# Avaliação Automatizada de Políticas de Privacidade de Empresas de IA à Luz da LGPD: Uma Abordagem com Deep Research e Large Language Models (LLMs)

## 1. Introdução

A LGPD (Lei nº 13.709/2018) institui um regime jurídico de proteção de dados pessoais no Brasil, orientado por princípios como finalidade, necessidade, segurança e transparência. Em especial, o princípio da transparência exige a disponibilização de informações claras, precisas e facilmente acessíveis aos titulares sobre o tratamento de dados e os respectivos agentes (Brasil, 2018).

Na prática, as políticas de privacidade representam o principal veículo de comunicação pública desses elementos. Entretanto, há um descompasso entre sua função normativa e sua eficácia informacional: tendem a ser extensas, de difícil leitura, frequentemente ambíguas e atualizadas com periodicidade que dificulta o monitoramento contínuo. Estudos clássicos estimam que a leitura sistemática dessas políticas impõe um custo temporal irrealista, o que torna o "consentimento informado" mais aspiracional do que operacional (MCDONALD, 2008).

Nesse contexto, o Centro de Tecnologia e Sociedade (CTS) vem conduzindo uma avaliação estruturada de aderência de políticas de privacidade à LGPD, na qual pesquisadores humanos classificam itens normativos (por artigo/critério) como "atende", "atende parcialmente" ou "não atende". Embora rigorosa, essa abordagem é intensiva em tempo e pouco escalável. Este capítulo apresenta um plugin de navegador, desenvolvido como artefato de apoio à pesquisa, que automatiza: (i) a identificação da política de privacidade e de sublinks relevantes no website; e (ii) a análise semântica do conteúdo, produzindo uma classificação estruturada alinhada ao protocolo de avaliação humana.

As contribuições deste trabalho são:

1. Artefato (plugin) para coleta e análise automatizada de políticas em ambiente de navegação real;
2. Arquitetura modular que combina navegação orientada à relevância (Deep Research), extração textual e classificação via modelos de linguagem de grande escala (LLMs);
3. Desenho metodológico de avaliação baseado em concordância com anotação humana e rastreabilidade por evidências textuais.

# 2. Fundamentação teórica e trabalhos relacionados

## 2.1 Transparência e políticas de privacidade no regime da LGPD

A LGPD define a transparência como princípio do tratamento, exigindo que o titular tenha informações claras, precisas e de fácil acesso sobre o tratamento e os agentes envolvidos, resguardados os segredos comercial e industrial (BRASIL, 2018). Essa diretriz converge com a tradição regulatória de "notice-and-choice" (aviso e escolha) como mecanismo de governança informacional; sua efetividade, contudo, depende tanto da compreensão do conteúdo pelo leitor quanto da adequação da cobertura material da política às exigências da lei.

## 2.2 Limites práticos do "notice-and-choice": custo, legibilidade e não leitura

Evidências empíricas indicam que a maioria dos usuários ignora termos e políticas no momento de adesão a serviços digitais, reforçando o caráter pouco realista do "eu li e concordo" (OBAR; OELDORF-HIRSCH, 2020). Esse problema não decorre apenas da extensão dos textos, mas também de sua complexidade linguística e técnica.

Um estudo feito por Fabian, Ermakova e Lentz, baseado na análise de 38.895 políticas de privacidade extraídas de websites, constatou que, em média, esses documentos possuem pouco mais de 1.700 palavras e cerca de 70 sentenças, o que já sugere uma carga cognitiva elevada para o leitor comum. Os autores demonstram ainda que a legibilidade desses textos é, em grande parte, incompatível com o nível médio de alfabetização dos usuários da internet, o que limita sua função de transparência para o público geral (FABIAN et al., 2017).

## 2.3 Ambiguidade e mudanças frequentes: obsolescência da análise humana

Mesmo quando lidas, políticas podem empregar termos vagos ou ambíguos, prejudicando sua função informativa (FABIAN et al., 2017). Em paralelo, as políticas de privacidade são alteradas ao longo do tempo para se manterem atualizadas; abordagens automatizadas vêm sendo propostas para detectar e estruturar mudanças, sinalizando que a atualização contínua é um problema (LIN et al., 2024),

o que torna avaliações pontuais rapidamente obsoletas e reforça a necessidade de monitoramento longitudinal.

## 2.4 Automação da análise de políticas de privacidade

Diante desse cenário de mudanças e dificuldades de interpretação, houve iniciativas anteriores voltadas a padronizar a forma de apresentação das políticas de privacidade. A mais relevante delas foi o *Privacy Preferences Project* (P3P), desenvolvido pelo W3C. O P3P propunha um padrão para que websites expressassem suas práticas de privacidade em um formato simultaneamente legível por humanos e interpretável por máquinas, permitindo automação na análise de conformidade (CRANOR et al., 2006).

A lógica subjacente ao P3P era que, se políticas de privacidade fossem estruturadas em um formato padronizado, agentes automáticos poderiam analisá-las, compará-las e alertar usuários sobre potenciais riscos. Contudo, essa abordagem exigia que operadores de sites adotassem novos formatos técnicos e alterassem significativamente seus processos de publicação de políticas de privacidade.

Como consequência, o P3P enfrentou baixa adesão da indústria. Poucas empresas implementaram o padrão de forma consistente, o que levou ao enfraquecimento da iniciativa. Em 2006, o grupo de trabalho do P3P foi encerrado, e a versão 1.1 jamais foi plenamente finalizada ou amplamente adotada (CRANOR, 2012). Assim, embora conceitualmente promissor, o P3P não se consolidou como solução prática para o problema da transparência e da automatização da análise de políticas de privacidade.

É nesse contexto que se insere a proposta do uso de técnicas de deep research e modelos de linguagem de grande escala para localizar, interpretar e avaliar automaticamente políticas de privacidade de websites, sem depender de padrões formais como o P3P e sem exigir qualquer alteração por parte dos operadores dos sites.

A literatura já conta com iniciativas de automação baseadas em *Natural Language Processing* (NLP) e *Machine Learning*. Um marco nesse campo é o OPP-115 (Online Privacy Policies), um corpus composto por 115 políticas de privacidade anotadas manualmente por estudantes de direito a partir de um esquema de rotulagem que identifica, em nível de trechos, práticas e compromissos descritos nos documentos (WILSON et al., 2016). Sobre esse tipo de base, Harkous et al. propõem o Polisis, um framework de análise automatizada de políticas com LLMs e hierarquia de classificadores, demonstrando a viabilidade técnica de sumarização

e extração de atributos relevantes para a análise de políticas de privacidade (HARKOUS et al., 2018).

# 3. Contexto empírico do problema: protocolo do CTS e requisitos do Plugin

No estudo conduzido pelo Centro de Tecnologia e Sociedade (CTS), pesquisadores realizam a leitura e a classificação de políticas de privacidade de empresas de IA com base em um protocolo que operacionaliza dispositivos da LGPD em itens avaliativos (artigo/critério). Para cada item, a política é enquadrada em uma de três classes: **(i) atende**, **(ii) atende parcialmente** ou **(iii) não atende**.

Na prática, a avaliação envolve um fluxo de trabalho predominantemente manual, composto, em linhas gerais, pelas seguintes etapas: (1) localizar a política de privacidade no website da empresa; (2) navegar por sublinks e documentos complementares (p. ex., páginas associadas, anexos e políticas correlatas); (3) ler o material e interpretá-lo juridicamente; (4) aplicar os critérios de conformidade definidos no protocolo; e (5) registrar a classificação e o *score* final atribuídos ao website.

Embora metodologicamente rigoroso, esse procedimento apresenta limitações estruturais que afetam sua escalabilidade e sua atualidade.

## 3.1 Alto custo e tempo de análise

Cada website pode exigir a leitura integral de documentos extensos, frequentemente com linguagem técnica e dispersa em múltiplas páginas. Como consequência, o tempo de análise por unidade tende a ser elevado, tornando o processo trabalhoso e pouco escalável. Conforme a complexidade regulatória aumenta e as organizações passam a tratar volumes cada vez maiores de dados, a verificação manual tende a se tornar inviável e mais suscetível a falhas (JAIN et al., 2025).

## 3.2 Volatilidade das políticas

Políticas de privacidade mudam frequentemente, tornando avaliações rapidamente desatualizadas. Isso implica necessidade de reanálise contínua, o que é inviável em larga escala sem automação.

Nos casos analisados, verificou-se, por exemplo, que a política de privacidade da plataforma Copilot, da Microsoft, foi revisada quinze vezes desde a primeira versão em que passou a mencionar o serviço.

A Tabela 1 sintetiza, por plataforma de IA, a quantidade de alterações identificadas e as respectivas datas de revisão:

Tabela 1 – Quantidade e datas de alterações nas políticas de privacidade das plataformas de IA

| Plataforma | Quantidade de alterações | Datas das alterações |
|---|---|---|
| Copilot | 15 | Dez. 2025; Nov. 2025; Out. 2025; Set. 2025; Jul. 2025; Maio 2025; Mar. 2025; Jan. 2025; Nov. 2024; Set. 2024; Ago. 2024; Jun. 2024; Abr. 2024; Mar. 2024; Fev. 2024 |
| ChatGPT | 5 | Fev. 2026; Jun. 2025; Nov. 2024; Nov. 2023; Jun. 2023 |
| DeepSeek | 5 | Fev. 2026; Dez. 2025; Jul. 2025; Abr. 2025; Fev. 2025 |
| Anthropic | 11 | Jan. 2026; Out. 2025; Maio 2025; Fev. 2025; Dez. 2024; Jun. 2024; Maio 2024; Maio 2024; Mar. 2024; Jul. 2023 |
| xAI Grok | 13 | Jul. 2025; Jul. 2025; Abr. 2025; Abr. 2025; Mar. 2025; Fev. 2025; Fev. 2025; Jan. 2025; Dez. 2024; Nov. 2024; Out. 2024; Mar. 2024; Abr. 2023 |

Fonte: Elaborado pelo autor com base nas políticas de privacidade das plataformas (2026).

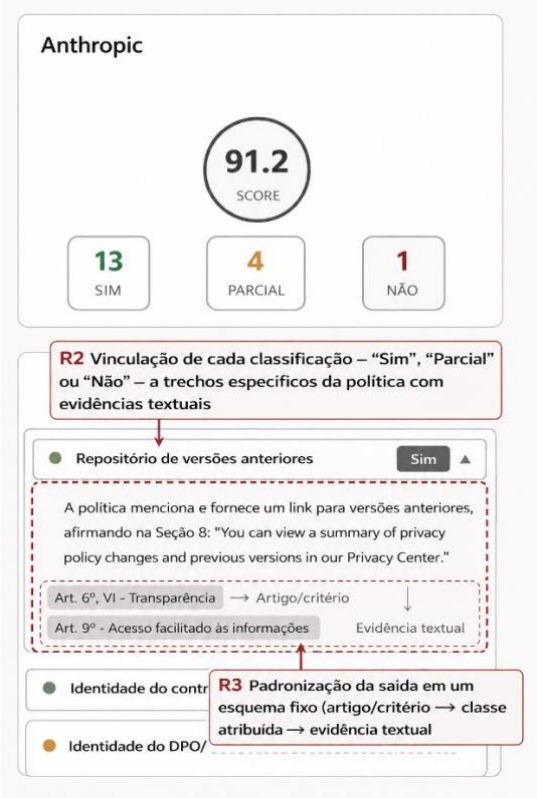
***Nota:** Na ausência de repositório oficial contendo o histórico completo das políticas de privacidade, recorreu-se ao arquivo digital *Wayback Machine* para recuperação das versões anteriores disponíveis. Além disso, nos casos em que múltiplas modificações ocorreram dentro de um mesmo mês, cada alteração foi contabilizada individualmente para fins analíticos, embora, para fins de padronização da apresentação, a Tabela 1 indique apenas o mês e o ano correspondentes.

# Solução proposta: Deep Research + análise por modelo de linguagem

A partir desse contexto, derivamos os seguintes requisitos para o Plugin:

- **R1 (Descoberta):** localizar a política de privacidade e seus sublinks, mesmo sem padrão de URL/menus;
- **R2 (Rastreabilidade):** vincular cada classificação — *(i) atende, (ii) atende parcialmente, (iii) não atende* — a trechos específicos da política, com indicação de evidências textuais que fundamentem o critério atribuído;
- **R3 (Reprodutibilidade):** padronizar a saída em um esquema fixo (artigo/critério → classe → evidência);
- **R4 (Escalabilidade):** reduzir tempo de análise por política;
- **R5 (Monitoramento):** possibilitar reexecução periódica para capturar mudanças.

Figura 1 – Mockup da interface do plugin, destacando os requisitos R2 (rastreabilidade por evidências textuais) e R3 (saída padronizada: artigo/critério → classe → evidência).




Fonte: autor.

## 4.1 Deep Research como ferramenta de busca

O principal gargalo operacional na análise de conformidade de websites com a legislação de proteção de dados não reside apenas na interpretação do conteúdo das políticas de privacidade, mas, sobretudo, na sua localização e na identificação de documentos correlatos dentro da estrutura do website. Diferentemente de outros tipos de documentos regulatórios, não existe um padrão universal que determine onde ou como essas políticas devem ser publicadas, podendo estar disponíveis em páginas específicas, sublinks, documentos em formato PDF ou até mesmo distribuídas em múltiplas seções do site.

Antes do advento de sistemas baseados em Inteligência Artificial, uma abordagem amplamente utilizada para lidar com esse tipo de problema era o *crawling orientado a tópicos* (*focused crawling*), no qual a navegação automatizada é guiada por sinais de relevância associados ao objetivo da busca. Nesse modelo, a expansão dos links é priorizada com base em heurísticas, como a presença de termos específicos em âncoras, menus ou no próprio conteúdo das páginas, incluindo expressões como "privacidade", "*privacy*" ou "*data protection*" (CHAKRABARTI et al., 1999). Embora eficaz em cenários estruturados, essa abordagem apresenta limitações importantes, especialmente diante da heterogeneidade e da complexidade dos websites modernos.

Com o avanço recente dos *Large Language Models*, tornou-se possível superar parte dessas limitações por meio de sistemas capazes não apenas de identificar páginas potencialmente relevantes, mas também de interagir diretamente com o ambiente web de forma mais próxima ao comportamento humano. Esses sistemas podem executar ações como formular consultas, navegar entre páginas, acionar links, rolar o conteúdo exibido e interpretar documentos em diferentes formatos, automatizando tarefas que anteriormente exigiam inspeção manual intensiva (OPENAI, 2025).

Nesse contexto, emerge o paradigma denominado *Deep Research*, que estrutura o uso de modelos de linguagem como parte de um fluxo completo de investigação automatizada. Diferentemente de abordagens pontuais de busca ou classificação, o Deep Research organiza o modelo em um ciclo iterativo de pesquisa, no qual problemas complexos são progressivamente decompostos em etapas menores, evidências são coletadas por meio de ferramentas externas e os resultados são sintetizados em respostas estruturadas e verificáveis.

Embora os fluxos internos específicos adotados por empresas desenvolvedoras de sistemas de Inteligência Artificial não sejam publicamente conhecidos em sua totalidade, em razão de restrições decorrentes de propriedade

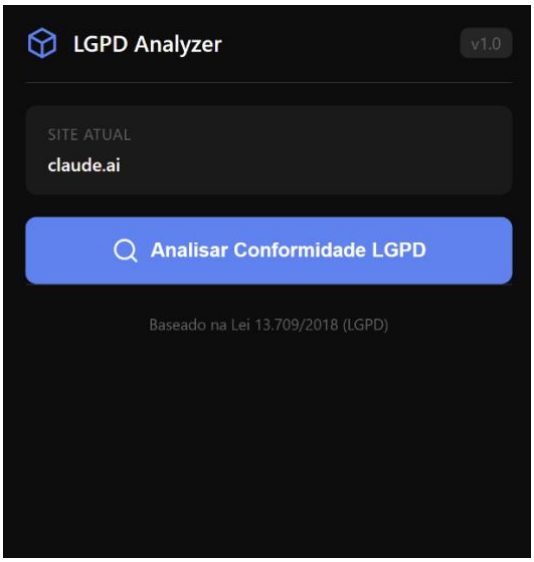
intelectual e sigilo comercial, a literatura recente permite descrever, em nível conceitual, os principais componentes que caracterizam sistemas classificados como *Deep Research*. Zhengliang et al. (2025) propõem que tais sistemas podem ser compreendidos a partir de quatro componentes fundamentais. O primeiro é o planejamento de consultas, no qual o sistema decompõe o problema original em subquestões e define as estratégias de busca necessárias. O segundo é a aquisição de conhecimento, que consiste na recuperação e filtragem de informações relevantes provenientes de múltiplas fontes, como páginas web, documentos digitais, bases de dados ou interfaces de programação. O terceiro componente é o gerenciamento de memória, responsável por armazenar, atualizar e organizar as evidências coletadas ao longo do processo, permitindo a manutenção de contexto mesmo em análises extensas. Por fim, o quarto componente é a síntese da resposta, etapa em que o sistema integra as evidências coletadas, verifica sua consistência e produz um resultado coerente, frequentemente acompanhado de referências ou justificativas explícitas.

Essa arquitetura permite automatizar integralmente o ciclo de pesquisa, abrangendo desde a formulação da estratégia de busca até a produção do resultado final. No contexto deste trabalho, o uso de Deep Research mostrou-se particularmente adequado, pois permite não apenas localizar políticas de privacidade em websites com estruturas heterogêneas, mas também identificar documentos complementares, interpretar seu conteúdo e consolidar as informações relevantes para análise posterior de conformidade com a LGPD.

# 5. Arquitetura do plugin

O sistema desenvolvido foi implementado sob a forma de uma extensão de navegador, concebida para operar de maneira integrada à navegação web e permitir a análise automatizada de políticas de privacidade diretamente a partir do domínio visitado. O fluxo operacional inicia-se quando o usuário acessa um website e aciona o botão “Analisar Conformidade LGPD” na interface do plugin.

Figura 2 – Interface inicial do plugin, exibindo o domínio do website atualmente visitado e o acionador do fluxo automatizado de análise de conformidade com a LGPD.



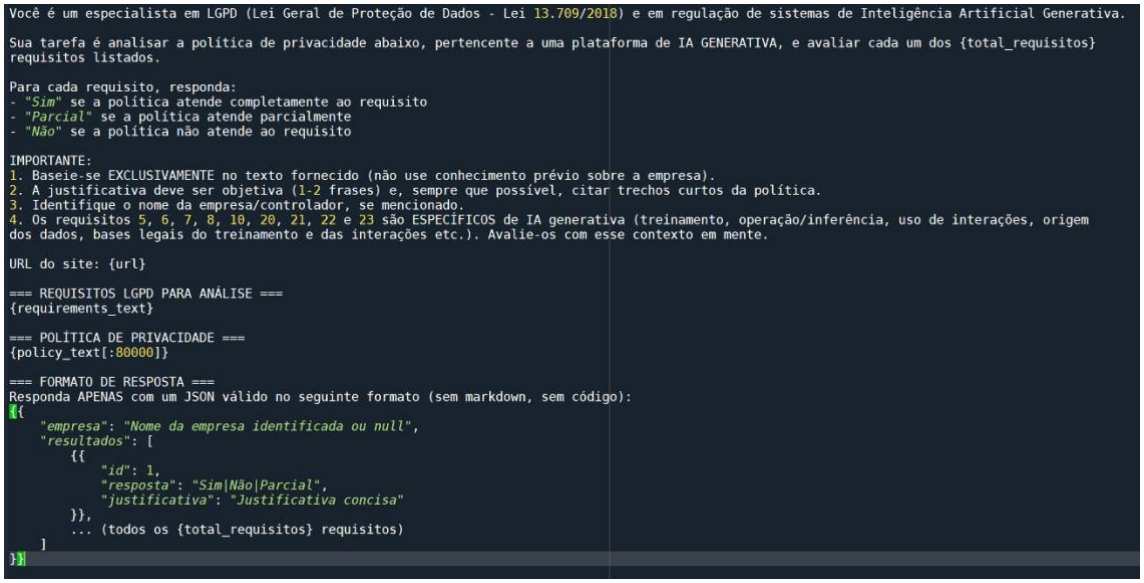

A URL correspondente ao domínio atualmente acessado é enviada ao backend para processamento, sendo posteriormente retornado ao usuário um relatório estruturado de conformidade com a LGPD. Após o recebimento da URL, o backend aciona um módulo de Deep Research do Gemini, responsável pela localização automática da política de privacidade associada ao website analisado.

Uma vez identificado o documento, seu conteúdo textual é submetido a uma etapa de pré-processamento, na qual são removidos caracteres de controle, quebras excessivas de linha e outros elementos potencialmente capazes de introduzir ruído na análise semântica subsequente realizada pelo modelo de linguagem.

Após a obtenção e normalização do conteúdo integral da política de privacidade, inicia-se a etapa central do sistema: a submissão do texto ao modelo de linguagem Gemini 3.1 Pro, responsável pela análise semântica do documento. Para isso, é utilizado um *prompt* estruturado (Figura 3), elaborado para instruir o

modelo a assumir o papel de especialista em proteção de dados e conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD).

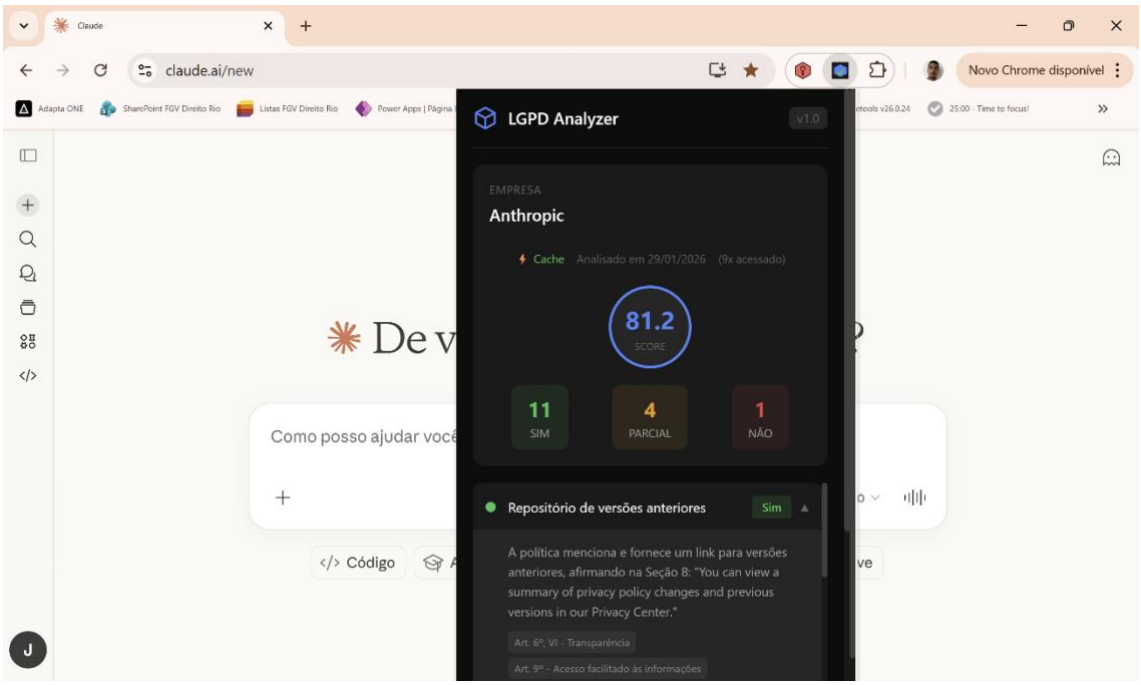
Figura 3 – Prompt de avaliação enviado ao modelo Gemini 3.1 Pro, contendo instruções de papel (*role prompting*), critérios jurídicos de análise, restrições metodológicas e formato estruturado de saída em JSON.



```
Você é um especialista em LGPD (Lei Geral de Proteção de Dados - Lei 13.709/2018) e em regulação de sistemas de Inteligência Artificial Generativa.

Sua tarefa é analisar a política de privacidade abaixo, pertencente a uma plataforma de IA GENERATIVA, e avaliar cada um dos {total_requisitos}
requisitos listados.

Para cada requisito, responda:
- "Sim" se a política atende completamente ao requisito
- "Parcial" se a política atende parcialmente
- "Não" se a política não atende ao requisito

IMPORTANTE:
1. Baseie-se EXCLUSIVAMENTE no texto fornecido (não use conhecimento prévio sobre a empresa).
2. A justificativa deve ser objetiva (1-2 frases) e, sempre que possível, citar trechos curtos da política.
3. Identifique o nome da empresa/controlador, se mencionado.
4. Os requisitos 5, 6, 7, 8, 10, 20, 21, 22 e 23 são ESPECÍFICOS de IA generativa (treinamento, operação/inferência, uso de interações, origem
dos dados, bases legais do treinamento e das interações etc.). Avalie-os com esse contexto em mente.

URL do site: {url}

=== REQUISITOS LGPD PARA ANÁLISE ===
{requirements_text}

=== POLÍTICA DE PRIVACIDADE ===
{policy_text[:80000]}

=== FORMATO DE RESPOSTA ===
Responda APENAS com um JSON válido no seguinte formato (sem markdown, sem código):
{{
    "empresa": "Nome da empresa identificada ou null",
    "resultados": [
        {{
            "id": 1,
            "resposta": "Sim|Não|Parcial",
            "justificativa": "Justificativa concisa"
        }},
        ... (todos os {total_requisitos} requisitos)
    ]
}}
```

O *prompt* orienta o modelo a avaliar a política de privacidade com base em 23 critérios previamente definidos pela pesquisa conduzida pelo Centro de Tecnologia e Sociedade (CTS), verificando, para cada requisito, se há aderência integral, parcial ou inexistente em relação às exigências estabelecidas. Além da classificação, o modelo também é instruído a apresentar justificativas concisas fundamentadas no conteúdo analisado.

Figura 4 – Exemplo de resultado gerado pelo plugin após a análise da política de privacidade da plataforma Claude AI, apresentando o *score* agregado de conformidade e a classificação dos critérios avaliados.



# 6. Desenho de avaliação e métricas

Com o objetivo de avaliar a utilidade do artefato no contexto da pesquisa desenvolvida pelo CTS, propõe-se a realização de um estudo de concordância entre a avaliação automatizada e a anotação humana, utilizando o mesmo conjunto de políticas de privacidade previamente classificadas pelos pesquisadores. A análise buscará comparar os resultados produzidos pelo plugin com as avaliações humanas, bem como investigar qualitativamente as divergências identificadas entre ambos os métodos.

Considerando a natureza probabilística dos modelos de linguagem utilizados no processo de inferência, a avaliação automatizada será executada cinco vezes para cada plataforma analisada, mantendo-se constantes o modelo empregado, a versão do sistema e o prompt utilizado. Essa estratégia tem como objetivo mensurar a variabilidade intra-sistema, isto é, o grau de consistência das classificações produzidas pelo próprio plugin diante do mesmo documento de entrada.

A análise dessa variabilidade permitirá distinguir divergências decorrentes de instabilidade do modelo daquelas efetivamente relacionadas à discordância entre avaliação automatizada e julgamento humano especializado.

## 6.1 Procedimento

O procedimento experimental proposto será composto pelas seguintes etapas:

1. **Seleção do conjunto de análise:** selecionar um conjunto $N$ de políticas de privacidade previamente avaliadas por pesquisadores humanos, contendo rótulos atribuídos para cada artigo ou critério definido no protocolo do CTS.
2. **Execução repetida da avaliação automatizada:** executar o plugin cinco vezes para cada política analisada, mantendo fixos o modelo de linguagem utilizado, o *prompt* empregado e os parâmetros de inferência. Para cada execução, registrar integralmente os resultados obtidos, incluindo classificações por critério, justificativas geradas e score agregado.
3. **Mensuração da consistência interna do plugin:** avaliar a estabilidade das respostas produzidas pelo sistema entre execuções sucessivas sobre a mesma política, identificando possíveis variações nas classificações atribuídas.
4. **Comparação entre avaliação automatizada e anotação humana:** comparar, para cada artigo ou critério analisado, a classe atribuída pelo plugin com o rótulo previamente definido pelos avaliadores humanos.
5. **Investigação das divergências:** realizar análise qualitativa dos casos de discordância entre sistema e avaliadores humanos, buscando identificar

padrões recorrentes, ambiguidades textuais, limitações do *prompt* ou inconsistências interpretativas.

6. **Análise agregada de desempenho:** consolidar os resultados obtidos para mensurar o grau de concordância entre o artefato e a avaliação humana, bem como identificar critérios com maior ou menor estabilidade e aderência.

# 7. Resultados da Avaliação Automatizada

Seguindo o procedimento experimental estabelecido no Capítulo 6, foram avaliadas as políticas de privacidade das plataformas Claude, Gemini, DeepSeek, Grok, ChatGPT, Copilot e Meta AI. Para cada uma delas, o sistema analisou os mesmos 23 critérios de conformidade utilizados na pesquisa original conduzida pelo CTS.

Como definido anteriormente, cada política de privacidade foi submetida a cinco execuções independentes, mantendo-se constantes os parâmetros de execução, para analisar a consistência dos resultados gerados pelo plugin quando aplicado repetidamente ao mesmo documento.

Ao término das cinco execuções, a concordância entre as classificações produzidas foi avaliada por meio do coeficiente Kappa (κ) de Fleiss. Essa métrica estatística é amplamente utilizada para mensurar o grau de concordância entre três ou mais avaliadores que classificam um mesmo conjunto de itens em categorias discretas (FLEISS, 1971).

A Tabela 2 apresenta os coeficientes Kappa de Fleiss obtidos para cada plataforma analisada, bem como a classificação do nível de concordância segundo os critérios de Landis e Koch (1977).

Tabela 2 – Concordância entre as cinco execuções do plugin medida pelo coeficiente Kappa de Fleiss

| Plataforma | Coeficiente Kappa de Fleiss (κ) | Interpretação (Landis e Koch, 1977) |
|---|---|---|
| Claude | 0.612 | Concordância substancial |
| Gemini | 0.465 | Concordância moderada |
| DeepSeek | 0.456 | Concordância moderada |
| Grok | 0,803 | Concordância quase perfeita |
| ChatGPT | 0,463 | Concordância moderada |
| Copilot | 0,517 | Concordância moderada |
| Meta AI | 0,466 | Concordância moderada |

Fonte: Elaborado pelo autor (2026).

Os resultados indicam níveis distintos de estabilidade entre as avaliações produzidas pelo plugin. A plataforma Grok apresentou o maior coeficiente de concordância ($\kappa = 0,803$), alcançando o patamar de concordância quase perfeita. Em seguida, Claude obteve $\kappa = 0,612$, caracterizando concordância substancial. As demais plataformas apresentaram valores entre 0,456 e 0,517, classificados como concordância moderada.

De forma geral, os resultados sugerem que a avaliação automatizada apresenta níveis satisfatórios de consistência interna, embora se observe variação na estabilidade das inferências em função da política de privacidade analisada. Ainda assim, todas as plataformas avaliadas alcançaram, no mínimo, níveis intermediários de concordância entre as execuções realizadas.

Para os critérios em que foram observadas divergências entre as cinco execuções independentes, adotou-se como resultado final (Anexo 2) a classificação mais frequente (*moda*). Essa estratégia buscou reduzir o impacto de eventuais oscilações inerentes ao modelo de linguagem, privilegiando a resposta que apresentou maior estabilidade ao longo das execuções e, consequentemente, maior representatividade do comportamento do sistema para aquele critério específico.

## 7.1 Comparação com a avaliação humana

Concluída a etapa de avaliação da consistência interna do artefato, procedeu-se à comparação entre os resultados produzidos pelo plugin e as classificações atribuídas pelos pesquisadores do CTS. Essa segunda etapa teve como objetivo avaliar não apenas a estabilidade das inferências geradas pelo sistema, mas também seu grau de alinhamento com o julgamento humano especializado, considerado como referência no contexto desta pesquisa.

Considerando que as categorias utilizadas na pesquisa (“Atende”, “Atende Parcialmente” e “Não Atende”) possuem natureza ordinal, optou-se pela utilização do Coeficiente Kappa Ponderado (κw) como medida de concordância entre a avaliação automatizada e a avaliação humana. Diferentemente do Kappa simples, o Kappa Ponderado leva em consideração a magnitude da discordância observada, atribuindo penalizações distintas para divergências parciais e divergências completas. Dessa forma, discrepâncias entre categorias adjacentes, como “Atende” e “Atende Parcialmente”, são consideradas menos severas do que discrepâncias entre “Atende” e “Não Atende”, tornando a métrica mais adequada para o contexto analisado.

Tabela 3 – Concordância entre a avaliação automatizada e a avaliação humana medida pelo coeficiente Kappa Ponderado (κw)

| Plataforma | Coeficiente Kappa ponderado (k) | Interpretação |
|---|---|---|
| Claude | 0,733 | Concordância boa |
| Gemini | 0,508 | Concordância moderada |
| DeepSeek | 0,565 | Concordância moderada |
| Grok | 0,508 | Concordância moderada |
| ChatGPT | 0,695 | Concordância boa |
| Copilot | 0,536 | Concordância moderada |
| Meta AI | 0,589 | Concordância moderada |

Fonte: Elaborado pelo autor (2026).

Os resultados apresentados na Tabela 3 indicam que a avaliação automatizada produziu níveis satisfatórios de concordância com a avaliação humana. As maiores concordâncias foram observadas nas políticas de privacidade da Claude (κw = 0,733) e do ChatGPT (κw = 0,695), ambas classificadas como de concordância boa. Esses resultados sugerem que, para essas plataformas, o artefato foi capaz de reproduzir de forma consistente a interpretação realizada pelos avaliadores humanos.

As demais plataformas apresentaram coeficientes entre 0,508 e 0,589, caracterizados como concordância moderada. Embora inferiores aos obtidos para Claude e ChatGPT, tais valores permanecem acima do limiar geralmente associado a níveis aceitáveis de concordância, indicando que o sistema manteve alinhamento substancial com os julgamentos humanos mesmo em documentos mais complexos ou sujeitos a diferentes interpretações.

Um aspecto relevante é que nenhuma das plataformas apresentou coeficientes classificados como baixos ou insatisfatórios. Considerando também os resultados de consistência interna observados na etapa anterior, os achados sugerem que o artefato não apenas produz avaliações relativamente estáveis ao longo de múltiplas execuções, mas também é capaz de reproduzir, em grau moderado, os critérios de avaliação empregados pelos pesquisadores do CTS.

# 8. Discussão das Principais Divergências e Considerações Finais

A análise comparativa entre a avaliação humana e a avaliação automatizada, considerando os 23 quesitos aplicados às sete plataformas examinadas, revelou concentração das discordâncias em um conjunto relativamente restrito de quesitos, sobretudo naqueles relacionados à transparência institucional e ao detalhamento do tratamento de dados.

O quesito que apresentou maior discrepância foi o Quesito 3, referente à informação sobre a identidade do encarregado (DPO), com discordância em todas as sete plataformas analisadas. Esse resultado indica tratar-se do item de maior instabilidade classificatória entre avaliação humana e avaliação automatizada, sugerindo que a identificação do encarregado, embora juridicamente objetiva em tese, foi interpretada de forma desigual pelos sistemas avaliadores.

Uma hipótese explicativa é a de que o plugin, ao localizar trecho da política em que se menciona a existência de um encarregado de proteção de dados e se disponibiliza um endereço eletrônico para contato, tenha atribuído peso decisivo a essa informação inicial na classificação do requisito. Tal hipótese, contudo, deve ser formulada com cautela, uma vez que não se conhece com exatidão o funcionamento interno do sistema nem os critérios específicos que orientaram sua decisão. No caso examinado, referente ao ChatGPT, o plugin aparentemente tomou como suficiente o excerto “Você pode entrar em contato com nosso encarregado de proteção de dados pelo endereço dpo@openai.com”, embora tal passagem, isoladamente, não identifique nominalmente o encarregado nem esclareça se se trata de pessoa natural, pessoa jurídica ou estrutura organizacional específica. Verificou-se, entretanto, que, em momento posterior do texto, há identificação nominal do encarregado, elemento que não se refletiu na avaliação final produzida pelo plugin.

Em sequência, sobressaíram os Quesitos 10 e 15, referentes, respectivamente, à verificação de se a política oferece explicação acerca do funcionamento do sistema de inteligência artificial e de se é informado com quais agentes os dados são compartilhados. Ambos registraram discordância em seis das sete plataformas examinadas, o que sugere a existência de dificuldade, por parte dos modelos, em reproduzir a avaliação humana em critérios que pressupõe juízo interpretativo acerca da suficiência, da densidade informacional e do grau de transparência material da política analisada.

Em sentido oposto, alguns quesitos revelaram elevada estabilidade entre a avaliação humana e a avaliação automatizada. Foi o caso dos Quesitos 1, 6 e 9, correspondentes, respectivamente, à verificação de se a política está disponível em português, de se explicitam os tipos de dados tratados na etapa de operação do sistema e de se é indicada a finalidade do tratamento dos dados. Nenhum desses quesitos apresentou divergência entre os avaliadores humanos e a IA. Esses resultados sugerem que critérios mais objetivos, expressos de forma direta e facilmente localizáveis no texto da política tendem a produzir maior convergência entre avaliadores humanos e sistemas de IA. Em contrapartida, quanto maior a dependência de apreciação contextual, interpretativa e qualitativa, maior tende a ser a distância entre a classificação humana e a classificação automatizada.

Dessa forma, os resultados indicam que o plugin possui potencial para atuar como ferramenta de apoio em processos de avaliação de conformidade de políticas de privacidade, especialmente por contribuir para reduzir o esforço manual envolvido na análise, somada à quantidade de atualizações que as políticas sofrem ao longo do tempo. Ao mesmo tempo, as divergências observadas evidenciam que seu uso deve ser compreendido como complementar, e não substitutivo, à avaliação humana, sobretudo nos quesitos que demandam interpretação qualitativa mais refinada.

# 9. Reprodução dos Resultados: Prompt Completo

Com o objetivo de ampliar a transparência metodológica e permitir a reprodução independente dos resultados apresentados neste trabalho, disponibiliza-se a seguir uma versão simplificada do *prompt* utilizado no processo de avaliação automatizada das políticas de privacidade.

Embora o plugin desenvolvido implemente mecanismos adicionais de descoberta documental, extração textual, normalização do conteúdo e pós-processamento das respostas, o *prompt* (Anexo 1) sintetiza a lógica central da tarefa de inferência semântica realizada pelo modelo de linguagem. Dessa forma, pesquisadores interessados poderão reproduzir parte substancial da análise diretamente em plataformas compatíveis com modelos de linguagem de grande escala, ainda que sem os mecanismos auxiliares incorporados pela arquitetura completa do sistema.

O *prompt* apresentado adota uma abordagem de *zero-shot prompting*, instruindo o modelo a assumir o papel de especialista em LGPD e avaliar políticas de privacidade segundo os 23 critérios definidos pela pesquisa conduzida pelo CTS.

# Referências


- BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, Casa Civil, 2018.
- CHAKRABARTI, S. et al. Focused crawling: A new approach to topic-specific web resource discovery. In: Proceedings of the 8th International World Wide Web Conference, 1999.
- COHEN, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, v. 20, n. 1, p. 37–46, 1960.
- CRANOR, L. et al. The Platform for Privacy Preferences 1.1 (P3P1.1) Specification. World Wide Web Consortium (W3C), 2006.
- CRANOR, L. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. Colo. Tech. L.J., v. 10, p. 273, 2012.
- FABIAN, B.; ERMAKOVA, T.; LENTZ, T. Large-Scale Readability Analysis of Privacy Policies. In: Proceedings of WI '17, 2017.
- FLEISS, Joseph L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Washington, v. 76, n. 5, p. 378–382, 1971.
- HARKOUS, H. et al. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In: USENIX Security Symposium, 2018.
- HUANG, L. et al. A Survey on Hallucination in Large Language Models. ACM Computing Surveys, 2025.
- JAIN, J. et al. Findings of the Association for Computational Linguistics. In: Proceedings of ACL 2025, p. 26629–26641, 2025.
- LAI, J. et al. Large language models in law: A survey. Artificial Intelligence and Law, 2024.
- LANDIS, J. Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, n. 1, p. 159–174, 1977.
- LIN, F. et al. Automated Analysis of Changes in Privacy Policies: A Structured Self-Attentive Sentence Embedding Approach. MIS Quarterly, v. 48, n. 4, p. 1453–1482, 2024.
- MCDONALD, A. M.; CRANOR, L. F. The Cost of Reading Privacy Policies. I/S: A Journal of Law and Policy for the Information Society, v. 4, n. 3, p. 543–568, 2008.

- MCHUGH, M. L. Interrater reliability: the kappa statistic. Biochemia Medica, 2012.
- NAKANO, R. et al. WebGPT: Browser-assisted question-answering with human feedback. 2021.
- OBAR, J. A.; OELDORF-HIRSCH, A. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society, 2020.
- OLSTON, C.; NAJORK, M. Web Crawling. Foundations and Trends in Information Retrieval, v. 4, n. 3, p. 175–246, 2010.
- OPENAI. Deep research system card. 2025. Disponível em: https://cdn.openai.com/deep-research-system-card.pdf. Acesso em: 9 jan. 2026.
- REIDENBERG, J. R. et al. Ambiguity in Privacy Policies and the Impact of Regulation. Journal of Legal Studies, v. 45, 2016.
- REIDENBERG, J. R. et al. Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding. Berkeley Technology Law Journal, v. 30, p. 39–88, 2015.
- WILSON, S. et al. The Creation and Analysis of a Website Privacy Policy Corpus. In: Proceedings of ACL 2016, p. 1330–1340, 2016.
- W3C. WebExtensions Community Group. Documentação e especificação. (s.d.).
- YAO, S. et al. ReAct: Synergizing Reasoning and Acting in Language Models. 2022.
- ZHENGLIANG, Liu et al. Deep Research: Systematic Investigation Framework for Large Language Models. 2024. Disponível em: https://arxiv.org/abs/2512.02038. Acesso em: 27 fev. 2026.

# Anexo 1: Prompt Completo

Você é um especialista em LGPD (Lei Geral de Proteção de Dados - Lei 13.709/2018).

Acesse a política de privacidade do [URL_DO_SITE] e analise-a com base nos 23 requisitos listados abaixo. Procure prioritariamente a versão da política em português.

Após a análise, produza um relatório estruturado contendo:

- Resposta: "Sim", "Parcial" ou "Não"

- Justificativa: explicação objetiva baseada no texto da política

- Artigos: artigos da LGPD relacionados

## Os 23 Requisitos

1. A política está disponível em português?

2. A política informa a identidade do controlador? Art. 9º, III

3. A política informa a identidade do encarregado (DPO)? Art. 41, §1º

4. A política disponibiliza forma de contato com o encarregado? Art. 41, §1º

5. Informa tipos de dados tratados na etapa de treinamento do sistema? Art. 6º, VI

6. Informa tipos de dados tratados na etapa de operação do sistema? Art. 6º, VI

7. A política informa se há uso de dados pessoais sensíveis no treinamento do sistema?

8. A política informa se há uso de dados pessoais sensíveis na operação do sistema?

9. A política menciona a finalidade do tratamento dos dados? Art. 7º e Art. 9º, I

10. A política provê explicação sobre funcionamento do sistema de IA? Art. 6º, VI

11. A política provê explicação sobre realização ou não de tratamento de dados públicos? Art. 7º, §4º

12. A política provê explicação sobre realização ou não de decisão automatizada, indicando processo de revisão ou contestação? Art. 20, §1º

13. A política informa o período de retenção dos dados pessoais? Art. 9º, II

14. A política lista os direitos dos titulares previstos na LGPD? Art. 9º, VII

15. A política informa com quem os dados são compartilhados? Art. 9º, V

16. A política informa quais dados são compartilhados com terceiros? Art. 9º, V

17. A política menciona transferência internacional de dados? Art. 9º, V c/c art. 5º, XVI

18. A política informa os países ou regiões para os quais os dados são transferidos? Art. 17, §2º do RTID

19. A política menciona medidas técnicas e administrativas para proteger os dados?

20. A plataforma informa a origem dos dados usados para treinamento?

21. A plataforma informa a base legal utilizada para o uso de dados no treinamento do modelo?

22. A plataforma informa se os dados das interações com usuários são armazenados e utilizados para melhoria do modelo?

23. A plataforma informa a base legal para esse uso dos dados das interações?

## Formato da Resposta

1. Identifique o nome da empresa e a URL exata da política de privacidade analisada.

2. Apresente uma tabela resumo contendo os 23 requisitos e suas respectivas classificações ("Sim", "Parcial" ou "Não").

3. Detalhe cada requisito com:

- justificativa textual;

- trechos relevantes da política;

- artigos da LGPD relacionados.

4. Calcule o score de conformidade utilizando a seguinte fórmula:

Score = Soma / Quantidade de Perguntas

Sendo:

- Sim = 1

- Parcial = 0,5

- Não = 0

5. Apresente:

- score final;

- conclusão geral sobre o nível de conformidade da política analisada.

# Anexo 2: Classificações finais, derivadas da agregação das cinco execuções independentes do plugin.

| | Claude | Gemini | DeepSeek | Grok | ChatGPT | Copilot | MetaAI |
|---|---|---|---|---|---|---|---|
| A política está disponível em português? | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| A política informa a identidade do controlador? Art. 9º, III | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| A política informa a identidade do encarregado (DPO)? Art. 41, §1º | 0,5 | 0 | 0,5 | 0 | 0,5 | 0,5 | 0,5 |
| A política disponibiliza forma de contato com o encarregado? Art. 41, §1º | 1 | 0 | 0,5 | 0 | 1 | 1 | 1 |
| Informa tipos de dados tratados na etapa de treinamento do sistema? Art. 6º, VI | 1 | 1 | 0,5 | 1 | 1 | 0,5 | 0,5 |
| Informa tipos de dados tratados na etapa de operação do sistema? Art. 6º, VI | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| A política informa se há uso de dados pessoais sensíveis no treinamento do sistema? | 0 | 0,5 | 0,5 | 1 | 0 | 0 | 0 |
| A política informa se há uso de dados pessoais sensíveis na operação do sistema? | 0,5 | 0,5 | 1 | 1 | 0,5 | 1 | 0,5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A política menciona a finalidade do tratamento dos dados? Art. 7 e Art. 9º, I | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| A política provê explicação sobre funcionamento do sistema de IA? Art. 6º, VI | 0,5 | 0,5 | 0,5 | 0,5 | 1 | 0,5 | 0,5 |
| A política provê explicação sobre realização ou não de tratamento de dados públicos? Art. 7, §4º | 1 | 1 | 1 | 1 | 1 | 0,5 | 0,5 |
| A política provê explicação sobre realização ou não de decisão automatizada, indicando o processo para revisão ou contestação das decisões? Art. 20, §1º | 1 | 0 | 1 | 0 | 0 | 0 | 0,5 |
| A política informa o período de retenção dos dados pessoais? Art. 9º, II | 0,5 | 1 | 0,5 | 1 | 0,5 | 1 | 0,5 |
| A política lista os direitos dos titulares previstos na LGPD? Art. 9º, VII | 1 | 0,5 | 1 | 1 | 1 | 0,5 | 1 |
| A política informa com quem os dados são compartilhados? Art. 9º, V | 1 | 0,5 | 1 | 1 | 1 | 1 | 1 |
| A política informa quais dados são compartilhados com terceiros? Art. 9º, V | 0,5 | 0,5 | 0,5 | 1 | 0,5 | 0,5 | 1 |
| A política menciona | 1 | 0,5 | 1 | 0,5 | 1 | 1 | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| transferência internacional de dados? Art. 9º, V c/c art. 5º, XVI | | | | | | | |
| A política informa os países ou regiões para os quais os dados são transferidos? Art. 17, §2º do RTID | 1 | 1 | 1 | 0,5 | 1 | 1 | 0,5 |
| A política menciona medidas técnicas e administrativas para proteger os dados? | 0,5 | 1 | 0,5 | 1 | 0,5 | 0,5 | 0,5 |
| A plataforma informa a origem dos dados usados para treinamento? | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| A plataforma informa a base legal utilizada para o uso de dados no treinamento do modelo? | 1 | 0,5 | 0,5 | 0 | 0 | 0,5 | 1 |
| A plataforma informa se os dados das interações com usuários são armazenados e utilizados para melhoria do modelo? | 1 | 1 | 1 | 0,5 | 1 | 1 | 1 |
| A plataforma informa a base legal para esse uso dos dados das interações? | 1 | 0,5 | 0,5 | 0 | 0 | 0,5 | 1 |