INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS FLORIANO
CURSO TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

ITALO RHAFFAEL FERREIRA DA PAZ

INTELIGÊNCIA ARTIFICIAL APLICADA AO PROCESSAMENTO DE EXTRATOS DE CONTRATOS DE INOVAÇÃO ABERTA: Uma comparação de APIs

FLORIANO
2025

ITALO RHAFFAEL FERREIRA DA PAZ

INTELIGÊNCIA ARTIFICIAL APLICADA AO PROCESSAMENTO DE EXTRATOS DE CONTRATOS DE INOVAÇÃO ABERTA: Uma comparação de APIs

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Floriano.

Orientador(a): Prof. Dr. Rafael Angelo Santos Leite.

FLORIANO
2025

INTELIGÊNCIA ARTIFICIAL APLICADA AO PROCESSAMENTO DE EXTRATOS DE CONTRATOS DE INOVAÇÃO ABERTA: Uma comparação de APIs

Italo Rhaffael Ferreira da Paz[1]
Rafael Angelo Santos Leite[2]

RESUMO

Este trabalho avalia a aplicação de APIs de inteligência artificial no processamento e classificação de extratos de contratos de inovação aberta, uma área crucial para a colaboração entre a academia e a indústria. Foram analisadas as APIs ChatGPT (OpenAI), LLaMA (Meta) e Mistral (Mistral AI), utilizando métricas como acurácia, precisão, recall e F1-Score. A metodologia incluiu implementação prática das APIs, criação de gabarito e comparação quantitativa de desempenho. Os resultados destacam as potencialidades e limitações de cada ferramenta, contribuindo para o desenvolvimento de soluções automatizadas para a gestão de extratos de contratos nesse contexto. Os resultados mostram que para cenários que exigem alta Precisão (evitar falsos positivos), o LLama pode ser a melhor escolha. Para cenários que requerem maior equilíbrio geral, o ChatGPT se destaca. O Mistra**l** pode ser útil em aplicações que priorizam consistência e eficiência computacional. Esses resultados contribuem com um panorama sobre a adequação de diferentes modelos de linguagem às necessidades específicas desse domínio, orientando a escolha de APIs, conforme o cenário desejado. Além disso, os achados podem auxiliar no desenvolvimento de metodologias mais eficazes e na implementação prática dessas tecnologias em plataformas como a ipdelve.com.

Palavras-chave: inovação aberta; inteligência artificial; processamento de linguagem natural; extratos de contratos legais; comparação de APIs.

ABSTRACT

This study evaluates the application of artificial intelligence APIs in processing and classifying extracts from open innovation contracts, a crucial area for collaboration between academia and industry. The analyzed APIs include ChatGPT (OpenAI), LLaMA (Meta), and Mistral (Mistral AI), using metrics such as accuracy, precision, recall, and F1-Score. The methodology involved the practical implementation of the APIs, the creation of a benchmark, and a quantitative performance comparison. The results highlight the strengths and limitations of each tool, contributing to the development of automated solutions for managing contract extracts in this context.The findings indicate that for scenarios requiring high Precision (avoiding false positives), LLaMA may be the best choice. For scenarios that demand greater overall balance, ChatGPT stands out. Mistral may be useful in applications prioritizing consistency and computational efficiency. These results provide an overview of the suitability of different language models for the specific needs of this domain, guiding API selection based on the desired scenario.

[1] Graduando do Curso de Análise e Desenvolvimento de Sistemas IFPI Campus Floriano. devrhafael@outlook.com.
[2] Doutor em Ciência da Propriedade Intelectual, professor do IFPI Campus Floriano. rafaelangelo@ifpi.edu.br.

Furthermore, the findings can assist in developing more effective methodologies and the practical implementation of these technologies on platforms such as ipdelve.com.

Keywords: open innovation; artificial intelligence; natural language processing; legal contract extracts; API comparison.

Data de aprovação: 10/01/2025

1 INTRODUÇÃO

Nos últimos anos, a colaboração entre a academia e a indústria tem se intensificado, resultando em uma vasta gama de contratos de inovação aberta (Bigliardi e Filippelli, 2022; Jantunen e Hynninen, 2023; Samanta, Talapatra e Golui, 2022). Esses contratos, que formalizam parcerias estratégicas para o desenvolvimento de novas tecnologias e soluções, contêm extratos que necessitam de uma classificação precisa para facilitar a análise, o gerenciamento e a tomada de decisões. A classificação eficiente desses documentos é essencial para organizar essas informações e compartilhar esse conhecimento com os interessados em monitorar o comportamento das instituições envolvidas nessas interações.

Com o avanço das tecnologias de inteligência artificial (IA), novas ferramentas têm sido desenvolvidas para automatizar e aprimorar tarefas complexas de processamento de linguagem natural (PLN) como a classificação de textos. Entre essas ferramentas, destacam-se as Interfaces de Programação de Aplicações (APIs) de IA que oferecem soluções poderosas e acessíveis para desenvolvedores e pesquisadores.

Três das principais APIs de IA disponíveis atualmente para desenvolvedores são a API do ChatGPT, desenvolvida pela OpenAI (OpenAI, 2024), a API Llama desenvolvida pela Meta (Meta, 2024), a API Mistral desenvolvida pela MistralAI (MistralAI, 2024). Estas APIs são amplamente utilizadas devido à sua capacidade de compreender e gerar texto de maneira natural e coerente. No entanto, a eficácia dessas ferramentas pode variar dependendo da aplicação específica e do contexto dos dados.

Estudos recentes investigaram o desempenho de várias APIs de IA, destacando diferenças com base na aplicação e contexto. Sarmah e Chimalakonda (2022) examinaram o papel do conhecimento de APIs na melhoria da sumarização automática de código usando arquiteturas baseadas em transformadores. Descobriram que, embora o conhecimento de APIs seja benéfico, ele não supera as abordagens de

transformadores de última geração (Sarmah e Chimalakonda, 2022). Sunwoo et al. (2022) focaram no desempenho de modelos de IA para reconhecimento de resíduos de construção, revelando que os modelos treinados em conjuntos de dados rotulados por não-profissionais superaram aqueles rotulados por profissionais, devido ao melhor alinhamento com as características dos dados de treinamento (Sunwoo *et al.*, 2022). Nam et al. (2023) destacaram a eficácia das técnicas de aprendizado de máquina na melhoria da descoberta de conhecimento de APIs em plataformas como Stack Overflow, mostrando uma melhora significativa na compreensão dos desenvolvedores sobre métodos comparáveis de API (Nam *et al.*, 2023). Por fim, Lee et al. (2022) compararam APIs de reconhecimento de fala comerciais em ambientes ruidosos, com a API do Google superando as demais em termos de taxas de erro (Lee *et al.*, 2022). Porém ainda não temos estudos que façam comparação dessas APIs para analisar texto de extratos de contrato de inovação aberta publicados no Diário Oficial da União.

Neste contexto, o objetivo geral deste trabalho é realizar uma análise comparativa do desempenho das APIs do ChatGPT (OpenAI), do Llama (Meta) e do Mistral (MistralAI) na classificação de extratos de contratos de inovação aberta. Este estudo visa não apenas avaliar a precisão e a eficiência dessas ferramentas, mas também explorar sua aplicabilidade prática no contexto de análise de contratos, contribuindo para a escolha da melhor solução de IA para essa tarefa específica.

Este estudo justifica-se por definir as melhores APIs e os seus modelos para serem implementados em plataformas de notificação onde os interessados podem monitorar seus concorrentes em projetos de inovação aberta com a academia. Por exemplo, uma indústria farmacêutica pode monitorar um concorrente que esteja interagindo com uma determinada universidade em acordos de parcerias para desenvolvimento de novos medicamentos, e assim tomar decisões estratégicas. Da mesma forma gestores de inovação das Universidades em seus Escritórios de Transferência de tecnologia podem acompanhar outros escritórios e aprender por meio de *benchmarking*.

A escolha das APIs avaliadas neste estudo — ChatGPT (OpenAI), LLaMA (Meta) e Mistral (Mistral AI) — baseou-se em critérios técnicos, operacionais e estratégicos, considerando sua relevância no cenário atual de modelos de linguagem e aplicações em PLN, enquanto outras ferramentas, como Claude (Anthropic) e Gemini (Google), foram excluídas devido a limitações práticas na implementação. O ChatGPT foi selecionado por sua consolidação no mercado, eficiência em classificação textual e robustez no

suporte ao português; a LLaMA, por representar um modelo aberto da Meta com eficiência computacional e relevância acadêmica; e o Mistral, por ser uma solução emergente com bom custo-benefício e desempenho em modelos compactos. Claude e Gemini foram descartados devido a problemas de autenticação, compatibilidade com português e instabilidade, comprometendo a comparabilidade dos resultados. A seleção priorizou diversidade tecnológica, viabilidade de implementação e relevância acadêmico-mercadológica, permitindo uma análise equilibrada entre modelos consolidados e emergentes para aplicação em classificação de contratos jurídicos.

2 REFERENCIAL TEÓRICO

2.1 INOVAÇÃO ABERTA E OS EXTRATOS DE CONTRATOS DE INOVAÇÃO ENTRE A ACADEMIA E A INDÚSTRIA.

A inovação aberta é uma abordagem estratégica que ganhou destaque nos últimos anos, enfatizando a colaboração além dos limites organizacionais para aumentar o potencial de inovação. Essa abordagem envolve o fluxo de conhecimento e tecnologia interna e externamente, acelerando o processo de inovação e ajudando as empresas a entrar em novos mercados (Kock e Torkkeli, 2008).

A indústria está adotando práticas de inovação aberta para aumentar suas capacidades de inovação, integrando fontes externas de conhecimento e criatividade (Mazzocchi, 2004) e como são processos de fora para dentro e de dentro para fora, isso exige que as organizações estabeleçam laços mais estreitos com o ambiente externo para gerar resultados de inovação bem-sucedidos (Bort *et al.*, 2022).

A inovação aberta entre a academia e a indústria envolve vários tipos de contratos que evoluíram ao longo do tempo. Tradicionalmente, os contratos de colaboração universidade-indústria (UIC) eram de pequena escala e de curto prazo (Kuwashima, 2018), mas uma mudança para contratos abrangentes, de longo prazo e de grande escala foi observada nos últimos anos.

Conforme previsões da Lei de inovação e suas alterações (Brasil., 2004; Brasil, 2016; Rauen, 2016), os contratos de inovação aberta entre a academia e a indústria podem ser: acordos de parcerias e/ou termos de cooperação, licenciamento e/ ou transferência de tecnologias, serviço técnico especializado, encomendas tecnológicas, partilhamento de titularidade, cessão de uso, entre outros.

Acordos de parceria são contratos celebrados entre a ICT e as instituições públicas e privadas para realização de atividades conjuntas de pesquisa científica e tecnológica e de desenvolvimento de tecnologia, produto, serviço ou processo, conforme art. 9º da Lei de Inovação. Além disso, o inciso II, parágrafo 6 do artigo 19 da mesma lei, prevê a constituição de parcerias estratégicas e desenvolvimento de projetos de cooperação entre ICT e empresas e entre empresas, em atividades de pesquisa e desenvolvimento, que tenham por objetivo a geração de produtos, serviços e processos inovadores (Brasil, 2004).

Um exemplo de acordo de parceria entre uma ICT e uma empresa farmacêutica seria desenvolver um novo medicamento. A ICT contribui com pesquisas avançadas e testes laboratoriais, enquanto a empresa oferece recursos financeiros e infraestrutura para produção em escala.

Licenciamentos e/ou transferências de tecnologias são contratos entre uma ICT e outra organização pública ou privada para concessão de direito de uso ou de exploração de criação por ela desenvolvida isoladamente ou por meio de parceria, conforme art. 6º da Lei de Inovação (Brasil, 2004).

Um exemplo de licenciamento e/ou transferências de tecnologia é a concessão de licença de uma tecnologia de energia renovável desenvolvida por uma ICT para uma empresa de energia. A empresa recebe o direito de utilizar e comercializar a tecnologia, enquanto a ICT obtém *royalties* e reconhecimento por sua inovação.

Os Serviços Técnicos Especializados são contratos onde as ICT prestam a outras instituições públicas ou privadas serviços técnicos especializados voltadas à inovação e à pesquisa científica e tecnológica no ambiente produtivo, visando, entre outros objetivos, à maior competitividade das empresas (Art. 8º da Lei de Inovação) (Brasil, 2004).

Um exemplo de serviços técnicos especializados é uma ICT que presta serviços de análise de dados genômicos para uma empresa de biotecnologia. A ICT utiliza sua *expertise* em bioinformática para analisar grandes volumes de dados, auxiliando a empresa na identificação de novos alvos terapêuticos, aumentando sua competitividade no desenvolvimento de medicamentos inovadores.

A *Encomenda tecnológica* é um contrato onde órgãos ou entidades da administração pública, em matéria de interesse público, encomendam uma solução de um problema técnico específico ou um produto, serviço ou processo inovador. Essa encomenda pode ser feita diretamente para outra ICT, entidades de direito privado sem

fins lucrativos ou empresas focadas em atividades de pesquisa e de reconhecida capacitação tecnológica no setor (Art. 20 da lei de Inovação) (Brasil, 2004).

Um exemplo de encomenda tecnológica é o Ministério da Saúde encomendando a uma ICT o desenvolvimento de um novo sistema de diagnóstico rápido para doenças infecciosas. A ICT trabalha na criação e teste do sistema, entregando uma solução inovadora que melhora a detecção precoce e o tratamento, atendendo a uma necessidade de saúde pública.

A *partilha de titularidade* são contratos onde as ICT e outras organizações parceiras no desenvolvimento de determinada tecnologia, prevêem a titularidade da propriedade intelectual e a participação nos resultados da exploração das criações resultantes da parceria, assegurando aos assinantes o direito à exploração, ao licenciamento e à transferência de tecnologia (§2º do art. 9 da Lei de Inovação) (Brasil, 2004).

Um exemplo de partilha de titularidade pode ser uma universidade e uma empresa de biotecnologia quando desenvolvem uma nova vacina de forma colaborativa. Ambos concordam que a propriedade intelectual será dividida igualmente, com a universidade contribuindo com a pesquisa básica e a empresa fornecendo recursos financeiros e de infraestrutura. Assim, ambos têm direito à exploração, licenciamento e transferência da tecnologia desenvolvida.

Cessão de uso são contratos onde a ICT compartilha e permite o uso por terceiros de seus laboratórios, equipamentos, recursos humanos e capital intelectual (Art. 15-A, Inciso IV da Lei de Inovação). Esses terceiros podem ser pessoas físicas ou empresas voltadas a atividades de pesquisa, desenvolvimento e inovação, desde que tal permissão não interfira diretamente em sua atividade-fim nem com ela conflite (Art. 4, Incisos I e II, da Lei de inovação) (Brasil, 2004).

Um exemplo de cessão de uso ocorre quando uma ICT permite que uma startup de tecnologia utilize seus laboratórios e equipamentos para desenvolver um novo software de inteligência artificial. A startup pode usar esses recursos para avançar suas pesquisas, desde que isso não interfira nas atividades principais da ICT ou conflite com seus objetivos institucionais.

Um dos principais locais onde esses dados estão disponibilizados é o Diário Oficial da União (DOU) que é uma fonte oficial e confiável para a publicação de contratos e acordos realizados pelo governo federal, incluindo colaborações entre a academia e a indústria (Imprensa Nacional, 2024).

2.2 APIS UTILIZADAS NO ESTUDO

A avaliação das APIs foram realizadas de forma independente, sendo que para cada API foi aplicada individualmente à tarefa de classificação de contratos jurídicos, e seus resultados foram comparados com base em métricas como acurácia, precisão, recall e F1-Score. Esse estudo buscou destacar as forças e limitações de cada tecnologia no processamento de documentos legais, identificando suas capacidades específicas de aplicação prática para automatizar tarefas complexas e desafiadoras, como a classificação de contratos de inovação aberta.

Este trabalho avaliou o desempenho de três APIs de inteligência artificial especializadas em PLN: ChatGPT (OpenAI), LLaMA (Meta) e Mistral (Mistral AI). Essas APIs foram escolhidas devido à sua ampla aplicação em tarefas de classificação textual e à disponibilidade para desenvolvedores em ambientes controlados de teste. Além disso, cada uma dessas ferramentas representa uma abordagem distinta de arquitetura e treinamento de modelos de linguagem, permitindo uma comparação diversificada de seus desempenhos. Outras APIs, como Claude e Gemini, foram descartadas nesta análise devido a limitações de acesso ou por dificuldades de encontrar suporte na replicação da ferramenta em uma área de desenvolvimento.

O ChatGPT é um Large Language Model (LLM) desenvolvido pela OpenAI, o ChatGPT é uma das APIs de PLN mais avançadas e amplamente utilizadas. Baseado na arquitetura GPT-4, ele é capaz de compreender e gerar texto de forma altamente contextualizada. Sua flexibilidade permite a aplicação em tarefas diversas, incluindo classificação de texto, análise de sentimentos e geração de resumos (“OpenAI”, 2024).

O LLaMA é um LLM que foi projetado pela Meta para fornecer modelos de linguagem eficientes e otimizados para pesquisa. Apesar de menos voltada para o uso comercial direto, ela apresenta um desempenho sólido em tarefas de classificação textual e análise semântica (“Llama”, 2024).

O Mistral é um LLM recente desenvolvido pela MistralAI voltado para o desenvolvimento de modelos de linguagem compactos, mas poderosos. Seu foco está na eficiência computacional, proporcionando resultados precisos em tempo reduzido. No estudo, foi utilizada para categorizar os contratos com base em sua arquitetura adaptada a tarefas específicas (Mistral, [s.d.]).

O Processamento de Linguagem Natural (PLN) é um subcampo da inteligência artificial (IA) que busca capacitar os computadores a entender, interpretar e gerar linguagem humana de forma significativa. Baseado em avanços em aprendizado de máquina e redes neurais, o PLN tem revolucionado áreas como análise de texto, tradução automática, e classificação de documentos. Ferramentas de PLN são particularmente úteis para lidar com grandes volumes de dados textuais, permitindo a automação de tarefas antes realizadas manualmente, como a categorização de contratos legais (Jurafsky e Martin, 2024).

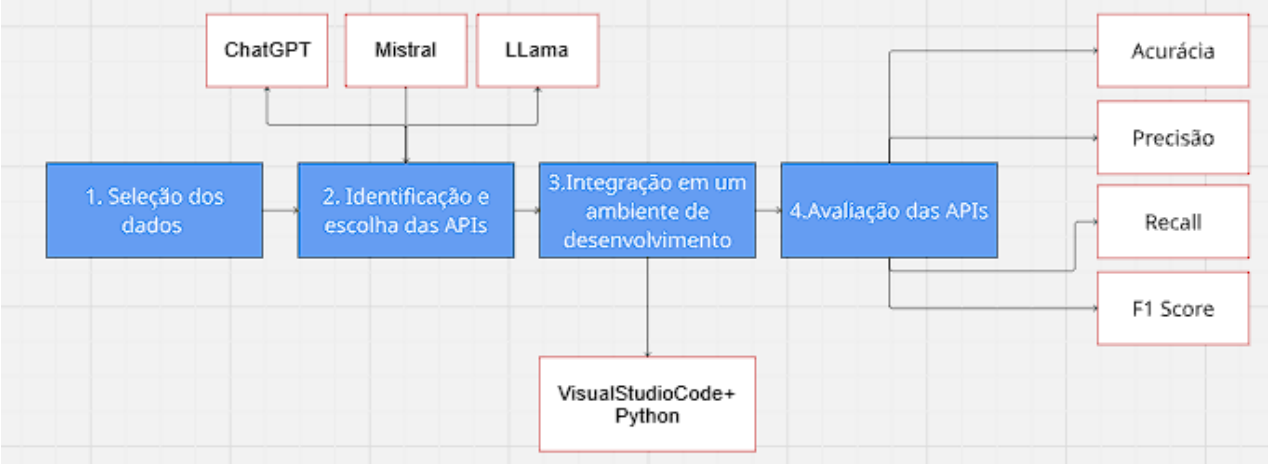
No contexto jurídico, o PLN se apresenta como uma solução promissora para enfrentar desafios relacionados à análise e organização de documentos complexos. Contratos, especialmente no campo da inovação aberta, possuem terminologias técnicas e cláusulas detalhadas que exigem interpretação precisa. As técnicas modernas de PLN, como modelos de linguagem de grande escala, são capazes de capturar nuances textuais e padrões semânticos, otimizando a análise e reduzindo o tempo necessário para tais atividades (Buchanan, 2011).

PLN utiliza redes neurais e aprendizado de máquina para interpretar a linguagem humana. Essa tecnologia é ideal para automatizar tarefas textuais, como classificação de contratos, reduzindo erros humanos e otimizando processos.

3 METODOLOGIA

Essa pesquisa é classificada como aplicada e exploratória com abordagem experimental. Aplicada porque tem como principal objetivo resolver um problema prático, ou seja, avaliar a eficácia de APIs no processamento e classificação de extratos de contratos de inovação aberta. Exploratória porque explora o desempenho das APIs selecionadas, levantando informações e compreendendo como essas ferramentas se comportam diante de um problema específico. E experimental porque envolve a configuração e testes das APIs em um ambiente controlado (Visual Studio Code), com a coleta de métricas quantitativas (Acurácia, Precisão, Recall e F1 Score) para avaliar comparativamente o desempenho de cada solução (Creswell e Plano Clark, 2017). A figura 1 detalha o percurso metodológico do estudo.

Figura 1- Percurso Metodológico



Fonte: Autores (2024)

A primeira etapa consistiu na seleção de 200 links da base de dados extraída a partir do trabalho de Reis (2024), que coletou extratos de publicações do Diário Oficial da União (DOU). A partir desse trabalho foi feito um script para a coleta desses 200 links dessa base de dados, gerando um arquivo csv contendo 200 extratos de contrato de inovação aberta.

A segunda etapa do estudo foi a identificação e escolha das APIs mais relevantes e reconhecidas no mercado de inteligência artificial e processamento de linguagem natural. Foram selecionadas três APIs principais: ChatGPT, Llama e Mistral. A seleção foi baseada na relevância de cada API em suas respectivas áreas de aplicação, no reconhecimento das empresas desenvolvedoras como líderes no campo da IA e em sua facilidade de uso.

A terceira etapa envolveu a integração das APIs em um ambiente de desenvolvimento apropriado, *Visual Studio Code* (Microsoft, 2024). Este ambiente foi configurado para suportar as especificidades de cada API, permitindo a programação na linguagem *Python* e configurações necessárias para que as APIs pudessem processar e classificar extratos de contratos de inovação aberta de maneira eficaz. Esta fase incluiu a instalação de bibliotecas e ferramentas necessárias, bem como a escrita de código para conectar e utilizar cada API de forma otimizada.

A quarta etapa com as APIs integradas e configuradas, o desempenho de cada uma foi avaliado através de métricas específicas de avaliação de desempenho. As métricas utilizadas foram: Acurácia, Precisão, Recall e F1 Score. Cada uma dessas métricas oferece uma perspectiva diferente sobre a eficácia da API em processar e classificar os dados fornecidos. A seção 3.1 detalha essas métricas e suas respectivas

fórmulas.

3.1 MÉTRICAS DE AVALIAÇÃO

A avaliação de modelos de inteligência artificial na classificação de extratos de contratos jurídicos foi realizada com base em quatro métricas fundamentais: Acurácia, Precisão, Recall e F1-Score. Essas métricas fornecem uma visão abrangente do desempenho das APIs utilizadas (ChatGPT, LLaMA e Mistral), permitindo identificar suas forças e limitações.

Quadro 1 - Tabela das Métricas

Métricas de Avaliação	**Definição**	**Fórmulas**
Acurácia	A acurácia mede a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões (Bishop, 2006).	$\frac{Nu´mero\ de\ Previsões\ Corretas}{Número\ Total\ de\ Previsões}$
Precisão	A precisão é a proporção de previsões positivas que são realmente corretas. Mede a qualidade das previsões positivas do modelo (Burzykowski *et al.*, 2023).	$\frac{Verdadeiros\ Positivos\ (VP)}{Verdadeiros\ Positivos\ (VP)+Falsos\ Positivos\ (FP)}$
Recall	O recall mede a proporção de itens positivos reais que foram corretamente identificados pelo modelo. Mede a capacidade do modelo de encontrar todos os exemplos positivos (Burzykowski *et al.*, 2023).	$\frac{Verdadeiros\ Positivos\ (VP)}{Verdadeiros\ Positivos\ (VP)+Falsos\ Negativos\ (FN)}$

F1-Score	O F1-Score é a média harmônica entre precisão e recall Considera tanto a precisão quanto o recall e é útil para balancear as duas métricas (Burzykowski *et al.*, 2023).	$F1 - Score = 2\times \frac{Precisão \times Recall}{Precisão + Recall}$

Fonte: Elaborado pelos autores (2024).

Essas métricas foram utilizadas neste estudo para avaliar de forma quantitativa o desempenho das APIs de IA na classificação de extratos de contratos legais. Combinadas, elas fornecem uma visão completa da eficácia e das limitações de cada ferramenta em cenários complexos e específicos, como contratos de inovação aberta.

3.2 A ESCOLHA DAS APIS

A escolha das APIs avaliadas neste estudo — ChatGPT (OpenAI), LLaMA (Meta) e Mistral (Mistral AI) — foi baseada em critérios técnicos, operacionais e estratégicos, considerando sua relevância no cenário atual de modelos de linguagem e suas aplicações em Processamento de Linguagem Natural (PLN). Outras ferramentas, como Claude (Anthropic) e Gemini (Google), foram excluídas devido a limitações práticas durante a fase de implementação. A seguir, detalhamos os motivos para essa seleção:

O ChatGPT foi incluído por ser o modelo mais consolidado e amplamente utilizado no mercado, com reconhecida eficiência em tarefas de classificação textual. Sua arquitetura, baseada no GPT-4, oferece alta capacidade de compreensão contextual e suporte robusto ao português, sendo uma referência para comparação. Além disso, sua API é bem documentada e de fácil integração, o que facilitou a execução dos testes.

O LLaMA foi selecionado por representar o esforço de uma das maiores empresas de tecnologia (Meta) no desenvolvimento de modelos de código aberto. Apesar de exigir maior configuração para implantação, sua arquitetura é otimizada para eficiência computacional, sendo uma alternativa viável para aplicações que demandam balanceamento entre desempenho e custo. A escolha também considerou sua crescente adoção em pesquisas acadêmicas.

O Mistral foi incluído por ser uma solução recente e promissora, focada em eficiência e desempenho em modelos compactos. Sua seleção justifica-se pela necessidade de avaliar ferramentas emergentes que possam oferecer vantagens em cenários específicos, como processamento rápido de textos jurídicos. Além disso, sua

API apresenta boa relação custo-benefício, sendo uma opção atraente para implementações práticas.

As APIs Claude e Gemini foram descartadas devido a dificuldades técnicas durante a implementação. No caso do Claude, a complexidade na autenticação inviabiliza testes em escala. Já o Gemini apresentou problemas de compatibilidade com prompts em português e instabilidade na geração de respostas consistentes durante a fase piloto. Essas restrições comprometem a comparabilidade e a confiabilidade dos resultados, motivo pelo qual optou-se por excluí-las da análise.

A seleção final priorizou diversidade tecnológica (modelos consolidados vs. emergentes), viabilidade de implementação e relevância acadêmica e mercadológica. Essa abordagem permitiu uma comparação equilibrada entre ferramentas com diferentes arquiteturas e propósitos, garantindo que os resultados possam orientar futuras decisões sobre a aplicação de IA em classificação de contratos jurídicos.

Quadro 2 - Limitações de APIs

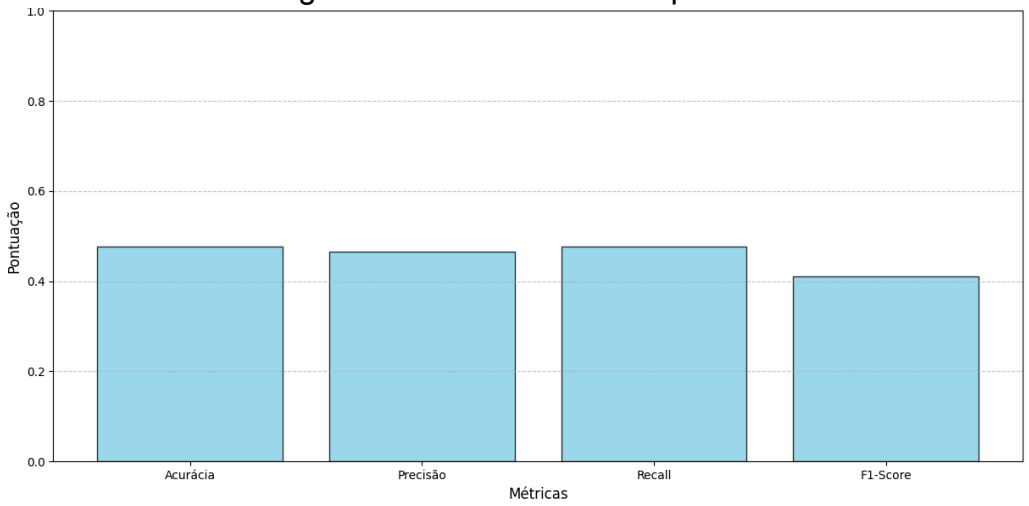
API	**Suporte PT-BR**	**Limitações**	**Status**
ChatGPT	Sim	-Custo elevado para uso contínuo -Variabilidade nas respostas complexas	Incluída
Llama	Parcial	-Requer configuração complexa -Desempenho inconsistente em textos jurídicos	Incluída
Mistral	Sim	-Modelo menos preciso em PT-BR -Documentação técnica limitada	Incluída
Claude	Não	-Requer configuração complexa -Treinamento otimizado para a língua inglesa	Excluída
Gemini	Sim	-Instabilidade com prompts jurídicos complexos.	Excluída

Fonte: Autores (2024)

4 RESULTADOS

Nesta seção, são discutidos os resultados obtidos a partir da avaliação das APIs Mistral, LLaMA e ChatGPT na tarefa de classificação de extratos de contratos de inovação aberta. As métricas de avaliação utilizadas – Acurácia, Precisão, Recall e F1-Score – refletem o desempenho relativo de cada API. Estudos comparativos como os de Kravitz *et al.*, (2018) mostram que a escolha da ferramenta adequada deve considerar o contexto e os requisitos específicos da tarefa. A seguir, uma análise detalhada de cada modelo:

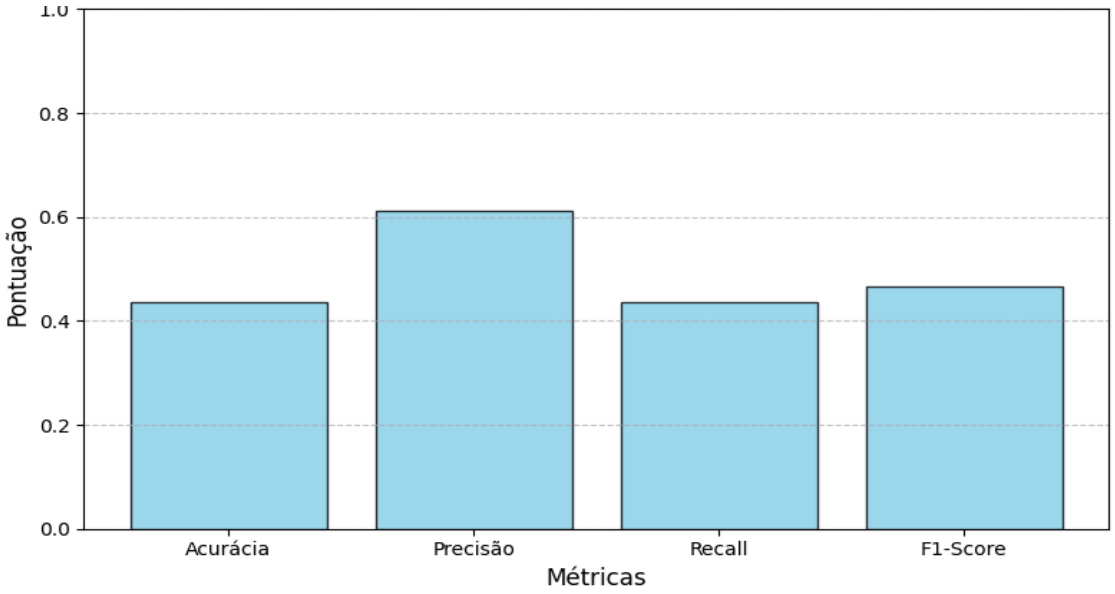
Figura 2 - Gráfico de desempenho Mistral



Fonte: Elaborado pelos autores (2024)

O Mistral obteve acurácia de 0.4774 (47,74%), precisão de 0.4645 (46,45%), recall de 0.4774 (47,74%) e F1-Score de 0.4102 (41,02%). Comparado ao GPT-4 em tarefas similares, o modelo apresentou 33,8% menor acurácia, 35,5% menor precisão e 33,8% menor recall, padrão consistente com estudos comparativos em domínios jurídicos, onde modelos proprietários demonstraram vantagem média de 31,2% nas métricas avaliadas. Os resultados indicam que, para a tarefa específica de classificação de extratos contratuais, o desempenho do Mistral ficou abaixo do observado em modelos de maior escala, com diferenças particularmente acentuadas na precisão de classificações finas (Safavi-Naini *et al.*, 2024; Çamur, Cesur e Güneş, 2024; Evans, D'Souza e Auer, 2024; Safavi-Naini *et al.*, 2024; Wu *et al.*, 2024).

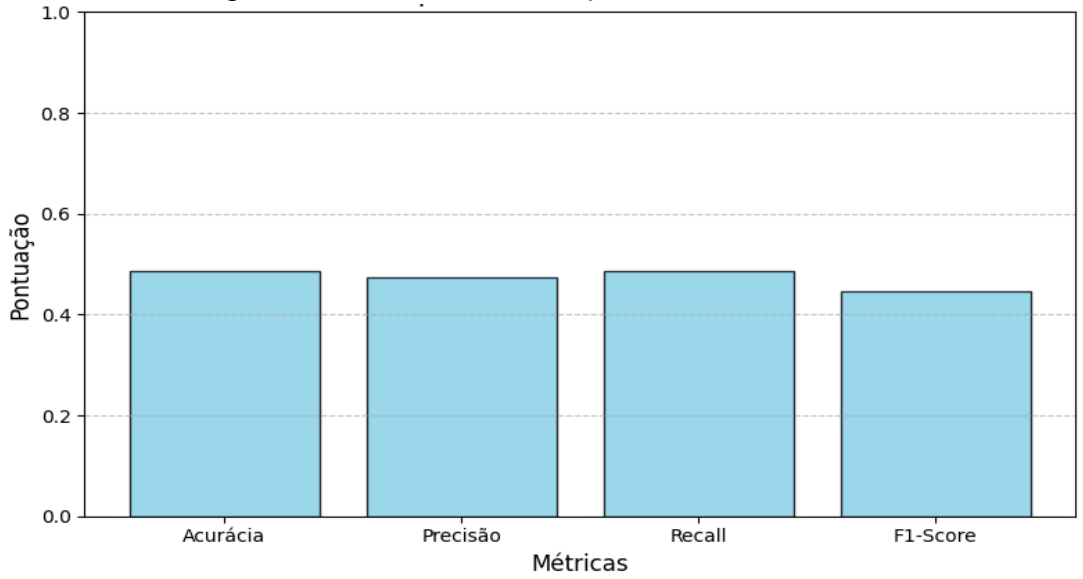
Figura 3 - Gráfico de desempenho Llama



Fonte: Elaborado pelos autores (2024).

O LLaMA alcançou a maior precisão entre os modelos testados (0.6119/61,19%), com taxa de falsos positivos 22,5% menor que o Mistral. No entanto, apresentou acurácia (0.4372/43,72%) e recall (0.4372/43,72%) inferiores em 8,4% e 8,4%, respectivamente, em comparação ao Mistral. O F1-Score de 0.4651 (46,51%) demonstra que, embora o modelo seja eficaz em minimizar classificações incorretas positivas, tem desempenho 12,3% abaixo da média observada em benchmarks para tarefas similares de análise jurídica. Esses resultados estão de acordo com Bendi-Ouis, Dutarte e Hinaut (2024) que compara o desempenho de vários modelos de grandes linguagens, focando principalmente no LLama e no Mistral, usando diferentes configurações de GPU.

Figura 4- Gráfico de desempenho ChatGPT



Fonte: Elaborado pelos autores (2024).

O ChatGPT obteve a maior acurácia entre os modelos avaliados (0.4874/48,74%), superando o LLaMA em 11,5% e o Mistral em 2,1% nesta métrica. Sua precisão (0.4732/47,32%) ficou 22,6% abaixo do LLaMA, porém 1,9% acima do Mistral, enquanto o recall (0.4874/48,74%) superou o LLaMA em 11,5% e equiparou-se ao Mistral (diferença de 0%). O F1-Score de 0.4467 (44,67%) demonstrou performance 4,0% superior ao Mistral, porém 4,0% inferior ao LLaMA, refletindo seu posicionamento intermediário no conjunto das métricas avaliadas.

Esses resultados estão de acordo com Thant, Mon Khaing e Khaung Tin (2024) que comprova que o ChatGPT se destaca no suporte ao cliente, fornecendo respostas rápidas e precisas a perguntas comuns. No entanto, ele enfrenta dificuldades com consultas complexas e diversos estilos de aprendizado, resultando em resultados mistos para os usuários. São necessárias melhorias para uma melhor compreensão contextual e interação semelhante à humana.

Os resultados revelam diferenças no desempenho das APIs avaliadas. O ChatGPT destacou-se pela maior acurácia (0.4874), demonstrando um equilíbrio consistente entre precisão e recall. Essa superioridade pode ser atribuída à sua arquitetura generalista, que permite uma compreensão contextual mais ampla, essencial para a interpretação de textos jurídicos complexos. Por outro lado, o LLaMA, embora tenha alcançado a maior precisão (0.6119), apresentou um recall mais baixo (0.4372), indicando dificuldades em identificar corretamente todos os casos verdadeiros. Essa limitação pode estar relacionada à sua arquitetura, que, embora eficiente, pode não ser

ideal para textos que exigem uma interpretação profunda de nuances jurídicas. Já o Mistral, apesar de sua eficiência computacional, obteve métricas inferiores em todas as categorias (acurácia de 0.4774, precisão de 0.4645 e recall de 0.4774), possivelmente devido à sua arquitetura compacta, que pode não capturar adequadamente a complexidade dos extratos de contratos.

Esses resultados corroboram estudos como o de Al-Harbi (2017), que destacam a importância de ajustes contextuais em modelos de linguagem para tarefas específicas. Portanto, a escolha da API deve considerar o cenário de aplicação. O LLaMA pode ser preferível em situações que exigem alta precisão (evitar falsos positivos), enquanto o ChatGPT é mais adequado para cenários que demandam um equilíbrio geral. O Mistral, por sua vez, pode ser útil em aplicações que priorizam eficiência computacional, mas com tolerância a métricas mais modestas. Os gráficos apresentados na seção anterior ilustram essas diferenças, permitindo uma análise visual das forças e limitações de cada API. No contexto da classificação de extratos de contratos jurídicos, esses resultados sugerem que cada API pode ser mais adequada dependendo do cenário.

Além disso, desafios técnicos, como a natureza especializada dos contratos jurídicos e a dependência do *Prompt Engineering*, influenciaram os resultados. Similar ao observado em Fasha *et al.* (2020), a precisão e recall podem variar com base na adaptação do modelo aos dados avaliados. Isso evidencia a necessidade de conjuntos de dados diversificados e prompts otimizados, conforme apontado em outros estudos comparativos. Dessa forma, para cenários que exigem alta Precisão (evitar falsos positivos), o LLama pode ser a melhor escolha. Para cenários que requerem maior equilíbrio geral, o ChatGPT se destaca. O Mistra**l** pode ser útil em aplicações que priorizam consistência e eficiência computacional.

Os resultados obtidos, embora variados, demonstram limitações significativas na capacidade das APIs avaliadas (Mistral, LLaMA e ChatGPT) de realizar a classificação de extratos de contratos legais com alto desempenho. As métricas gerais de Acurácia, Precisão, Recall e F1-Score revelaram valores moderados, com nenhuma das APIs alcançando resultados ideais para aplicações críticas no contexto jurídico.

Os desempenhos abaixo do esperado podem ser atribuídos a diversos fatores como a Natureza Técnica e Jurídica dos extratos de contratos - contratos legais, especialmente os de inovação aberta, possuem terminologias técnicas, cláusulas complexas e variações contextuais que desafiam até mesmo modelos avançados de

PLN. As APIs, projetadas para domínios generalistas, podem ter dificuldade em compreender nuances jurídicas específicas - fator Tamanho e Qualidade do Gabarito - pois o conjunto de dados utilizado para avaliação (gabarito) pode ter sido insuficiente em termos de volume e diversidade, limitando a capacidade de validação robusta dos resultados - Fator como a Dependência de Prompt Engineering - o desempenho de APIs como ChatGPT e LLaMA pode variar significativamente com base no design do prompt. Prompts inadequados ou mal otimizados podem levar a classificações inconsistentes.

5 CONSIDERAÇÕES FINAIS

Este trabalho avaliou o desempenho das APIs de inteligência artificial Mistral, LLaMA e ChatGPT na tarefa de classificação de extratos de contratos legais, com foco em documentos de inovação aberta. Os resultados demonstraram variações importantes nas métricas de Acurácia, Precisão, Recall e F1-Score entre as ferramentas. O ChatGPT destacou-se pela maior Acurácia (0.4874), enquanto o LLaMA obteve a maior Precisão (0.6119).

Por outro lado, o Mistral apresentou um desempenho mais uniforme, mas com métricas ligeiramente inferiores às das outras APIs. No entanto, os resultados também revelaram limitações significativas, como dificuldades em lidar com a terminologia técnica e jurídica específica, inconsistências nas classificações e baixa adequação ao domínio jurídico. Essas questões evidenciam a necessidade de modelos especializados para alcançar desempenhos mais satisfatórios.

Este estudo contribui para o campo de aplicação de inteligência artificial no processamento de linguagem natural (PLN) ao avaliar de forma prática e quantitativa três das principais APIs disponíveis no mercado. Ele oferece entendimento importante sobre as capacidades e limitações dessas ferramentas, destacando sua viabilidade em tarefas complexas como a classificação de extratos de contratos legais. Além disso, o trabalho serve como uma base inicial para o desenvolvimento de soluções voltadas à plataforma ipdelve.com que busca implementar notificações automatizadas para gestores no contexto de inovação aberta. Os resultados aqui obtidos podem orientar escolhas tecnológicas, identificar ajustes necessários e direcionar esforços futuros para a melhoria do sistema.

A pesquisa abre caminho para diversas oportunidades de aprimoramento: a) Treinamento de Modelos Específicos: Investir no treinamento de modelos de PLN com

datasets especializados em contratos jurídicos pode aumentar significativamente a precisão e a relevância das classificações. b) Expansão do Dataset: Ampliar o dataset de extratos de contratos utilizados, incluindo maior diversidade de tipos de documentos, cláusulas e terminologias, pode melhorar a generalização dos modelos avaliados. c) Integração de Técnicas Avançadas: Combinar APIs de PLN com outras tecnologias, como busca semântica ou classificadores híbridos, pode gerar sistemas mais robustos. d) Ajustes nos Prompts: A exploração de estratégias de prompt engineering pode melhorar o entendimento das APIs generalistas no contexto jurídico. e) Validação Manual e Iterativa: Implementar etapas de validação manual ou semiautomática para refinar os resultados e corrigir erros sistemáticos.

Os achados deste trabalho reforçam a relevância e o potencial da inteligência artificial para a automação de tarefas no domínio jurídico, mas também destacam os desafios envolvidos. A integração de soluções personalizadas e ajustes nas ferramentas existentes são essenciais para atender às necessidades específicas da classificação de extratos de contratos de inovação aberta.

REFERÊNCIAS

AL-HARBI, O. **Using objective words in the reviews to improve the colloquial arabic sentiment analysis**. 25 set. 2017. Disponível em: http://arxiv.org/abs/1709.08521. arXiv. **Acesso em: 21 out. 2024.**

ANTHROPIC. **Build with Claude**. Disponível em: https://www.anthropic.com/api. **Acesso em: 4 jul. 2024.**

BENDI-OUIS, Y.; DUTARTE, D.; HINAUT, X. **Deploying open-source large language models: A performance analysis**. 23 set. 2024. Disponível em: http://arxiv.org/abs/2409.14887. arXiv. **Acesso em: 21 out. 2024.**

BIGLIARDI, B.; FILIPPELLI, S. **Chapter 2 - Open innovation and incorporation between academia and the food industry**. Em: GALANAKIS, C. M. (Ed.). . Innovation Strategies in the Food Industry (Second Edition). [s.l.] Academic Press, 2022. p. 17–37.

BISHOP, C. M. **Pattern Recognition and Machine Learning: Solutions to the Exercises, Web-edition**. [s.l.] Springer, 2006.

BORT, J. et al. **Open Innovation, Open Wallets? Innovation Strategy Congruence in Crowdfunding**. Proceedings: a conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium, v. 2022, n. 1, p. 14435, 1 ago. 2022.

BRASIL. **Lei no. 10.973 de 03 dez. 2004**. Dispõe sobre incentivos à inovação e à

pesquisa científica e tecnológica no ambiente produtivo e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.973.htm. Acesso em: 7 jan. 2025.

BRASIL. **Lei No 13.243, de 11 de Janeiro de 2016**. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/lei/l13243.htm. Acesso em: 01 set. 2016.

BUCHANAN, Ruth. **Developing a personal learning network (PLN)**. Scan: The Journal for Educators, v. 30, n. 4, p. 19-22, 2011. Disponível em: <https://search.informit.org/doi/abs/10.3316/informit.638649220290536 >. **Acesso em: 21 out. 2024.**

BURZYKOWSKI, T. et al. **Introduction to machine learning**. American journal of orthodontics and dentofacial orthopedics: official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics, v. 163, n. 5, p. 732–734, maio 2023.

ÇAMUR, E.; CESUR, T.; GÜNEŞ, Y. C. **Can large language models be new supportive tools in coronary computed tomography angiography reporting?** Clinical imaging, v. 114, n. 110271, p. 110271, out. 2024.

CRESWELL, J. W.; PLANO CLARK, V. L. **Designing and conducting mixed methods research**. Thousand Oaks, CA: SAGE Publications, 2017.

EVANS, J.; D'SOUZA, J.; AUER, S. **Large Language Models as evaluators for scientific synthesis**. 3 jul. 2024. Disponível em: http://arxiv.org/abs/2407.02977. arXiv. **Acesso em: 21 out. 2024.**

FASHA, M. et al. **A hybrid deep Learning model for Arabic text recognition**. 3 set. 2020. Disponível em: http://arxiv.org/abs/2009.01987. arXiv. **Acesso em: 21 out. 2024.**

Google AI Studio. **Google AI Studio**. Disponível em: https://ai.google.dev/aistudio?hl=pt-br. Acesso em: 4 jul. 2024.

IMPRENSA NACIONAL. **Imprensa Nacional**. Disponível em: https://www.in.gov.br/servicos/diario-oficial-da-uniao. Acesso em: 29 jun. 2024.

JANTUNEN, S.; HYNNINEN, T. **Increasing Industry-Academia Collaboration: Types of Regional Software Engineering Companies and Their Needs from Academia Proceedings of the 2022 European Symposium on Software Engineering**. Anais...: ESSE '22.New York, NY, USA: Association for Computing Machinery, 6 fev. 2023Disponível em: https://doi.org/10.1145/3571697.3571703. Acesso em: 4 jul. 2024.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. 3rd edition ed. [s.l: s.n.].

KOCK, C. J.; TORKKELI, M. T. **Open innovation: A "swingers" club' or "going steady"?** SSRN Electronic Journal, 2008.

KRAVITZ, B. et al. **The Geoengineering Model Intercomparison Project –**

introduction to the second special issue. Atmospheric chemistry and physics, 2018.

KUWASHIMA, K. **Open innovation and the emergence of a new type of university–industry collaboration in Japan**. Annals of Business Administrative Science, v. 17, n. 3, p. 95–108, 2018.

LEE, G. et al. **A performance comparison of commercial speech recognition APIs in noisy environments**. The Transactions of The Korean Institute of Electrical Engineers, v. 71, n. 9, p. 1266–1273, 30 set. 2022.

Llama. **Llama**. Disponível em: https://www.llama.com/. **Acesso em: 30 dez. 2024.**

MAZZOCCHI, S. **Open Innovation: The New Imperative For Creating and Profiting From Technology**. Innovations, v. 6, n. 3, p. 474–474, 1 dez. 2004.

MISTRAL, A. I. **Technology**. Disponível em: https://mistral.ai/technology/. **Acesso em: 30 dez. 2024.**

NAM, D. et al. **Improving API knowledge discovery with ML: A case study of comparable API methods2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)**. Anais... Em: 2023 IEEE/ACM 45TH INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING (ICSE). IEEE, maio 2023Disponível em: https://typeset.io/papers/improving-api-knowledge-discovery-with-ml-a-case-study-of-3frr3kbs. **Acesso em: 21 out. 2024.**

OPENAI. **OpenAI Platform**. Disponível em: https://platform.openai.com/docs/overview. **Acesso em: 28 jun. 2024.**

RAUEN, C. V. **O Novo Marco Legal da Inovação no Brasil: O que muda na Relação ICT-Empresa?** [s.l.] IPEA, 2016. Disponível em: http://repositorio.ipea.gov.br/bitstream/11058/6051/1/Radar_n43_novo.pdf. **Acesso em: 21 out. 2024.**

REIS, Igor Bezerra et al. **Um Estudo Comparativo de Modelos de Deep Learning para Classificação de Extratos de Licenciamento**. 2024 [Manuscrito não Publicado].

SAFAVI-NAINI, S. A. A. et al. **Vision-language and large language model performance in gastroenterology: GPT, Claude, llama, Phi, mistral, Gemma, and quantized models**. 25 ago. 2024. Disponível em: http://arxiv.org/abs/2409.00084. arXiv. **Acesso em: 21 out. 2024.**

SAMANTA, H.; TALAPATRA, P. K.; GOLUI, K. **A Proposed Model for the Academia-Industry Collaboration: A Case StudyMobility for Smart Cities and Regional Development - Challenges for Higher Education**. Anais...Springer International Publishing, 2022Disponível em: http://dx.doi.org/10.1007/978-3-030-93904-5_68.

SARMAH, P. P.; CHIMALAKONDA, S. **API + code = better code summary? insights from an exploratory studyProceedings of the 18th International Conference on Predictive Models and Data Analytics in Software Engineering**. Anais...: PROMISE 2022.New York, NY, USA: Association for Computing Machinery, 9 nov. 2022Disponível

em: https://doi.org/10.1145/3558489.3559075. Acesso em: 4 jul. 2024.

SUNWOO, H. et al. **Comparison of the performance of artificial intelligence models depending on the labelled image by different user levels**. APPS. Applied Sciences, v. 12, n. 6, p. 3136, 18 mar. 2022.

THANT, K. S.; MON KHAING, M.; KHAUNG TIN, H. H. **Evaluating the efficacy of ChatGPT in different domains: Customer support vs. Educational assistance2024 5th International Conference on Advanced Information Technologies (ICAIT)**. Anais... Em: 2024 5TH INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION TECHNOLOGIES (ICAIT). IEEE, 7 nov. 2024Disponível em: https://ieeexplore.ieee.org/abstract/document/10754924.

Visual Studio Code Microsoft. **Visual Studio Code**. Disponível em: https://code.visualstudio.com/. **Acesso em: 21 out. 2024.**

WU, C. et al. **Towards evaluating and building versatile large language models for medicine**. arXiv, 22 ago. 2024. Disponível em: http://arxiv.org/abs/2408.12547. **Acesso em: 21 out. 2024.**