



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

MATHEUS RIBEIRO DE ARAUJO

**AUTOMAÇÃO DA ANÁLISE NA COLIDÊNCIA VISUAL DE MARCAS: Uma
abordagem com LLMs**

FLORIANO

2026

MATHEUS RIBEIRO DE ARAUJO

AUTOMAÇÃO DA ANÁLISE NA COLIDÊNCIA VISUAL DE MARCAS: Uma
abordagem com Ilms

Trabalho de Conclusão de Curso (TCC)
apresentado como exigência para aprovação na
disciplina Trabalho de Conclusão de 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 A S Leite .
Coorientador(a): Prof. MSc. Igor Bezerra .

FLORIANO

2026

AUTOMAÇÃO DA ANÁLISE DE COLIDÊNCIA VISUAL DE MARCAS: Um Método Baseado em LLMS

Matheus Ribeiro de Araujo¹

Rafael Angelo Santos Leite²

Igor reis Bezerra³

RESUMO

A análise de colidência visual entre marcas figurativas é uma das atividades mais complexas do exame de registros marcários, pois exige a comparação simultânea de múltiplos elementos gráficos e depende, em grande medida, da interpretação subjetiva do examinador. Diante do crescente volume de pedidos de registro, este trabalho teve como objetivo desenvolver e avaliar um método automatizado, baseado em modelos de linguagem multimodal de grande escala, para apoiar a identificação de semelhança visual entre marcas figurativas. A metodologia compreendeu quatro etapas principais: a construção de um conjunto de dados derivado, composto exclusivamente por pares de marcas figurativas extraídas de casos reais brasileiros; o ajuste fino de quatro modelos multimodais de código aberto, por meio da técnica de adaptação de baixa ordem; o desenvolvimento de um modelo explicador, orientado por engenharia de prompt e por aprendizado com few-shot; e a avaliação quantitativa e qualitativa dos resultados. O modelo Qwen2 alcançou o melhor desempenho, com acurácia de 86% e elevada capacidade de identificar conflitos reais, sendo selecionado para gerar relatórios técnicos fundamentados. Tais relatórios foram avaliados de forma positiva por especialistas em propriedade intelectual. Conclui-se que o método proposto reduz a subjetividade, amplia a transparência e constitui ferramenta promissora de apoio à análise marcária, ainda que dependa de revisão humana.

Palavras-chave: marcas figurativas; colidência visual; modelos de linguagem multimodais; propriedade intelectual; aprendizado de máquina.

ABSTRACT

Abstract

The visual similarity assessment of figurative trademarks is one of the most challenging tasks in trademark examination, as it requires the simultaneous comparison of multiple graphical elements and largely depends on the examiner's subjective interpretation. Given the growing number of trademark applications, this study aimed to develop and evaluate an automated method based on large multimodal language models to support the identification of visual similarity between figurative trademarks. The proposed methodology comprised four main stages: the construction of a derived dataset composed exclusively of pairs of figurative trademarks extracted from real Brazilian trademark cases; the fine-tuning of four open-source multimodal models using the Low-Rank Adaptation (LoRA) technique; the development of an explanation model guided by prompt engineering and

¹ Matheus Ribeiro de Araujo . IFPI. ribeiro790j@gmail.com

² Rafael Angelo Santos Leite. IFPI. rafaelangelo@ifpi.edu.br

³ Igor Bezerra Reis. IFPI. igor.bezerra@ifpi.edu.br

few-shot learning; and the quantitative and qualitative evaluation of the results. Among the evaluated models, Qwen2 achieved the best performance, reaching an accuracy of 86% and demonstrating a high capability to identify actual trademark conflicts. Consequently, it was selected to generate technically grounded explanatory reports, which were positively evaluated by intellectual property specialists. The results indicate that the proposed method reduces subjectivity, enhances transparency, and represents a promising decision-support tool for figurative trademark similarity analysis, although human review remains essential.

Keywords: figurative trademarks; visual similarity; multimodal language models; intellectual property; machine learning.

Data de aprovação: 07/07/2026

1 INTRODUÇÃO

Em mercados cada vez mais competitivos, a identidade visual de uma empresa tornou-se um de seus patrimônios mais valiosos. É por meio dela que o consumidor reconhece a origem de um produto ou serviço, diferencia concorrentes e constrói vínculos de confiança que se traduzem em preferência e fidelização (Vesnin; Levshun; Chechulin, 2023). Esse reconhecimento é juridicamente assegurado pela Marca Registrada (MR), ativo intangível que confere ao titular o direito exclusivo de uso e impede que terceiros se beneficiem indevidamente de sua reputação (Trappey; Trappey; Lin, 2020). Proteger uma MR é, portanto, proteger não apenas um sinal gráfico, mas todo o valor econômico e simbólico associado a ele, evitando prejuízos decorrentes de infrações à Propriedade Intelectual (PI).

Embora a forma de regulação varie entre jurisdições, há notável convergência internacional quanto à função essencial da marca. No Brasil, o Instituto Nacional da Propriedade Industrial (INPI) define a marca como o sinal distintivo destinado a identificar a origem e a distinguir produtos ou serviços de outros idênticos, semelhantes ou afins, classificando-a em quatro tipos: nominativa, figurativa, mista e tridimensional (INPI, 2023). Entendimento análogo é adotado pelo *Canadian Intellectual Property Office* (CIPO), que descreve a marca como a combinação de letras, palavras, sons ou desenhos que diferencia uma empresa de suas concorrentes (CIPO, 1976), e pelo *United States Patent and Trademark Office* (USPTO), para o qual o registro permite que o público identifique e distinga a procedência comercial dos produtos (Uspto, 2024). A *World Intellectual Property Organization* (WIPO), por sua vez, adota definição ainda mais ampla, reconhecendo como marca qualquer sinal capaz de distinguir produtos ou serviços no mercado.

Dada essa importância estratégica, o exame prévio de registrabilidade ocupa posição central na prevenção de conflitos comerciais e jurídicos. Cabe ao INPI verificar se cada pedido atende aos requisitos legais, entre eles a distintividade e a ausência de colidência com marcas anteriores (INPI, 2023). Quando dois sinais guardam semelhança suficiente para induzir o consumidor a erro ou associação indevida, configura-se o risco de confusão que pode dar origem a oposições, processos de nulidade e litígios. Essa proximidade pode manifestar-se em diferentes dimensões: fonética, textual, ideológica ou visual, frequentemente reforçada pela própria afinidade mercadológica entre os produtos envolvidos.

Entre todas essas dimensões, a análise das marcas figurativas, compostas exclusivamente por desenhos, figuras ou símbolos, é particularmente desafiadora. Diferentemente das marcas nominativas, cuja comparação pode apoiar-se em critérios fonéticos e textuais relativamente objetivos, a avaliação figurativa depende da percepção simultânea de forma, silhueta, composição, traços e disposição dos elementos gráficos. Trata-se de um julgamento de natureza subjetiva, no qual diferentes examinadores podem chegar a conclusões distintas diante do mesmo par de sinais (Showkatramani et al., 2018a; Trappey; Trappey; Lin, 2020), comprometendo a previsibilidade e a uniformidade das decisões.

Esse desafio é agravado pelo crescimento contínuo do volume de pedidos submetidos ao INPI, que pressiona a capacidade de análise e amplia a necessidade de instrumentos capazes de tornar o exame mais ágil, transparente e consistente. Nesse cenário, os avanços recentes em Inteligência Artificial (IA), especialmente os Modelos de Linguagem de Grande Escala (LLMs) em suas versões multimodais, abrem novas possibilidades. Ao integrar a interpretação de imagens à capacidade de raciocínio e geração de texto, esses modelos não apenas classificam a similaridade entre sinais, mas também são capazes de explicar, em linguagem acessível, os motivos que fundamentam tal julgamento (Zhao et al., 2024; Banwait; Prabha, 2026).

Diante dessa lacuna, o presente trabalho propõe um método automatizado para apoiar a análise de colidência visual entre marcas figurativas. A solução emprega LLMs multimodais para classificar o grau de similaridade entre pares de marcas e gerar relatórios técnicos que justifiquem a presença ou ausência de conflito. O sistema foi treinado a partir de casos reais de indeferimento do INPI e validado por especialistas em PI, com o propósito de reduzir a subjetividade, aumentar a eficiência e conferir maior transparência ao processo de exame marcário, atuando como ferramenta de apoio à decisão e não como substituto do julgamento humano.

Além desta introdução, este trabalho está organizado em quatro seções. A segunda seção apresenta o referencial teórico, abordando o conceito e os tipos de marcas, as diferentes dimensões de similaridade consideradas no exame marcário e os fundamentos dos LLMs, incluindo suas variantes multimodais. A terceira seção descreve a metodologia adotada, detalhando a construção do conjunto de dados CMAD Visual, o treinamento dos modelos classificadores, o desenvolvimento do

modelo explicador e os critérios de avaliação. A quarta seção expõe e discute os resultados obtidos, tanto na classificação quanto na geração de relatórios técnicos. Por fim, a quinta seção apresenta as considerações finais, sintetizando as contribuições, reconhecendo as limitações do estudo e apontando desdobramentos futuros.

2 REFERENCIAL TEÓRICO

Esta seção reúne os fundamentos teóricos necessários à compreensão do método proposto. Inicialmente, discute-se o conceito de marca e suas formas de apresentação, com ênfase nas marcas figurativas, objeto central deste estudo. Em seguida, apresentam-se as diferentes dimensões de similaridade consideradas na análise de colidência , fonética, textual, ideológica e visual , destacando os critérios adotados pelo INPI . Por fim, abordam-se os LLMs e sua evolução para arquiteturas multimodais, bem como as técnicas de ajuste fino .

2.1 MARCAS

De acordo com o INPI, órgão responsável por garantir a PI no Brasil conforme determinado Art. 2º da Lei 5.648, de 11 de dezembro de 1970, a marca é um sinal distintivo cujas funções principais são: identificação da origem, distinção de serviços ou produtos de outros idênticos, semelhantes ou afins de origem diversa (INPI, 2023). A apresentação de uma marca pode ser nominativa, ou verbal, conforme mostra a figura 1. Neste tipo de marca, o que vai ter mais destaque é sua composição de uma ou mais palavras no sentido amplo do alfabeto romano, abrangendo neologismo, combinações de letras, algarismos romanos, arábicos, contanto que os elementos não apresentem forma fantasiosa ou figurativa.

Figura 1 - Exemplo de marcas nominativas

TREZENTOS E OITENTA	CORONA	VARIG
XIX	Atlântica	IRREVERENTE & CIA

Fonte: INPI (2023).

As marcas também podem receber a classificação figurativa ou emblemática, conforme ilustrado na Figura 2. Neste tipo de classificação é destaque: desenhos, imagens, figuras, símbolos. Outrossim, qualquer forma fantasiosa, figurativa de letra, algarismo isoladamente, acompanhado por desenho, imagem, figura, símbolo. Ademais, palavras compostas por letras de alfabetos distintos da língua vernácula, como hebraico, cirílico, árabe, igualmente ideogramas, tais como japonês e chinês (INPI, 2023).

Figura 2 - Exemplo de marcas figurativas



Fonte: Elaborado pelo autor (2026).

Uma marca mista, ou composta, é o sinal constituído pela combinação de elementos nominativos e figurativos ou mesmo apenas por elementos nominativos cuja grafia se apresente sob forma fantasiosa ou estilizada, como exemplificado na Figura 3.

Figura 3 - Exemplo de marcas mista



Fonte: Elaborado pelo autor (2026).

Todas essas formas utilizadas para identificar os produtos ou serviços de uma empresa. Sua definição apresenta elevada consonância com diversos institutos de PI ao redor do mundo, como o USPTO, o INPI, o CIPO e a WIPO. A proteção desses elementos visuais é fundamental, visto que a similaridade entre marcas representa uma das principais causas de indeferimento de pedidos de registro (Showkatramani et al., 2018).

2.2 TIPOS DE SIMILARIDADE

2.2.1 Similaridade Fonética e Textual

De acordo com o manual de marcas do INPI para a análise de colidência de marcas é levado em consideração se os sinais podem causar confusão ou associação indevida no consumidor, outro ponto que merece destaque é que em regra geral a análise de colidência entre os sinais baseia-se na avaliação da impressão geral dos conjuntos e não apenas em seus elementos individuais, sendo levado em conta, ao mesmo tempo, os aspectos gráficos, ideológicos, fonéticos e dos signos comparados.

A similaridade fonética refere-se à semelhança existente na pronúncia das marcas. Nessa análise, são observados aspectos como sequência silábica, ritmo, tonicidade e entonação das palavras presentes nos sinais em comparação. Mesmo quando apresentam grafias distintas, duas marcas podem produzir impressões sonoras muito próximas, aumentando o risco de confusão entre os consumidores.

A Figura 4 apresenta exemplos de marcas que demonstram essa proximidade fonética. As expressões “CAVALINHO AZUL”, “CAVALINHO BLUE”, “CAVALO AZULADO” e “KAVALLO AZULADO”, apesar das diferenças ortográficas e da presença de termos em idiomas distintos, mantêm forte aproximação sonora e conceitual, especialmente pela repetição dos elementos “cavalo/cavalinho” e “azul/azulado/blue”. Situação semelhante ocorre com as marcas “JOLY”, “JOLLI” e “JOLIE”, cujas pequenas alterações gráficas não eliminam a semelhança fonética percebida pelo público consumidor (INPI, 2023).

Figura 4 - Exemplo de similaridade Fonética

CAVALINHO AZUL para assinalar roupas infantis	X	CAVALINHO BLUE CAVALO AZULADO KAVALLLO AZULADO para assinalar roupas
JOLY para assinalar laticínios	X	JOLLI JOLIE para assinalar iogurte

Fonte: INPI (2023)

Por sua vez, a Figura 5 evidencia situações em que há coincidência textual entre marcas utilizadas para identificar produtos ou serviços semelhantes. As marcas “JOGO DO AMOR”, utilizadas respectivamente para perfumes e cosméticos, apresentam identidade textual completa, aumentando significativamente a possibilidade de associação pelo público. O mesmo ocorre com “DOMINGO NO PARQUE”, voltada para organização de programas de televisão e organização de eventos, bem como com “*MISTER EGO*”, utilizada para serviços de salão de cabeleireiro e estética pessoal. Outro exemplo relevante é a marca “*SPREEM*”, utilizada para assinalar produtos eletroeletrônicos em ambos os registros apresentados. Nesse caso, verifica-se identidade textual, fonética e mercadológica, configurando elevado potencial de confusão entre os sinais.

Figura 5 - Exemplo de similaridade fonética

JOGO DO AMOR para assinalar perfumes	X	JOGO DO AMOR para assinalar cosméticos
DOMINGO NO PARQUE para assinalar organização de programas de televisão	X	DOMINGO NO PARQUE para assinalar organização de eventos
MISTER EGO para assinalar serviços de salão de cabeleireiro	X	MISTER EGO para assinalar serviços de estética pessoal
SPREEM para assinalar eletroeletrônicos	X	SPREEM para assinalar eletroeletrônicos

Fonte: INPI (2023)

2.2.3 Similaridade Ideológica

Também pode ocorrer colidência marcária por meio da reprodução ou imitação ideológica, situação em que sinais distintos do ponto de vista gráfico ou fonético transmitem uma mesma ideia ou conceito ao consumidor. Nesse tipo de análise, considera-se o conteúdo semântico das marcas e a percepção conceitual que elas despertam no público, uma vez que diferentes expressões podem remeter ao mesmo significado e, conseqüentemente, gerar associação indevida entre produtos ou serviços (INPI, 2023).

A Figura 6 apresenta exemplos de marcas que, embora possuam diferenças visuais e textuais, demonstram proximidade ideológica significativa. As expressões “CIDADE DOS AUTOMÓVEIS” e “CIDADE DOS CARROS”, por exemplo, fazem referência ao mesmo conceito mercadológico relacionado ao comércio de veículos. Apesar da substituição do termo “automóveis” por “carros”, ambos os sinais transmitem essencialmente a mesma ideia ao consumidor, podendo gerar associação quanto à origem empresarial dos serviços prestados.

Figura 6 - Exemplo de similaridade Ideológica

CIDADE DOS AUTOMÓVEIS para assinalar comércio de veículos	X	CIDADE DOS CARROS para assinalar comércio de veículos ou peças automotivas
 para assinalar ferramentas manuais	X	BLACK HORSE para assinalar comércio de artigos de ferragem
TRÊS IGREJAS para assinalar leite	X	 para assinalar queijos

Fonte: INPI (2023)

Outro exemplo presente na figura 6 refere-se à comparação entre a marca figurativa composta pela imagem de um cavalo e a expressão nominativa “*BLACK HORSE*”. Ainda que um dos sinais seja predominantemente figurativo e o outro verbal, ambos evocam diretamente a figura do animal “cavalo”, estabelecendo

semelhança conceitual capaz de aproximar os conjuntos marcários perante o público consumidor.

Situação semelhante ocorre com a marca “TRÊS IGREJAS” e o sinal figurativo composto pela representação gráfica de três edificações religiosas. Nesse caso, embora um dos sinais seja textual e o outro exclusivamente figurativo, ambos remetem ao mesmo conceito visual e semântico, reforçando a possibilidade de associação indevida entre os produtos identificados.

Além disso, a reprodução ideológica pode ocorrer mesmo quando as marcas utilizam idiomas diferentes. Um exemplo frequentemente citado no manual de marcas do INPI é a comparação entre “*PERFECT RUN*” e “CORRIDA PERFEITA”, expressões que, apesar de apresentarem diferenças linguísticas, possuem equivalência semântica direta. Nesses casos, o grau de familiaridade do consumidor com idiomas estrangeiros torna-se um fator relevante na avaliação da possibilidade de confusão ou associação indevida entre os sinais marcários (INPI, 2023).

2.2.4 Similaridade Visual

O uso de formas geométricas, imagens, cores e combinações cromáticas semelhantes pode aumentar o risco de confusão ou associação indevida entre marcas. Dessa forma, a análise dos aspectos gráficos dos sinais constitui um elemento essencial na verificação da possibilidade de colidência marcária.

A Figura 7 apresenta exemplos em que, mesmo sem identidade textual ou fonética, é possível observar proximidade visual entre os elementos gráficos das marcas. Nos conjuntos apresentados, nota-se a utilização de figuras semelhantes, como criaturas aladas e capacetes estilizados, cujas composições visuais compartilham características próximas de formato, silhueta e disposição gráfica. Essas semelhanças podem levar o consumidor a estabelecer associações indevidas entre os sinais marcários.

Figura 7 - Exemplo de marcas visualmente semelhantes.





Fonte: Elaborado pelo autor (2026).

Embora essa análise seja especialmente relevante em marcas figurativas, mistas e tridimensionais, a comparação dos elementos visuais desempenha papel fundamental na identificação de possíveis conflitos marcários. Nesses casos, aspectos como formato, composição gráfica, símbolos, traços, cores e disposição visual dos elementos podem gerar proximidade entre os sinais, aumentando a possibilidade de associação indevida ou confusão por parte do público consumidor (Inpi, 2023). Essa diretriz nacional está em plena consonância com os parâmetros da WIPO, que estabelece que as marcas também englobam logotipos, figuras e imagens, e que o principal objetivo de seu registro é justamente permitir o bloqueio de sinais "semelhantes a ponto de causar confusão", garantindo que os consumidores não sejam enganados (Wipo, 2019).

2.3 MODELOS DE LINGUAGEM DE GRANDE ESCALA (LLMs)

As LLMs representam um dos avanços mais significativos da IA contemporânea, promovendo a transição de modelos especializados em tarefas específicas para modelos de fundação capazes de executar múltiplas tarefas de forma generalista (Zhao et al., 2024). O desenvolvimento desses modelos está fundamentado principalmente na arquitetura *Transformer*, a qual revolucionou o campo do Processamento de Linguagem Natural (PLN) ao substituir arquiteturas sequenciais, como Redes Neurais Recorrentes (RNNs) e *Long Short-Term Memory* (LSTM), por mecanismos paralelizáveis baseados em atenção (Banwait; Prabha, 2026).

A principal inovação introduzida pela arquitetura *Transformer* é o mecanismo de *self-attention* (autoatenção), responsável por permitir o processamento paralelo de sequências e a atribuição dinâmica de pesos aos diferentes elementos do texto. (Vaswani et al., 2017).

Esse mecanismo possibilita a captura de dependências de longo alcance e a construção de representações contextuais mais ricas e eficientes (Banwait; Prabha, 2026; Zhao et al., 2024).

Os modelos baseados em *Transformer* podem ser classificados em três categorias principais: modelos baseados apenas em codificadores (*encoder-only*), representados pelo *Bidirectional Encoder Representations from Transformers* (BERT); modelos codificador-decodificador (*encoder-decoder*) e modelos baseados exclusivamente em decodificadores (*decoder-only*), exemplificados pelas famílias GPT e LLaMA (Banwait; Prabha, 2026). Atualmente, a arquitetura decoder-only consolidou-se como o principal paradigma para tarefas de geração de texto e raciocínio em domínio aberto, impulsionada pelas chamadas leis de escala (*scaling laws*), segundo as quais o aumento do número de parâmetros e do volume de dados tende a produzir ganhos significativos de desempenho (Zhao et al., 2024).

Apesar dos avanços alcançados e da emergência de capacidades como o Aprendizado em Contexto (*In-Context Learning*), o treinamento e a implantação de LLMs ainda enfrentam desafios computacionais e metodológicos relevantes (Zhao et al., 2024). Entre as principais limitações encontra-se o fenômeno da alucinação, caracterizado pela geração de conteúdos linguisticamente coerentes, porém factualmente incorretos ou inconsistentes com o contexto apresentado, comprometendo a confiabilidade dos modelos em aplicações críticas (Banwait; Prabha, 2026; Liu; Leong, 2026).

Além disso, sobre leis de escalabilidade em modelos de linguagem demonstram que aumentam em complexidade do modelo não garantem ganhos proporcionais de desempenho, evidenciando um padrão de retornos decrescentes (Hoffmann et al., 2022). Evidências empíricas corroboram este fenômeno, demonstrando que otimizações arquitetônicas, como multi-input e inicialização de pesos apropriada, superam o simples aumento do número de camadas (Trappey et al., 2020), enquanto limitações de modelos treinados em dados históricos reforçam a necessidade de contexto específico (Bui, 2025), evidenciando a necessidade crítica de equilíbrio entre arquitetura, volume de dados e qualidade do treinamento.

Nesse contexto, técnicas de Ajuste Fino Eficiente em Parâmetros (*Parameter-Efficient Fine-Tuning*, PEFT) têm sido amplamente empregadas para reduzir os custos associados à adaptação de modelos para domínios específicos (Banwait; Prabha, 2026). Um exemplo dessa aplicação é a análise de colidência de

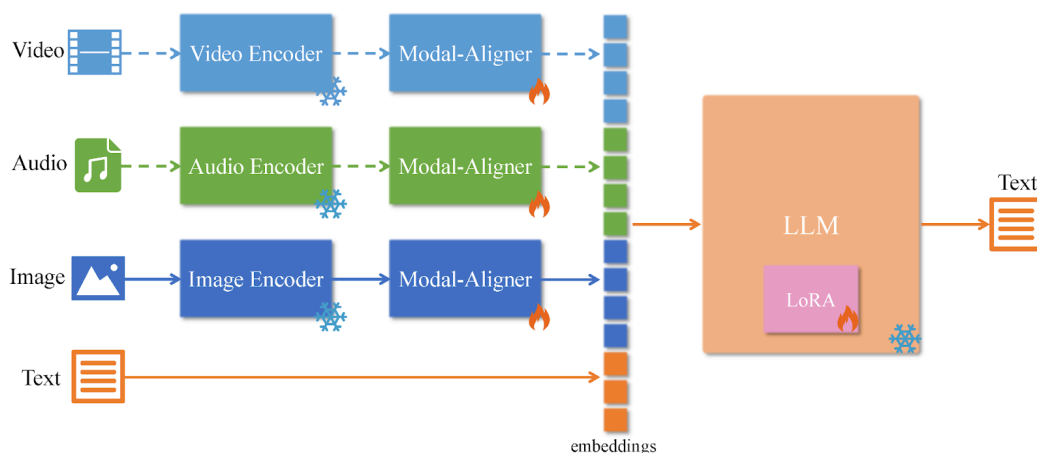
marcas proposta neste trabalho. Entre essas técnicas, destaca-se a Adaptação de Baixa Ordem (*Low-Rank Adaptation*, LoRA), que consiste na inserção de matrizes de atualização de baixa dimensionalidade nos módulos de atenção, mantendo congelados os parâmetros originais do modelo pré-treinado (Rezaei et al., 2025; Zhao et al., 2024). Evidências experimentais indicam que essa abordagem é capaz de reduzir significativamente o número de parâmetros treináveis e os custos computacionais, preservando níveis de desempenho comparáveis ou superiores aos obtidos por meio do ajuste fino completo (Rezaei et al., 2025). Dessa forma, o LoRA viabiliza a adoção de LLMs em ambientes com restrições de infraestrutura computacional, sem comprometer substancialmente a precisão dos resultados.

2.3.1 LLMs Multimodais

O advento e o crescimento exponencial de dados heterogêneos impulsionaram a evolução dos LLMs para o chamado estágio de fusão multimodal. Nesse novo paradigma, sinais textuais, visuais, auditivos e códigos de programação são alinhados e processados em um espaço de representação semântica unificado (Liu; Leong, 2026; Wang et al., 2023), conforme ilustrado na figura 8. Modelos de vanguarda: GPT-4V, LLaVA, Flamingo, Gemini e arquiteturas recentes baseadas em Mistura de Especialistas (MoE), têm se destacado por fundir capacidades avançadas de compreensão linguística com mecanismos sofisticados de percepção sensorial (Banwait; Prabha, 2026; Wang et al., 2023).

Diferentemente dos sistemas tradicionais de visão computacional, que se limitam predominantemente à extração de características em nível de pixel, os LLMs multimodais, transcendem a identificação de padrões ao incorporar seu vasto "cérebro" linguístico para interpretar os dados de maneira profundamente contextualizada (Banwait; Prabha, 2026). Essa sinergia entre modalidades possibilita a realização de inferências lógicas complexas sobre objetos, cenários e relações semânticas, culminando em uma verdadeira cognição multimodal que aproxima significativamente o processamento computacional da percepção e intuição humanas (Banwait; Prabha, 2026; Liu; Leong, 2026).

Figura 8 – Modelos de Linguagem Multimodais.



Fonte: Huang et.al (2024,p.10)

Esses modelos têm alcançado o estado da arte em tarefas transversais complexas, tais como Resposta a Perguntas Visuais (*Visual Question Answering*, VQA), síntese intermodal e raciocínio baseado em elementos gráficos (Wang et al., 2023). Ao quebrar as barreiras entre diferentes formatos de dados, tais avanços ampliam drasticamente o potencial de aplicação da IA. Essa evolução viabiliza desde a automação de sistemas corporativos e análises documentais rigorosas até inovações em instituições como o INPI.

Para compreender como esses modelos processam simultaneamente texto e imagem, é necessário examinar seu funcionamento interno. Em sua maioria, os LLMs multimodais de visão e linguagem adotam uma arquitetura composta por três módulos principais: um codificador de modalidade (*modality encoder*), um módulo conector (*connector* ou projetor multimodal) e o modelo de linguagem propriamente dito (Yin et al., 2024; Huang et al., 2024).

O codificador visual é um *Vision Transformer* (ViT) pré-treinado em larga escala por aprendizado contrastivo imagem-texto (Yin et al., 2024). Nesse tipo de arquitetura, a imagem de entrada é dividida em pequenos blocos de dimensão fixa, denominados *patches*, e cada bloco é convertido em um vetor numérico, o *embedding* visual, de modo que a imagem passa a ser representada como uma sequência de vetores, análoga a uma sequência de tokens textuais (Khan et al., 2022).

Como os modelos de linguagem foram concebidos para perceber apenas texto, e o codificador visual opera em um espaço de representação distinto, esses embeddings visuais não podem ser consumidos diretamente pelo LLM. É o módulo conector que cumpre o papel de ponte entre as duas modalidades, projetando os embeddings visuais para o espaço de representação que o modelo de linguagem é capaz de interpretar, sendo comumente implementado por meio de camadas de projeção, como perceptrons multicamadas (Yin et al., 2024).

Na estratégia mais difundida, conhecida como fusão em nível de token (*token-level fusion*), as características extraídas pelo codificador são transformadas em tokens visuais e concatenadas aos tokens de texto da entrada, formando uma única sequência multimodal que é processada pelo LLM, o qual gera a resposta token a token, de maneira autoregressiva (Yin et al., 2024). Em outras palavras, do ponto de vista do modelo de linguagem, a imagem passa a ser tratada como uma espécie de pseudo-texto inserido no prompt, permitindo que o mecanismo de autoatenção da arquitetura *Transformer* relaciona livremente elementos visuais e textuais em todas as camadas do modelo (Lin et al., 2022).

Essa abordagem modular é vantajosa porque evita o custo proibitivo de treinar um modelo multimodal do zero: parte-se de um codificador visual e de um LLM já pré-treinados, e o conector, módulo leve e treinável, é o principal responsável pelo alinhamento entre as modalidades, tipicamente ajustado em etapas sucessivas de pré-treinamento de alinhamento sobre pares imagem-texto e de ajuste por instruções multimodais (Zhang et al., 2024).

O Qwen2-VL, segue esse mesmo paradigma geral, mas introduz o mecanismo de resolução dinâmica (*Naive Dynamic Resolution*), que converte imagens de diferentes resoluções em quantidades variáveis de tokens visuais, além do *Multimodal Rotary Position Embedding* (M-RoPE), que favorece a fusão de informações posicionais entre texto e imagem (Wang et al., 2024), características particularmente relevantes para o domínio de marcas figurativas, no qual os logotipos variam consideravelmente em proporção, nível de detalhe e resolução.

Contudo, adaptar modelos dessa escala a domínios específicos, como a análise de colidência marcária, impõe custos computacionais elevados, uma vez que o ajuste fino completo exigiria a atualização de bilhões de parâmetros. É nesse cenário que se inserem as técnicas de *Parameter-Efficient Fine-Tuning* (PEFT), cuja premissa é adaptar o modelo atualizando apenas uma fração mínima de seus

parâmetros, mantendo o restante congelado (Banwait; Prabha, 2026; Zhao et al., 2024).

Entre essas técnicas, destaca-se a LoRA, que parte do pressuposto de que a atualização dos pesos necessária para especializar o modelo em um novo domínio possui posto (*rank*) intrinsecamente baixo. Assim, em vez de atualizar diretamente uma matriz de pesos pré-treinada, o LoRA a mantém congelada e aprende uma atualização decomposta em duas matrizes de baixa dimensionalidade (Hu et al., 2022).

De maneira complementar ao PEFT, a quantização ataca o custo computacional por outra frente: a redução da precisão numérica dos pesos do modelo, originalmente armazenados em ponto flutuante de 32 ou 16 bits, para formatos de menor precisão, como 8 ou 4 bits, diminuindo proporcionalmente o consumo de memória e o custo de inferência e treinamento (Dettmers et al., 2023). Nesse contexto, destaca-se a técnica *Quantized Low-Rank Adaptation* (QLoRA), que combina as duas estratégias: os pesos pré-treinados são quantizados em 4 bits e mantidos congelados, enquanto os adaptadores LoRA, mantidos em maior precisão, são treinados por meio da retropropagação dos gradientes através do modelo quantizado (Dettmers et al., 2023).

A abordagem introduz ainda o formato 4-bit *NormalFloat* (NF4), otimizado para pesos com distribuição normal, e a dupla quantização, que quantiza as próprias constantes de quantização, reduzindo ainda mais a pegada de memória (Dettmers et al., 2023). É exatamente essa combinação de quantização e LoRA que fundamenta as versões quantizadas dos modelos disponibilizadas pela biblioteca *Unsloth*, utilizadas na etapa de treinamento deste trabalho, viabilizando o ajuste fino de LLMs multimodais de bilhões de parâmetros em infraestruturas com recursos computacionais limitados.

Além da arquitetura dos modelos, a avaliação do desempenho de sistemas de classificação requer métricas capazes de quantificar objetivamente a qualidade das predições realizadas. Em tarefas de classificação binária, como a identificação de colidência visual entre marcas figurativas, as métricas mais empregadas são Acurácia, Precisão, *Recall* e *F1-Score*, pois fornecem uma visão complementar sobre o comportamento do modelo diante de acertos e erros de classificação (Powers, 2011) conforme representado na Figura 9 .

Figura 9 – Métricas de Avaliação para Classificação.

$$1. \text{ Acurácia (Accuracy)}_{\text{Acurácia}} = \frac{VP + VN}{VP + VN + FP + FN}$$

$$2. \text{ Precisão (Precision)}_{\text{Precisão}} = \frac{VP}{VP + FP}$$

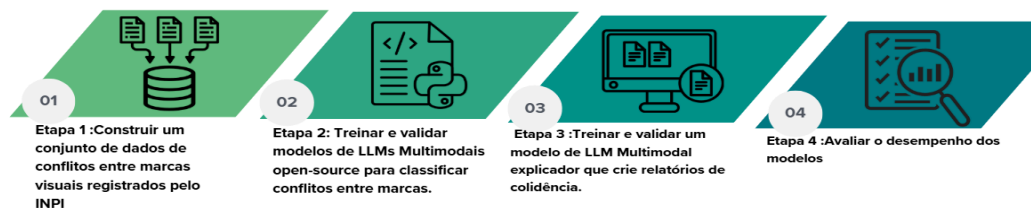
$$3. \text{ Recall (Sensibilidade/Recall)}_{\text{Recall}} = \frac{VP}{VP + FN}$$

$$4. \text{ F1-Score (F1-Medida)}_{F1} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}$$

Fonte: Elaborada pelo autor (2026)

3 METODOLOGIA

A execução do projeto seguiu uma metodologia estruturada, com ciclos iterativos de preparação de dados, treinamento e validação. O processo foi dividido em cinco etapas principais: (1) construção do conjunto de dados; (2) treinamento de modelos classificadores multimodais; (3) desenvolvimento do modelo explicador e (4) análise comparativa dos modelos, conforme ilustrado na figura 10.

Figura 10: Etapas dos objetivos

Fonte: Elaborado pelo autor (2026).

3.1 ETAPA 1 - CONJUNTO DE DADOS

Inicialmente, foi realizada a coleta e a estruturação do conjunto de dados. Como ponto de partida, empregou-se o CMAD, desenvolvido por Reis et al. (2024), que reúne casos reais de conflitos de marcas registrados no INPI. Entretanto, essa base é composta predominantemente por marcas nominativas e mistas, o que limita sua aplicação no treinamento de LLMs voltados à análise de conflitos envolvendo marcas estritamente figurativas. Isso ocorre porque, em pares que contêm elementos nominativos, a colidência pode decorrer de fatores fonéticos, textuais ou ideológicos, e não exclusivamente dos elementos gráficos.

Como o objetivo deste trabalho é a análise de colidência visual, optou-se por construir um conjunto de dados, denominado CMAD Visual, composto unicamente por pares de marcas figurativas, nas quais a similaridade resulta de aspectos como forma, silhueta, composição gráfica, símbolos e disposição dos elementos. Para a construção do CMAD Visual, adotou-se a mesma metodologia de coleta proposta por Reis et al. (2024), preservando suas etapas e ferramentas, porém com a inclusão de um critério de filtragem que restringe a base às marcas cuja forma de apresentação seja figurativa.

Dessa maneira, garante-se que tanto a marca indeferida quanto a marca já registrada sejam exclusivamente figurativas, assegurando que o rótulo de conflito reflita uma colidência de natureza visual. A elaboração do dataset seguiu a sequência de quatro etapas apresentada na Figura 11.

Figura 11: Etapas da Criação do Dataset



Fonte: Elaborado pelo autor (2026).

3.1.1 Download dos Arquivos XML

A primeira etapa consistiu na obtenção dos dados-fonte. A Revista da Propriedade Industrial (RPI) é o documento oficial publicado semanalmente pelo INPI, que dispõe de uma seção dedicada aos pedidos de registro de marcas e, em especial, aos despachos de indeferimento (INPI, 2023). Os arquivos correspondentes a esses despachos foram feito o download em um processo de *web scraping* implementado por meio de três scripts em Python.

O primeiro script realiza o acesso automatizado ao portal oficial do INPI (<http://revistas.inpi.gov.br/rpi/>), a fim de identificar o número da edição mais recente da RPI disponível na tabela de publicações. Com esse identificador, o segundo módulo do mesmo script constrói dinamicamente a URL do arquivo compactado correspondente, realiza o download do pacote e extrai o arquivo XML para uma pasta temporária local. Esses arquivos encontram-se no formato XML, estruturado de modo a viabilizar o processamento automático das informações. Cada arquivo XML contém os processos publicados na semana, incluindo os campos relativos ao número do processo, ao titular, à classificação e ao texto complementar do despacho, que é utilizado como entrada para o algoritmo da etapa seguinte.

O segundo script executa a extração e a filtragem dos dados relevantes a partir do XML. Para restringir a análise aos casos de colidência macária, aplica-se um filtro textual baseado na expressão-chave "A marca reproduz ou imita os seguintes registros de terceiros", descartando trechos anteriores que não são pertinentes ao estudo. A partir do trecho filtrado, empregam-se expressões regulares para identificar os números dos processos conflitantes citados no despacho. Os resultados são consolidados em arquivos CSV tabulados, contendo três colunas: o processo indeferido, os processos conflitantes e o texto do despacho filtrado, exemplo figura 12.

Figura 12: Arquivo CSV Contendo os Processos

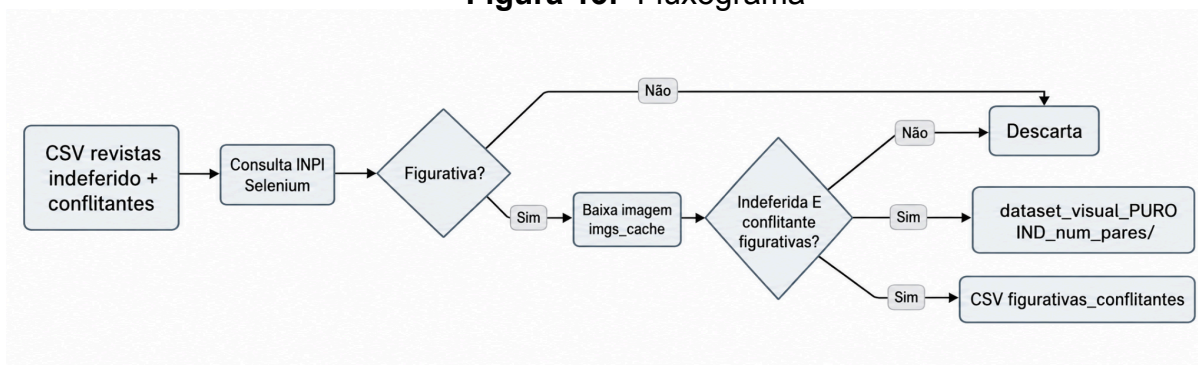
Processo	Indeferido	Processo	Conflitante	Texto do Despacho (Filtrado)
1				
2	936696494	926391968		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
3	936680547	817615898		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
4	936652616	924624159		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
5	936654438	910707430		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
6	936654708	912901160		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
7	936559080	911837213	911837272	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
8	936559080	911837337		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
9	936486015	929842537	929763009	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
10	936728698	926509756		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
11	936729287	921119330		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
12	936678348	902020358	932819443	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
13	936719761	933971702	933678207	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
14	936719761	933971761	933971583	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
15	936720298	920488935		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
16	936141778	900892285		"A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irreg
17	936414723	909136815		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
18	936415355	903626136		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
19	936387599	824817540		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
20	936378484	903670259		"A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irreg
21	936171758	823477002	936098430	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
22	936171758	900444967	936113693	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
23	936171758	900445068	936097760	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
24	936171758	936097710	922278555	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
25	936171758	922450706		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
26	936421312	927313499		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
27	936636130	901106275		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
28	936035323	916447669		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
29	936636262	919426522		A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregi
30	936541270	810056020	930724569	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
31	936287454	928387739	916447413	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port
32	936665483	908756283	907649700	A marca reproduz ou imita os seguintes registros de terceiros, sendo, port

Fonte: Elaborado pelo autor (2026).

O terceiro script orquestra a coleta em lote e integra os dois módulos anteriores. A partir da edição mais recente identificada, percorre retroativamente um conjunto definido de edições, por exemplo, as revistas anteriores, verificando previamente se o CSV correspondente já foi gerado, a fim de evitar reprocessamento. Para cada edição pendente, aciona sequencialmente o *download* do XML e a análise dos despachos.

3.1.2 Identificação de Conflitos e Coleta de Dados

A principal adaptação em relação à metodologia original foi introduzida nesta etapa: aplicou-se um filtro sobre a forma de apresentação das marcas, conforme ilustrado no fluxograma, representado na figura 13, de modo a reter apenas os processos cuja apresentação é do tipo figurativa, descartando-se as marcas nominativas e mistas. Esse procedimento delimita o escopo da base ao domínio puramente visual, coerente com o objetivo de analisar conflitos entre elementos gráficos e não entre denominações verbais isoladas.

Figura 13: Fluxograma

Fonte: Elaborado pelo autor (2026).

O arquivo recebe como entrada os arquivos CSV gerados na etapa anterior, armazenados na pasta revistas. A partir desses registros, o algoritmo realiza a coleta detalhada de metadados e imagens no portal público de consulta do INPI, por meio de automação com *Selenium WebDriver* (Chrome). Para cada número de processo, o robô acessa a página de busca, insere o pedido, abre a ficha detalhada e extrai, das tabelas HTML, os campos relativos ao número do processo, à denominação, à situação, à apresentação, à natureza, à classe de Nice, à classe nacional, à classificação de Viena e à data de depósito.

Após a leitura do campo Apresentação, aplica-se o critério de inclusão central desta etapa: apenas processos cuja apresentação é igual a "figurativa" seguem adiante; marcas nominativas, mistas ou sem apresentação identificada são descartadas imediatamente, sem download de imagem nem registro na base final. Para os processos aceitos, a imagem da logomarca é obtida diretamente do portal do INPI.

Esse cache global evita consultas repetidas ao portal quando o mesmo processo aparece em múltiplos conflitos ou em execuções subsequentes do script; complementarmente, mantém-se um cache em memória que registra, para cada processo já visitado, se é figurativo e qual o caminho da imagem correspondente.

A lógica de pareamento opera da seguinte forma: para cada linha do CSV de entrada, consulta-se uma única vez o processo indeferido; se ele for figurativo e possuir imagem disponível, itera-se sobre cada processo conflitante da linha, consultando-o individualmente e aplicando o mesmo filtro de apresentação.

Um par só é registrado quando ambas as marcas, indeferida e conflitante são figurativas e possuem imagem. Os pares válidos são organizados em duas saídas

complementares. Essa estrutura agrupa visualmente, por marca indeferida, todas as marcas anteriores com as quais ela colidiu, facilitando inspeção manual, rotulagem e treinamento de modelos de similaridade visual. Paralelamente, o algoritmo gera um arquivo CSV intermediário tabulado por revista. Cada linha representa um par candidato figurativa × figurativa, reunindo lado a lado os dados completos das duas marcas: processo, denominação, situação, apresentação, natureza, classes de Nice e nacional, classificação de Viena e data de depósito e revista de origem.

Ao final, essa etapa constitui a ponte entre a extração textual dos despachos e a análise visual das marcas: o CSV intermediário com os pares candidatos de marcas figurativas em conflito e as pastas organizadas em dataset visual puro servem de entrada para a coleta detalhada de dados e para as etapas subsequentes de comparação visual e geração textual.

3.1.3 Geração da base CMAD Visual

Concluída a coleta, cada revista processada deu origem a um arquivo intermediário com os pares candidatos daquela publicação. Esses arquivos foram consolidados em um único CSV, que passou a funcionar como índice mestre dos pares conflitantes do CMAD Visual. Para preservar a procedência de cada par, acrescentou-se a coluna arquivo origem, indicando a edição da RPI de onde foi extraído e permitindo identificar registros repetidos entre edições.

Ao final, a base CMAD Visual ficou constituída por dois conjuntos complementares de artefatos. O primeiro são as imagens das logomarcas, armazenadas em cache e organizadas por processo indeferido em dataset visual puro, com cada marca figurativa gravada em JPEG e indexada pelo número do processo. O segundo é a tabela unificada de pares conflitantes (tabela unificada.csv), que reúne lado a lado os metadados das duas marcas de cada par.

Cada artefato cumpre um papel distinto no fluxo do projeto: a tabela atua como índice, rótulo e instrumento de curadoria, determinando quais marcas colidem entre si e organizando a estrutura de diretórios; as imagens constituem a entrada efetiva do treinamento. Como o objetivo do trabalho é a análise de colidência estritamente visual, são os elementos gráficos e os metadados textuais que alimentam os modelos multimodais nas etapas seguintes. Dessa forma, a tabela unificada governa a montagem e a rotulagem dos pares, enquanto as imagens de

cache são o dado bruto efetivamente processado pelos LLMs multimodais durante o ajuste fino.

5.2. ETAPA 2 - TREINAMENTO MODELO CLASSIFICADOR MULTIMODAL

Nesta etapa, foram utilizados modelos LLMs *open-source* multimodais disponíveis na biblioteca *Unsloth* (Unsloth AI, 2025), que oferece suporte nativo à quantização de pesos e à integração com técnicas de Fine-tuning eficiente em parâmetros. Os modelos selecionados para os experimentos foram: Llama 3.2 Vision (Meta), Qwen 2 VL (Alibaba Cloud), Pixtral 12B (Mistral) e LLaVA (Microsoft), conforme detalhado na Tabela 1 juntamente com suas respectivas empresas desenvolvedoras e tamanhos em parâmetros.

Tabela 1 - Modelo de LLMs selecionados para avaliação

Modelo	Empresa	Tamanho	Referência
Llama 3.2 Vision	Meta	1B	https://huggingface.co/unsloth/Llama-3.2-1B-Vision-Instruct-bnb-4bit
Qwen 2 VI	Alibaba Cloud	2B	https://huggingface.co/unsloth/Qwen2-VL-7B-Instruct-bnb-4bit
Pixtral 12B	Mistral	12B	https://huggingface.co/unsloth/Pixtral-12B-2409-bnb-4bit
Llava	Microsoft	1,5 (7 B)	https://huggingface.co/unsloth/llava-1.5-7b-hf-bnb-4bit

Fonte: Elaborado pelo autor (2026).

A escolha da biblioteca *Unsloth* justifica-se por três fatores principais: disponibilização de versões quantizadas dos modelos, reduzindo os requisitos de memória GPU sem perdas significativas de desempenho; integração nativa com o framework de ajuste fino supervisionado e compatibilidade com a técnica de Adaptação de Baixa Ordem (*Low-Rank Adaptation*, LoRA). A adoção do LoRA elimina a necessidade de re-treinar a totalidade dos parâmetros do modelo pré-treinado, uma vez que insere matrizes de atualização de baixa dimensionalidade nos módulos de atenção, mantendo os pesos originais congelados e reduzindo substancialmente os custos computacionais associados ao treinamento (Hu et al., 2022; Rezaei et al., 2025).

Todos os modelos foram submetidos a um processo de ajuste fino supervisionado (*Supervised Fine-Tuning*, SFT), utilizando o conjunto de dados CMAD Visual construído na etapa anterior. O objetivo do treinamento foi especializar cada modelo na tarefa de classificação binária de similaridade visual entre pares de marcas figurativas, de modo que o modelo seja capaz de determinar se dois logotipos apresentam grau de semelhança suficiente para configurar colidência marcária. Os experimentos de ajuste fino foram conduzidos no *Google Colaboratory*, utilizando uma GPU NVIDIA T4 com 15 GB de memória de vídeo (VRAM), 12,7 GB de memória RAM e 112,6 GB de armazenamento. O ambiente de software baseou-se em *Python*, *Unsloth*, quantização de 4 bits foi realizada por meio da biblioteca *bitsandbytes*.

Cada par de marcas foi apresentado ao modelo como uma única imagem composta, na qual a marca indeferida (Marca A) e a marca registrada anterior (Marca B) são posicionadas lado a lado, respectivamente à esquerda e à direita. Essa imagem foi acompanhada de uma instrução textual idêntica nos quatro experimentos, tanto no treinamento quanto na inferência, reproduzida no Quadro 1

Na primeira frente, o desempenho do modelo classificador foi mensurado de forma quantitativa por meio das métricas Acurácia, Precisão, *Recall* e *F1-Score*, conforme observa-se na figura 9, amplamente adotadas na literatura para tarefas de classificação (Powers, 2011). Essas métricas permitiram aferir, de maneira objetiva, a capacidade do modelo em identificar corretamente pares de marcas figurativas em situação de colidência visual.

3.3 DESENVOLVIMENTO DO MODELO EXPLICADOR MULTIMODAL

Concluída a etapa de classificação, o modelo que apresentou o melhor desempenho nas métricas de Acurácia, Precisão, *Recall* e *F1-Score* foi selecionado para a fase de desenvolvimento do modelo explicador. Nessa etapa, o objetivo é gerar relatórios técnicos detalhados que fundamentam a decisão em linguagem acessível e estruturada, alinhada aos critérios de análise utilizados pelo INPI.

A estratégia adotada combinou dois elementos principais: (1) um *prompt* de sistema (*system prompt*), responsável por definir o papel do modelo e as diretrizes gerais de análise; e (2) um *prompt* de usuário estruturado, que delimita o formato esperado da resposta e os critérios que devem ser abordados em cada relatório. O *prompt* de sistema foi elaborado com o objetivo de posicionar o modelo como um especialista em análise de marcas do INPI, conferindo-lhe o contexto institucional necessário para produzir justificativas tecnicamente coerentes com as diretrizes do Manual de Marcas (INPI, 2023). O Quadro 1 apresenta o *prompt* de sistema utilizado.

Quadro 1 — Prompt de Sistema (*System Prompt*) do Modelo Explicador

Você é um especialista em propriedade intelectual com profundo conhecimento nas diretrizes do Instituto Nacional da Propriedade Industrial (INPI). Sua função é analisar pares de marcas figurativas e gerar relatórios técnicos detalhados sobre a existência ou ausência de colidência visual entre elas, fundamentando sua análise nos critérios estabelecidos pelo Manual de Marcas do INPI. Sua análise deve ser clara, objetiva e acessível a profissionais da área jurídica e administrativa. Considere sempre a impressão geral que os conjuntos causam no consumidor medianamente atento, avaliando simultaneamente os aspectos gráficos, de composição, de forma, de silhueta, cromáticos e de disposição dos elementos visuais.

Fonte: Elaborado pelo autor (2026).

O *prompt* de usuário foi estruturado para apresentar o par de imagens ao modelo e delimitar os critérios de análise que devem ser contemplados na resposta. A estrutura adotada orienta o modelo a analisar cinco dimensões visuais específicas e a emitir uma conclusão fundamentada, em consonância com os aspectos avaliados pelo INPI na análise de colidência de marcas figurativas (INPI, 2023). O Quadro 2 apresenta o template de *prompt* de usuário utilizado.

Quadro 2 — Template do Prompt de Usuário

Sua análise deve contemplar obrigatoriamente os seguintes aspectos:

1. **Forma e Silhueta:** Descreva as formas geométricas e a silhueta de cada marca. Identifique semelhanças e diferenças estruturais entre os conjuntos.
2. **Composição Gráfica:** Analise a disposição dos elementos visuais, as proporções e a estrutura compositiva de cada sinal.
3. **Elementos Visuais Específicos:** Identifique e compare símbolos, traços, figuras e demais elementos gráficos presentes em ambas as marcas.
4. **Paleta de Cores:** Avalie a semelhança cromática entre os sinais, considerando que marcas em preto e branco são comparadas em sua forma, independentemente da cor.
5. **Impressão Geral do Consumidor:** Considere se um consumidor medianamente atento poderia confundir ou associar indevidamente as marcas ao visualizá-las separadamente.
6. **Conclusão:** Indique se há ou não colidência visual e fundamente sua decisão com base nos elementos analisados.

Responda seguindo obrigatoriamente o seguinte formato:

ANÁLISE DE COLIDÊNCIA VISUAL

Descrição da Marca Indeferida: [descrição objetiva dos elementos visuais]

Descrição da Marca Registrada: [descrição objetiva dos elementos visuais]

1. **Forma e Silhueta:** [análise comparativa]
2. **Composição Gráfica:** [análise comparativa]
3. **Elementos Visuais Específicos:** [análise comparativa]
4. **Paleta de Cores:** [análise comparativa]
5. **Impressão Geral do Consumidor:** [análise]
6. **Conclusão:** [COLIDÊNCIA IDENTIFICADA / SEM COLIDÊNCIA IDENTIFICADA]
— [justificativa técnica resumida]

Fonte: Elaborado pelo autor (2026).

Para orientar o modelo quanto ao padrão de linguagem técnica, à profundidade das análises e à estrutura narrativa esperada, foram inseridos no contexto do *prompt* dois exemplos de relatórios previamente gerados pelo ChatGPT 5.5. válidos por especialista em PI. Essa estratégia, conhecida como few-shot prompting (Brown et al., 2020; Ma et al., 2023), consiste em fornecer exemplos de

entrada e saída para induzir o modelo a reproduzir padrões desejados de resposta, sendo amplamente utilizada para melhorar a consistência e a qualidade das saídas produzidas por grandes modelos de linguagem em tarefas especializadas.

Os exemplos foram construídos para representar dois cenários distintos da tarefa de classificação: um caso de colidência confirmada e um caso de ausência de colidência. Os *prompts* desenvolvidos para a tarefa de classificação na primeira fase foram submetidos ao GPT-5.5o. Os quadros 3 e 4 apresentam os exemplos de few-shot utilizados.

Quadro 3 — Exemplo *Few-Shot* 1: Colidência Visual Identificada

ANÁLISE DE COLIDÊNCIA VISUAL

Descrição da Marca Indeferida: Figura de criatura fantástica alada posicionada em perfil voltado para a direita, composta por traços robustos em preto, com destaque para a envergadura das asas abertas e a expressividade do conjunto.

Descrição da Marca Registrada: Figura de ave de rapina com asas abertas em posição de voo, representada em silhueta negra com traços dinâmicos e composição centralizada.

1. Forma e Silhueta: Ambas as marcas apresentam silhuetas de criaturas aladas com envergadura de asas pronunciada como elemento visual dominante. A semelhança na forma geral das asas abertas, na orientação da figura e no contorno do corpo cria impressão visual próxima, capaz de aproximar os conjuntos marcários perante o público consumidor.
2. Composição Gráfica: As duas marcas adotam composição centralizada com figura única, preenchimento sólido em preto e proporções similares entre o corpo e a envergadura das asas. A ausência de elementos textuais ou ornamentais adicionais reforça a concentração da percepção visual nos elementos figurativos centrais.
3. Elementos Visuais Específicos: Os elementos de asas abertas, corpo robusto e postura erguida são compartilhados como elementos dominantes em ambas as marcas. A presença de detalhes como penas estilizadas e membros inferiores, ainda que com traços distintos, não elimina a aproximação visual entre os conjuntos.
4. Paleta de Cores: Ambas as marcas utilizam exclusivamente a cor preta sobre fundo branco, eliminando a paleta cromática como fator de distinção entre os sinais marcários.
5. Impressão Geral do Consumidor: Um consumidor medianamente atento, ao visualizar as marcas em momentos distintos, poderia associá-las como provenientes da mesma origem empresarial, dado o conjunto visual semelhante formado pela criatura alada em silhueta escura com estrutura compositiva similar.
6. Conclusão: COLIDÊNCIA IDENTIFICADA — As marcas compartilham elementos visuais dominantes de forma, silhueta, composição e paleta cromática que, avaliados em conjunto, configuram risco de confusão e associação indevida pelo público consumidor, nos termos do art. 124, XIX, da LPI.

Fonte: Elaborado pelo autor (2026).

Quadro 4 — Exemplo *Few-Shot* 2: Sem Colidência Visual Identificada

ANÁLISE DE COLIDÊNCIA VISUAL

Descrição da Marca Indeferida: Representação estilizada de um leão em posição frontal, composta por linhas geométricas angulares com juba proeminente e traços modernos em preto.

Descrição da Marca Registrada: Composição circular formada por elementos florais estilizados, com traços delicados e simétricos dispostos radialmente em torno de um núcleo central, em preto.

1. Forma e Silhueta: As marcas apresentam silhuetas fundamentalmente distintas. Enquanto a marca indeferida exibe figura animal com forma orgânica e verticalidade marcada, a marca registrada adota composição circular e simétrica de natureza ornamental. Não há aproximação visual relevante entre as silhuetas dos conjuntos.
2. Composição Gráfica: A composição da marca indeferida é assimétrica e dinâmica, centrada na representação figurativa do animal. A marca registrada apresenta composição radialmente simétrica, de caráter decorativo, sem elementos figurativos de origem animal. As estruturas compositivas são marcadamente diferentes.
3. Elementos Visuais Específicos: Os elementos dominantes de cada marca são de natureza completamente distinta: figura animal versus padrão floral ornamental. Não há elementos gráficos comuns que possam estabelecer aproximação visual entre os conjuntos marcários.
4. Paleta de Cores: Ambas as marcas utilizam o preto como cor predominante; contudo, a semelhança cromática isolada não é suficiente para caracterizar colidência quando os demais elementos visuais são significativamente distintos entre si.
5. Impressão Geral do Consumidor: A impressão geral produzida pelas marcas é notadamente diferente. A marca indeferida remete a força e identidade animal, enquanto a marca registrada transmite sofisticação ornamental. Um consumidor medianamente atento não confundiria os conjuntos nem os associaria à mesma origem empresarial.
6. Conclusão: SEM COLIDÊNCIA IDENTIFICADA — As marcas apresentam elementos visuais dominantes de natureza distinta, silhuetas incompatíveis e impressões gerais divergentes, não configurando risco de confusão ou associação indevida pelo público consumidor.

Fonte: Elaborado pelo autor (2026).

Após a definição dos prompts e dos exemplos de *few-shot*, o modelo explicador foi executado sobre o conjunto de teste do CMAD Visual, buscando gerar automaticamente um relatório textual estruturado para cada par de marcas figurativas avaliado.

3.4 AVALIAÇÃO DO DESEMPENHO DOS MODELOS

Na segunda frente, os relatórios gerados pelo modelo explicador passaram por uma validação qualitativa conduzida por especialistas em PI. Para operacionalizar essa avaliação, foi desenvolvido um formulário no *Google Forms*, no qual cada item apresenta, lado a lado, o par de imagens de marcas analisadas e a justificativa textual gerada automaticamente pelo modelo. Dessa forma, o avaliador tinha acesso simultâneo ao insumo visual e à explicação produzida, contextualizando sua análise de maneira integrada.

. As respostas dos quatros especialistas em PI foram coletadas por meio da Escala *Likert* de cinco pontos, variando de "Discordo totalmente" a "Concordo totalmente", possibilitando a quantificação do grau de concordância dos especialistas em relação à qualidade das explicações geradas.

4 RESULTADOS E DISCUSSÃO

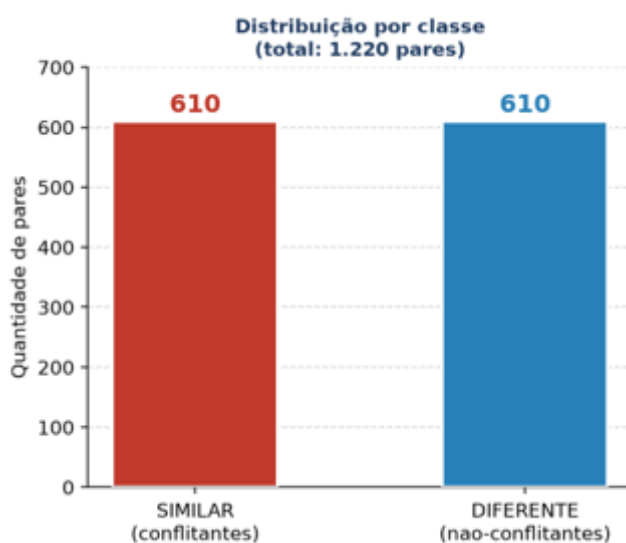
4.1 CONSTRUÇÃO DO CONJUNTO DE DADOS CMAD VISUAL

A raspagem das RPI resultou em 610 pares de conflitos figurativos, ou seja, casos em que o INPI reconheceu oficialmente a colidência visual entre as marcas. Esses pares constituem as amostras positivas (SIMILAR) do conjunto de dados, representação .

Para que o modelo aprendesse a distinguir marcas colidentes de marcas distintas, foi necessário compor também um conjunto de amostras negativas (DIFERENTE), formado por pares de marcas figurativas entre as quais não há colidência reconhecida pelo INPI. Esses pares negativos foram gerados a partir do próprio acervo de marcas figurativas coletadas, combinando-se aleatoriamente marcas que não constam como conflitantes entre si nos despachos do INPI. Ao manter as classes "SIMILAR" e "DIFERENTE" com a mesma quantidade de amostras, evita-se que o desbalanceamento de classes enviesasse o treinamento e a avaliação dos modelos.

Somando-se os 610 pares conflitantes (SIMILAR) aos 610 pares não conflitantes (DIFERENTE), o CMAD Visual totaliza 1.220 pares. Esses pares foram montados a partir de um acervo de 1.079 imagens de logomarcas, das quais 741 correspondem a marcas figurativas únicas efetivamente pareadas na base. Cabe observar que nem toda marca figurativa coletada pôde integrar um par: quando o único conflito registrado para uma marca figurativa se dava com uma marca mista ou nominativa, não era possível formar um par figurativa × figurativa, e tais casos foram descartados, coerentemente com o objetivo de restringir a base à colidência de natureza estritamente visual. A Figura 14 apresenta a composição e a divisão do conjunto de dados.

Figura 14 — Composição e divisão do CMAD Visual



Fonte: Elaborado pelo autor (2026).

A consolidação da base resultou em um arquivo CSV mestre com 610 registros e 21 colunas, no qual cada linha corresponde a um par de conflito reconhecido pelo INPI (classe SIMILAR), reunindo lado a lado as informações estruturadas da marca indeferida e da marca registrada que motivou o indeferimento. Esse arquivo funciona como índice dos pares positivos e como instrumento de rastreabilidade e rotulagem. Os 610 pares negativos (classe DIFERENTE) não integram esse índice mestre, por não decorrerem de despachos de indeferimento: eles são gerados separadamente, por combinação aleatória das

marcas figurativas do acervo, no momento da montagem do conjunto de treinamento e avaliação. Dessa forma, o total de 1.220 pares que compõem o CMAD Visual resulta da união dos 610 pares positivos catalogados no CSV mestre com os 610 pares negativos produzidos à parte. A Tabela 2 apresenta as colunas geradas.

Tabela 2 — Estrutura da tabela *Dataset.csv*

Coluna	Descrição
N Processo	Número do processo da marca indeferida (polo requerente)
Marca Indeferida	Nome/denominação da marca indeferida
Situação	Situação da marca indeferida
Apresentação	Tipo de apresentação (Figurativa — único tipo no dataset)
Natureza	Natureza da marca (De Produto / De Serviço)
Classe de Nice	Classificação Internacional de Niza do polo indeferido
Classe Nacional	Classificação nacional de produtos/serviços
Classificação de Viena	Códigos de Viena dos elementos figurativos (polo indeferido)
Data de Depósito	Data de depósito do pedido de registro indeferido
N Processo.1	Número do processo da marca registrada (polo bloqueador)
Marca Registrada	Nome/denominação da marca registrada anteriormente
Situação.1	Situação administrativa da marca registrada
Apresentação.1	Tipo de apresentação da marca registrada
Natureza.1	Natureza da marca registrada
Classe de Nice.1	Classificação de Nice da marca registrada
Classe Nacional.1	Classificação nacional da marca registrada
Classificação de Viena.1	Códigos de Viena dos elementos figurativos (polo registrado)
Data de Depósito.1	Data de depósito da marca registrada
Revista	Edição da RPI de origem
Visual	Indicador de colidência visual confirmada pelo INPI
arquivo_origem	Arquivo CSV de origem — rastreabilidade por edição

Fonte: Elaborado pelo autor (2026).

As colunas referentes ao número do processo, situação administrativa, natureza da marca, classes de Nice e classificação nacional permitem contextualizar juridicamente cada conflito, enquanto os campos de data de depósito possibilitam análises temporais e a identificação da anterioridade registral das marcas. Adicionalmente, a coluna *Revista* registra a edição da RPI da qual o conflito foi extraído, e a coluna preserva a procedência dos dados, garantindo rastreabilidade e permitindo a reconstituição do processo de coleta.

Entre os atributos armazenados, destaca-se a Classificação de Viena, presente em ambos os conflitos. Esse sistema de classificação, adotado internacionalmente pela OMPI, organiza os elementos figurativos das marcas em categorias, divisões e seções padronizadas. Dessa forma, símbolos, animais, figuras humanas, formas geométricas e outros componentes visuais são representados por códigos específicos, permitindo descrever semanticamente o conteúdo gráfico das logomarcas.

Dessa maneira, além da percepção visual fornecida pelos logotipos, o modelo passou a receber uma descrição estruturada dos elementos gráficos, favorecendo a geração de relatórios mais detalhados e tecnicamente consistentes.

Assim, o Dataset não atua apenas como uma tabela de armazenamento dos pares conflitantes, mas como um índice estruturado capaz de integrar informações administrativas, temporais e visuais, servindo tanto para o treinamento dos modelos classificadores quanto para o suporte à geração automática de explicações fundamentadas sobre os conflitos marcários, representação na Figura 15 do CMAD visual.

Figura 15 — CMAD Visual



Fonte: Elaborado pelo autor (2026).

4.2 DESEMPENHO DOS MODELOS CLASSIFICADORES

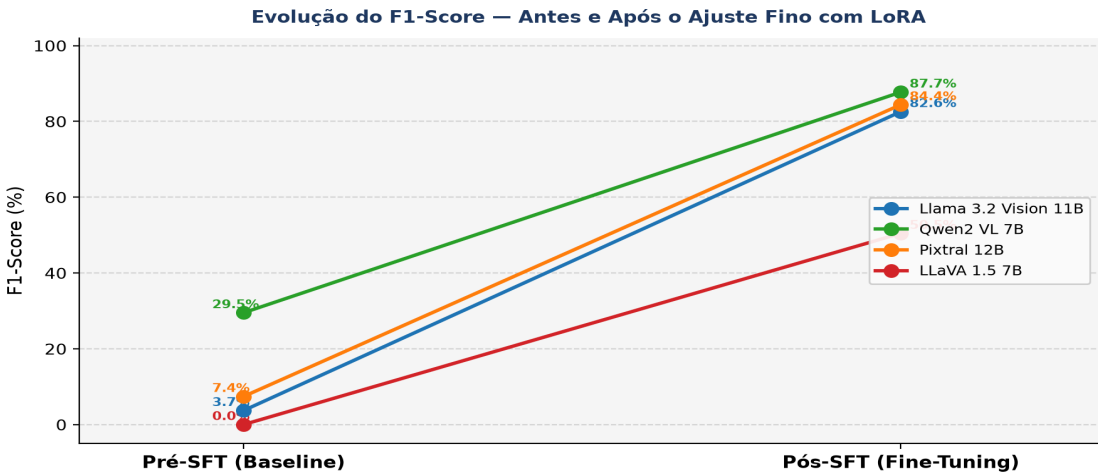
Foram avaliados quatro modelos LLMs multimodais open-source após ajuste fino supervisionado com LoRA sobre o CMAD Visual. As métricas de Acurácia, Precisão, *Recall* e *F1-Score* foram registradas antes e após o SFT, permitindo mensurar o ganho de desempenho. A Tabela 3 sintetiza os resultados antes e após o ajuste fino, e a Figura 16 apresenta a evolução do *F1-Score* para cada modelo.

Tabela 3 — Métricas dos modelos após o ajuste fino supervisionado (SFT)

Modelo	Acurácia (antes → depois)	Precisão (antes → depois)	Recall (antes → depois)	F1 (antes → depois)
Llama 3.2 Vision 11B	48.0%→81.0%	33.3%→77.6%	2.0%→88.2%	3.7%→82.6%
Qwen2 VL 7B	57.0%→86.0%	90.0%→79.4%	17.6%→98.0%	29.5%→87.7%
Pixtral 12B	50.0%→83.0%	66.7%→79.3%	3.9%→90.2%	7.4%→84.4%
LLaVA 1.5 7B	48.0%→45.0%	0.0%→46.7%	0.0%→54.9%	0.0%→50.5%

Fonte: Elaborado pelo autor (2026).

Figura 16 — Evolução do F1-Score antes e após o SFT com LoRA



Fonte: Elaborado pelo autor (2026).

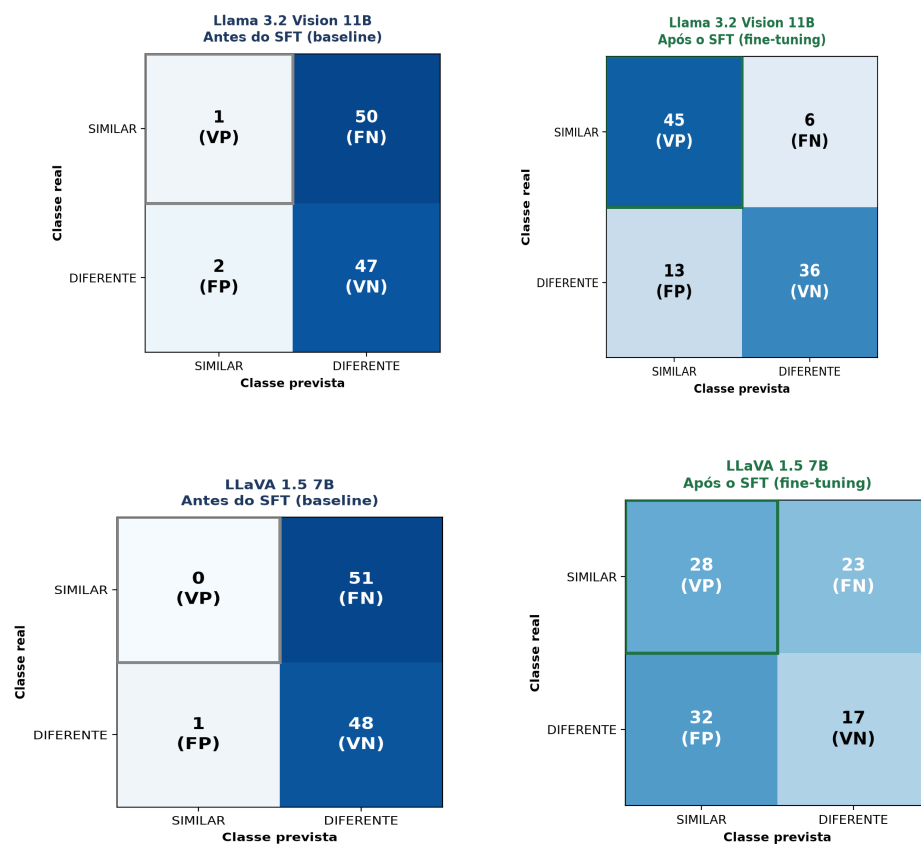
A análise dos resultados evidencia que o processo de ajuste fino supervisionado produziu ganhos expressivos em três dos quatro modelos. O Qwen2 VL 7B destacou-se como o melhor classificador, alcançando Acurácia de 86,0%, F1-Score de 87,72% e, notavelmente, *Recall* de 98,04% após o SFT. Esse elevado *Recall* indica que o modelo identifica corretamente quase a totalidade dos pares em situação de colidência visual, critério de maior relevância prática no contexto do INPI, onde a não identificação de um conflito real representa risco jurídico e econômico superior ao de um falso positivo.

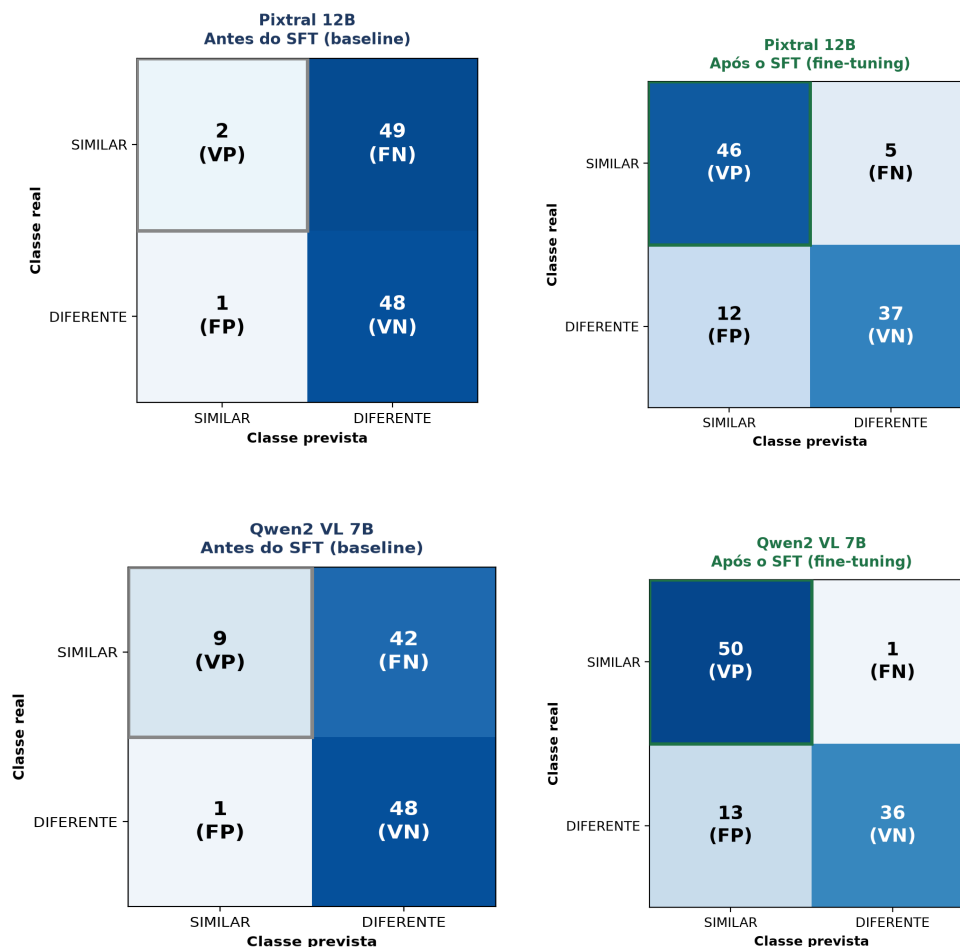
O Pixtral 12B obteve o segundo melhor desempenho, com F1-Score de 84,40% e *Recall* de 90,20%. O Llama 3.2 Vision 11B atingiu F1 de 82,57% e Acurácia de 81,0%. Ambos demonstraram forte capacidade de adaptação ao domínio de marcas figurativas mediante fine-tuning eficiente com LoRA.

O LLaVA 1.5 7B apresentou resultados significativamente inferiores, com *F1-Score* de 50,45% após o SFT e Acurácia de 45,0%, valor inferior à própria baseline (48,0%). Esse comportamento sugere que a arquitetura LLaVA possui menor capacidade de adaptação a tarefas de comparação visual especializada por meio de LoRA, possivelmente em razão de limitações no alinhamento entre seu encoder visual e o decoder de linguagem.

As matrizes de confusão (Figura 17) mostram a distribuição exata de acertos e erros nos pares de avaliação.

Figura 17 — Matrizes de confusão antes e após o SFT (100 pares por modelo)





Fonte: Elaborado pelo autor (2026).

A análise das matrizes de confusão revela um padrão consistente entre os modelos avaliados: antes do ajuste fino, todos apresentaram forte viés conservador, classificando a maioria dos pares como diferente comportamento evidenciado pelo baixíssimo número de Verdadeiros Positivos (VP) e pelo elevado número de Falsos Negativos (FN). Esse resultado indica que, em sua configuração pré-treinada, os modelos tendem a subestimar a ocorrência de colidência visual entre marcas figurativas.

Após o SFT, observa-se uma reversão expressiva desse padrão, especialmente nos modelos Qwen2 VL 7B, Pixtral 12B e Llama 3.2 Vision 11B. O Qwen2 VL 7B apresentou o ganho mais significativo, reduzindo os Falsos Negativos de 42 para apenas 1, em um conjunto de 100 pares de teste, o que se traduz no elevado *Recall* de 98,04% reportado na Tabela 3. Esse resultado é particularmente relevante no contexto do INPI, uma vez que a falha em identificar uma colidência

real (Falso Negativo) representa um risco jurídico e econômico mais crítico do que um Falso Positivo, que tende a ser corrigido em etapas posteriores de revisão.

O Pixtral 12B e o Llama 3.2 Vision 11B seguiram tendência semelhante, com reduções expressivas nos FN de 49 para 5 e de 50 para 6, respectivamente, confirmando que o ajuste fino supervisionado com LoRA foi eficaz em corrigir o viés conservador inicial e em especializar esses modelos para a tarefa de comparação visual.

Em contraste, o LLaVA 1.5 7B manteve um número elevado de erros em ambas as classes mesmo após o SFT, sugerindo um quadro de subajuste (*underfitting*) ao domínio. Esse comportamento aponta para limitações estruturais do modelo, possivelmente relacionadas ao alinhamento entre seu encoder visual e o decoder de linguagem, que o tornaram menos responsivo ao processo de especialização via LoRA, mesmo diante do mesmo conjunto de dados e da mesma estratégia de treinamento aplicada aos demais modelos.

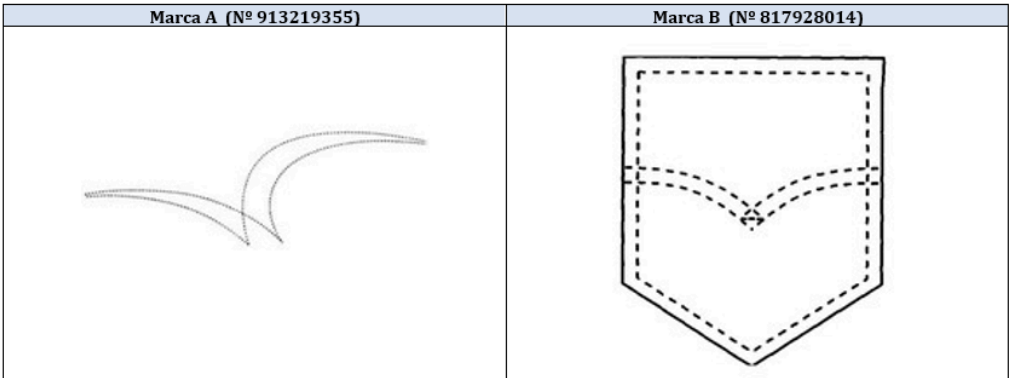
4.4 ANÁLISE QUALITATIVA DO MODELO EXPLICADOR

Conforme apresentado na Seção 4.2, o modelo Qwen2 VL 7B obteve o melhor desempenho entre os classificadores avaliados, superando os demais em todas as métricas mensuradas após o ajuste fino supervisionado (Acurácia de 86,0%, Precisão de 79,4%, *Recall* de 98,0% e *F1-Score* de 87,7%). Em razão desse resultado, esse foi o único modelo conduzido para a etapa de desenvolvimento do explicador multimodal, conforme descrito na Seção 5.3, sendo responsável por gerar os relatórios técnicos de colidência visual analisados a seguir.

Para essa etapa, o modelo foi executado sobre pares de marcas figurativas do conjunto de teste do CMAD Visual, recebendo como entrada o *prompt* de sistema Quadro 1, o template de *prompt* de usuário Quadro 2 e os dois exemplos de few-shot Quadros 3 e 4, conforme o fluxo de inferência detalhado na Seção 5.3. A cada par de imagens fornecido, a marca indeferida e a marca registrada apontada como conflitante, o modelo produziu como saída um relatório estruturado em seis tópicos, contemplando a descrição objetiva de cada marca, a análise comparativa de forma e silhueta, composição gráfica, elementos visuais específicos, paleta de cores e impressão geral do consumidor, encerrando com uma conclusão fundamentada sobre a existência ou não de colidência visual.

A figura 18 apresenta um dos exemplos real de par de marcas figurativas extraído do conjunto de teste, acompanhado do relatório efetivamente gerado pelo Qwen2 VL 7B para esse par, ilustrando o tipo de explicação produzida pelo modelo na prática.

Figura 18 — Relatório de Colidência Visual Gerado pelo Modelo Explicador (Qwen2 VL 7B)



A análise visual revela que as marcas possuem elementos figurativos da mesma família temática segundo a Classificação de Viena, suficientes para gerar associação indevida na percepção do consumidor. Adicionalmente, ambas atuam em classes mercadológicas próximas (NCL: NCL(10) 25 vs NCL(8) 25), o que amplifica o risco de confusão quanto à origem comercial dos produtos/serviços. Portanto, considera-se que **HÁ COLIDÊNCIA VISUAL** entre as marcas, sendo cabível o indeferimento do pedido de registro mais recente.

Fonte: Elaborado pelo autor (2026).

O caso apresentado evidencia características importantes do comportamento do modelo explicador. Em primeiro lugar, observa-se que o modelo manteve a estrutura de conclusão exigida pelo *prompt* de usuário (Quadro 2), emitindo um veredito explícito, COLIDÊNCIA IDENTIFICADA, seguido de uma justificativa técnica resumida, em conformidade com o formato de saída definido na Seção 5.3. Esse resultado indica que a engenharia de prompt, combinada à estratégia de *few-shot prompting*, foi eficaz em induzir o modelo a reproduzir o padrão estrutural estabelecido pelos exemplos validados (Quadros 3 e 4), mesmo diante de um par de marcas com elementos figurativos distintos dos utilizados como referência.

Em segundo lugar, chama atenção o fato de que a fundamentação apresentada pelo modelo não se restringiu exclusivamente aos aspectos visuais

propriamente ditos , forma, silhueta, composição gráfica e elementos figurativos , mas incorporou também informações administrativas complementares, como a Classificação de Viena e a Classificação de Nice (NCL) dos dois processos. Esse comportamento confirma o que foi descrito anteriormente: ao incluir os códigos da Classificação de Viena no *prompt* como informação textual complementar às imagens, o modelo passou a articular a percepção visual dos logotipos com metadados estruturados sobre a natureza dos elementos figurativos e a proximidade mercadológica entre os pedidos, produzindo uma justificativa mais próxima dos critérios multifatoriais efetivamente empregados pelo INPI na análise de colidência (INPI, 2023).

A validação qualitativa dos relatórios gerados pelo modelo explicador foi conduzida por quatro especialistas em PI, que avaliaram um conjunto de seis relatórios de colidência visual produzidos pelo Qwen2 VL 7B. Para cada relatório, os avaliadores atribuíram uma nota em Escala Likert de cinco pontos, variando de 1 (Discordo totalmente) a 5 (Concordo totalmente), além de poderem registrar comentários abertos sobre a explicação apresentada. Ao todo, foram coletadas 24 avaliações, cuja síntese é apresentada na Tabela 4.

Tabela 4 — Avaliação dos especialistas por relatório (Escala Likert de 1 a 5)

Relatório	Esp. 1	Esp. 2	Esp. 3	Esp. 4	Média
Relatório 1	5	4	5	5	4,75
Relatório 2	1	3	2	3	2,25
Relatório 3	5	1	5	5	4,00
Relatório 4	2	5	4	4	3,75
Relatório 5	5	5	5	5	5,00
Relatório 6	5	5	5	5	5,00
Média geral					4,13

Fonte: Elaborado pelo autor (2026).

O conjunto das avaliações revela uma percepção predominantemente positiva quanto à qualidade das explicações geradas pelo modelo. Das 24 notas atribuídas, 18 (75%) situam-se nos níveis 4 e 5 da escala, sendo que 15 delas (62,5%) corresponderam à pontuação máxima. A média geral de 4,13 pontos indica que, de modo geral, os especialistas consideraram os relatórios claros, tecnicamente corretos e alinhados aos critérios de análise de colidência empregados pelo INPI.

O desempenho, contudo, não foi homogêneo entre os relatórios. Os Relatórios 5 e 6 alcançaram concordância unânime, recebendo nota máxima de

todos os avaliadores (média 5,0), o que sugere que, nesses casos, a justificativa textual produzida pelo modelo foi percebida como plenamente adequada. O Relatório 1 também obteve avaliação elevada (média 4,75). Os comentários associados a esses relatórios reforçam essa percepção: um dos especialistas considerou a justificativa correta em razão do "alto risco de confundir o consumidor devido à grande similaridade entre ambas", enquanto outro destacou a coerência da conclusão de ausência de colidência, observando que, embora ambas as marcas apresentassem círculos, a presença de um formato em "S" em apenas uma delas justifica a distinção entre os sinais.

Em contrapartida, os Relatórios 2, 3 e 4 apresentaram maior dispersão entre as notas, evidenciando divergência de opinião entre os especialistas. O Relatório 2 obteve a menor média (2,25) e concentrou as principais críticas: um avaliador registrou não enxergar similaridade visual entre as marcas, e outro apontou que faltou descrever melhor as semelhanças visuais, considerando a análise "genérica demais". O Relatório 3 foi o mais controverso, recebendo nota máxima de três avaliadores e nota mínima de um quarto especialista, que identificou uma falha relevante: embora o modelo tenha acertado quanto à ausência de colidência, sua justificativa mencionou elementos inexistentes nas imagens (peças de roupas e calçados), quando as figuras apresentadas correspondiam apenas a uma lebre e a pegadas.

Esse último apontamento é particularmente significativo, pois ilustra o fenômeno da alucinação descrito na Seção 2.3, no qual o modelo gera conteúdo linguisticamente coerente, porém factualmente inconsistente com a entrada fornecida. Ainda que o veredito de classificação tenha sido correto, a descrição equivocada dos elementos visuais compromete a confiabilidade técnica do relatório e reforça a necessidade de revisão humana na etapa de fundamentação, ademais ajustes dos pesos. De modo geral, os resultados indicam que o modelo explicador é capaz de produzir justificativas estruturadas e bem avaliadas na maioria dos casos, mas ainda apresenta limitações pontuais na precisão descritiva dos elementos gráficos, especialmente em pares de marcas com baixa similaridade visual.

5 CONSIDERAÇÕES FINAIS

Diante do crescente volume de pedidos submetidos ao INPI, e considerando que a similaridade visual figura entre as principais causas de indeferimento, a

ausência de ferramentas capazes de aliar elevada capacidade de classificação a explicações compreensíveis sobre suas decisões representa uma lacuna relevante tanto do ponto de vista técnico quanto institucional. Foi nesse contexto que se inseriu o problema de pesquisa investigado neste trabalho.

Este trabalho consistiu em desenvolver e avaliar um método baseado em LLMs para apoiar a identificação de colidência visual entre marcas figurativas. Os resultados obtidos ao longo da pesquisa demonstram que esse objetivo foi alcançado, tendo sido igualmente cumpridos os objetivos específicos previamente estabelecidos: construiu-se um conjunto de dados de conflitos visuais entre marcas registradas pelo INPI; treinaram-se e validaram-se modelos multimodais open-source voltados à classificação de colidência; desenvolveu-se um modelo explicador capaz de gerar relatórios técnicos fundamentados; e realizou-se a análise comparativa do desempenho dos modelos avaliados.

Entre as contribuições do trabalho, destacam-se quatro resultados principais. O primeiro é a construção do CMAD Visual, base derivada do CMAD original e composta exclusivamente por pares de marcas figurativas, totalizando 1.220 pares, 610 casos de colidência reconhecida pelo INPI e 610 pares sem colidência, a partir de 1.079 imagens e 741 marcas figurativas únicas, acompanhados de metadados estruturados como a Classificação de Viena e a Classificação de Nice. O segundo é o ajuste fino supervisionado, com a técnica LoRA, de quatro LLMs multimodais (Llama 3.2 Vision, Qwen2 VL, Pixtral e LLaVA), permitindo mensurar de forma objetiva o ganho de desempenho proporcionado pela especialização no domínio.

O terceiro é a identificação do Qwen2 VL 7B como o modelo de melhor desempenho, com Acurácia de 86,0%, Precisão de 79,4%, *Recall* de 98,0% e *F1-Score* de 87,7%, evidenciando que a qualidade do pré-treinamento multimodal pode ser mais determinante para a tarefa do que o número absoluto de parâmetros. O quarto é o desenvolvimento de um modelo explicador que, combinando engenharia de *prompt* e *few-shot prompting*, produziu relatórios técnicos estruturados que obtiveram avaliação predominantemente positiva por parte de especialistas em propriedade intelectual, com média geral de 4,13 em uma escala Likert de cinco pontos.

Quanto ao seu impacto, o método proposto oferece subsídios concretos para a modernização do processo de exame de marcas. No âmbito institucional, pode atuar como ferramenta de apoio ao INPI e a escritórios de propriedade intelectual,

reduzindo o tempo de análise, aumentando a transparência das decisões e mitigando parte da subjetividade inerente à avaliação visual. Em vez de substituir o julgamento humano, a solução posiciona-se como instrumento de suporte ao examinador, fornecendo uma classificação inicial acompanhada de justificativa textual fundamentada. Adicionalmente, a própria construção do CMAD Visual constitui um ativo de pesquisa relevante, ao disponibilizar uma base estruturada de conflitos visuais entre marcas figurativas que pode subsidiar estudos futuros na área.

Apesar dos resultados alcançados, o trabalho apresenta limitações que devem ser reconhecidas: foram avaliados apenas quatro LLMs multimodais de código aberto, todos treinados utilizando a estratégia LoRA. Dessa forma, os resultados obtidos não permitem afirmar que o desempenho observado se estenda a outras arquiteturas, modelos proprietários ou diferentes técnicas de ajuste fino.

A validação qualitativa também apresenta limitações. A avaliação dos relatórios foi realizada por quatro especialistas em propriedade intelectual, número suficiente para uma análise exploratória, mas ainda reduzido para permitir conclusões estatísticas mais robustas sobre a qualidade das explicações produzidas.

Como desdobramentos futuros, pretende-se ampliar a base de dados por meio da construção de um novo conjunto de dados contendo marcas mistas e tridimensionais, permitindo investigar cenários de colidência que envolvem simultaneamente elementos figurativos, nominativos e tridimensionais. Também se prevê a integração da LLM treinada na plataforma IPDelve, com o objetivo de otimizar um sistema de apoio à análise de colidência de marcas. Além disso, pretende-se avaliar novas LLMs multimodais e investigar estratégias adicionais de ajuste fino, buscando aprimorar tanto o desempenho na classificação quanto a qualidade e a confiabilidade das explicações geradas. Essas iniciativas visam ampliar a aplicabilidade da solução proposta, consolidando seu potencial como ferramenta de apoio ao processo de análise de marcas e fomentando novas pesquisas no âmbito entre I.A e P.I.

REFERÊNCIAS

AI, U. *Unsloth: Efficient Fine-Tuning and Deployment for Large Language Models*. 2024. <<https://github.com/unslothai/unsloth>>. Accessed: 2025-10-10.

BANWAIT, H. S.; PRABHA, C. *Transformer to Generative Pretrained Transformer-5: advancements, limitations, and future directions in large language model design*. In: **INTERNATIONAL CONFERENCE ON IMAGE PROCESSING AND CAPSULE NETWORKS (ICIPCN)**, 6., 2026. *Proceedings [...]*. [S.l.]: IEEE, 2026. p. 1156-1160. Disponível em: <<https://doi.org/10.1109/ICIPCN67432.2026.11438782>>. Acesso em: 26 out. 2025.

BROWN, T. et al. *Language models are few-shot learners*. In: LAROCHELLE, H. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2020. v. 33, p. 1877-1901. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

BUI, L. V. *Advancing patent law with generative AI: human-in-the-loop systems for AI-assisted drafting, prior art search, and multimodal IP protection*. **World Patent Information**, v. 80, p. 102341, 2025. Disponível em: <https://doi.org/10.1016/j.wpi.2025.102341>. Acesso em: 2 jul. 2026.

CIPO. **Instituto de Propriedade Intelectual do Canadá (CIPO)**. 1976. Disponível em: <https://www.ic.gc.ca/eic/site/cipointernet-internetopic.nsf/eng/h_wr00003.html. acesso em : 26 out.2025>.

DETTMERS, T.; PAGNONI, A.; HOLTZMAN, A.; ZETTLEMOYER, L. QLoRA: *efficient finetuning of quantized LLMs*. In: **ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS)**, 37., 2023, New Orleans. *Proceedings [...]*. [S. l.]: Curran Associates, 2023. Disponível em: <https://arxiv.org/abs/2305.14314>. Acesso em: 2 jul. 2026.

HOFFMANN, J.; BORGEAUD, S.; MENSCH, A. et al. *Training compute-optimal large language models*. In: **INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR)**, 10., 2022. *Proceedings [...]*. Virtual: ICLR, 2022. p. 1-23. Disponível em: <https://arxiv.org/abs/2203.15556>. Acesso em: 2 jul. 2026.

HU, E. J.; SHEN, Y.; WALLIS, P.; ALLEN-ZHU, Z.; LI, Y.; WANG, S.; WANG, L.; CHEN, W. **LoRA: low-rank adaptation of large language models**. In: **INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR)**, 2022. Disponível em: <<https://arxiv.org/abs/2106.09685>>. Acesso em: 26 out. 2025.

HUANG, Dawei; XU, Ziyang; YU, Yutao; ZHENG, Cheng; YU, Hao; CAO, Mingliang; XU, Qian; LIU, Yutong; LIU, Wenbin; HU, Zhen. **From Large Language Models to Large Multimodal Models: A Literature Review**. *Applied Sciences*, v. 14, n. 12, art. 5068, 2024. DOI: 10.3390/app14125068. Disponível em: <https://www.mdpi.com/2076-3417/14/12/5068>. Acesso em: 27 jun. 2026.

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL (Brasil). Manual de marcas. Rio de Janeiro: INPI, 2023. Disponível em: <<https://manualdemarcas.inpi.gov.br/>>. Acesso em: 11 set. 2025.

KHAN, S.; NASEER, M.; HAYAT, M.; ZAMIR, S. W.; KHAN, F. S.; SHAH, M. *Transformers in vision: a survey*. **ACM Computing Surveys**, v. 54, n. 10S, art. 200, p. 1-41, 2022. Disponível em: <https://doi.org/10.1145/3505244>. Acesso em: 2 jul. 2026.

LIN, T.; WANG, Y.; LIU, X.; QIU, X. *A survey of transformers*. **AI Open**, v. 3, p. 111-132, 2022. Disponível em: <https://doi.org/10.1016/j.aiopen.2022.10.001>. Acesso em: 2 jul. 2026.

LIU, Z.; LEONG, W. *The development history, technological optimization, and multi-domain applications of large language models*. **Journal of Logistics, Informatics and Service Science**, v. 13, n. 2, p. 43-61, 2026. Disponível em: <<https://doi.org/10.33168/JLISS.2026.0203>>. Acesso em: 26 out. 2025.

MA, Huan et al. *Prompting de poucos disparos guiado por imparcialidade para modelos de linguagem de grande porte*. **Advances in Neural Information Processing Systems**, v. 36, p. 43136-43155, 2023. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2023/file/8678da90126aa58326b2fc0254b33a8c-Paper-Conference.pdf>. Acesso em: 10 out. 2025.

MACÍAS, W.; CERVIÑO, J. *Trademark dilution and its practical effect on purchase decision*. **Spanish Journal of Marketing - ESIC**, v. 21, n. 1, p. 1-13, 2017. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2444969517300057>. Acesso em: 26 out. 2025.

POWERS, D. M. W. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37-63, 2011. Disponível em: <<https://arxiv.org/abs/2010.16061>>. Acesso em: 26 out. 2025.

REIS, I. B. et al. *Conflicting marks archive dataset: a dataset of conflicting marks from the Brazilian Intellectual Property Office*. **Data**, v. 9, n. 2, 2024. Disponível em: <https://www.mdpi.com/2306-5729/9/2/33>. Acesso em: 26 out. 2025.

REZAEI, Z.; SAMGHABADI, S. S.; AMINI, M. A.; BANAD, Y. M. *A computationally efficient biomedical text processing framework for pharmacovigilance: integrating low-rank adaptation and interpretable AI for adverse drug reaction detection*. **Medical & Biological Engineering & Computing**, v. 64, p. 755-780, 2025. Disponível em: <<https://doi.org/10.1007/s11517-025-03477-w>>. Acesso em: 26 out. 2025.

SHOWKATRAMANI, G. et al. *Trademark image similarity search*. In: STEPHANIDIS, C. (ed.). **HCI International 2018 – Posters' Extended Abstracts**. Cham: Springer International Publishing, 2018. p. 199-205. (Communications in

Computer and Information Science, v. 850). Disponível em: <https://doi.org/10.1007/978-3-319-92270-6_27>. Acesso em: 26 out. 2025.

TRAPPEY, C. V.; TRAPPEY, A. J. C.; LIN, S. C.-C. *Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies*. **Advanced Engineering Informatics**, v. 45, p. 101120, 2020. Disponível em: <<https://doi.org/10.1016/j.aei.2020.101120>>. Acesso em: 26 out. 2025.

UNSLOTH AI. **Unsloth: Open source fine-tuning & RL for LLMs**. 2025. Disponível em: <https://unsloth.ai/>. Acesso em: 26 out. 2025.

USPTO. **United States Patent and Trademark Office**. 2024. Acesso: 15 de janeiro de 2024. Disponível em: <<https://www.uspto.gov/trademarks/basics>>. Acesso em: 26 out. 2025.

VASWANI, A. et al. *Attention is all you need*. In: **ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS)**, 31., 2017, Long Beach. *Proceedings [...]*. [S. l.]: Curran Associates, 2017. p. 5998-6008. Disponível em: <https://arxiv.org/abs/1706.03762>. Acesso em: 2 jul. 2026.

VESNIN, D.; LEVSHUN, D.; CHECHULIN, A. *Trademark similarity evaluation using a combination of ViT and local features*. **Information**, v. 14, n. 7, p. 398, 2023. Disponível em: <<https://doi.org/10.3390/info14070398>>. Acesso em: 26 out. 2025.

WANG, P. et al. *Qwen2-VL: enhancing vision-language model's perception of the world at any resolution*. **arXiv preprint**, arXiv:2409.12191, 2024. Disponível em: <https://arxiv.org/abs/2409.12191>. Acesso em: 2 jul. 2026.

WANG, Z.; LONG, C.; XU, S.; GAN, B.; SHI, W.; CAO, Z.; CHUA, T.-S. **LGM3A 2023: the 1st Workshop on Large Generative Models Meet Multimodal Applications**. In: *ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA (MM '23)*, 31., 2023, Ottawa, ON, Canada. *Proceedings [...]*. New York: ACM, 2023. p. 10166-10167. Disponível em: <<https://doi.org/10.1145/3581783.3610953>>. Acesso em: 26 out. 2025.

WIPO. **World Intellectual Property Indicators 2019**. [S.l.], 2019. Disponível em: <https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2019.pdf>. Acesso em: 26 out. 2025.

YIN, S.; FU, C.; ZHAO, S.; LI, K.; SUN, X.; XU, T.; CHEN, E. *A survey on multimodal large language models*. **National Science Review**, v. 11, n. 12, p. nwae403, 2024. Disponível em: <https://doi.org/10.1093/nsr/nwae403>. Acesso em: 2 jul. 2026.

ZHANG, D.; YU, Y.; DONG, J.; LI, C.; SU, D.; CHU, C.; YU, D. *MM-LLMs: recent advances in multimodal large language models*. In: **FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: ACL 2024**, 2024, Bangkok. *Proceedings [...]*. Bangkok: Association for Computational Linguistics,

2024. p. 12401-12430. Disponível em:
<https://aclanthology.org/2024.findings-acl.738/>. Acesso em: 2 jul. 2026.

ZHAO, H. et al. *Explainability for large language models: a survey*. **ACM Transactions on Intelligent Systems and Technology**, v. 15, n. 2, art. 20, p. 1-38, 2024. Disponível em: <<https://doi.org/10.1145/3639372>>. Acesso em: 26 out. 2025.