, J. C. Facto dataset: A dataset of user reports for faulty computer components. In *Anais do VI Dataset Showcase Workshop*. SBC, pp. 1–12, 2024.

Townsend, V., Xie, D., Huang, P., and Cole, L. Structured Information Extraction from Complex Scientific Text with Fine-Tuned Large Language Models. *arXiv preprint arXiv:2212.05238*, December, 2023. [Online]. Disponível em: `https://arxiv.org/abs/2212.05238`.

Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K. A., Ceder, G., and Jain, A. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling* 59 (9): 3692–3702, 2019.

Zhang, X., Wang, Y., Xu, Y., and Zhang, J. Adaptive Weight Quantization for Efficient Neural Network Inference. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Zhang, Y., Kumar, S., Singh, D., and Jain, A. LMDX: Language Model-based Document Information Extraction and Localization. arXiv preprint arXiv:2303.01234, 2023.