

RELATÓRIO TRIMESTRAL DE MONITORAMENTO DE IA

Arts. 41 e 42 da Resolução CNJ nº 615/2025

Documento:	Relatório Anual de Monitoramento de Soluções de IA
Código:	IA-07 (Pós-Produção / Sustentação)
Base normativa:	Arts. 41, 42, 18 e 25 da Resolução CNJ nº 615/2025
Período de referência:	Abril a Junho de 2026
Processo SEI:	0007600-34.2024.8.01.0000
Responsável:	SETIC/SUDIG/DIDIA
Versão / Data:	1.0 — 22/07/2026

HISTÓRICO DE REVISÕES

Versão	Data	Autor	Descrição
1.0	22/07/2026	Gabriel Guedes	Versão inicial
1.0	22/07/2026	Salomão Mafalda	Revisão

0. Orientações Gerais

Este relatório consolida o monitoramento contínuo das soluções de IA em produção no TJAC (Art. 41), com periodicidade mínima anual. Pode ser elaborado por solução ou de forma consolidada pela Divisão de IA e RPA. As informações alimentam a publicação no catálogo do CNJ (Art. 25) e dialogam com o relatório anual do Comitê Nacional (Art. 18).

1. Identificação e Visão Geral do Projeto

Nome	ADA – Assistente Digital Ampliada
Tribunal	TJAC SETIC/Divisão IA e RPA
Situação	Em produção
Classificação risco	Baixo (Art. 11 §3º CNJ 615)

2. Arquitetura Técnica e Modelos (LLMOps)

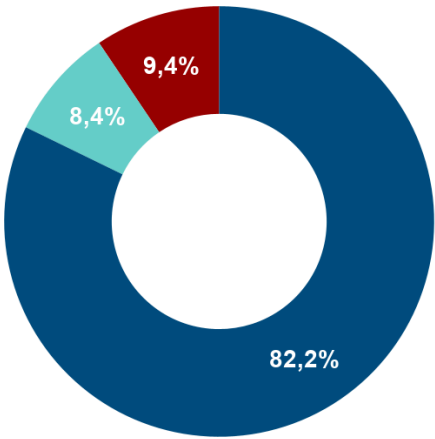
No período, a solução operou com três modelos de linguagem, com predominância do Gemini 2.5 Pro. A tabela abaixo apresenta os modelos utilizados e as camadas de mitigação de alucinação empregadas, que atuam de forma combinada: a base RAG fornece contexto documental confiável, os embeddings garantem busca por similaridade semântica, a engenharia de prompt orienta o comportamento do modelo, e a validação humana (Arts. 32 e 34 da Resolução CNJ nº 615/2025) confere a etapa final de checagem na entrega ao usuário.

Modelos	Fornecedor / Hospedagem	Uso no Período
gemini-2.5-pro	GOOGLE / Serpro	10.458
gemini-3.1-preview	GOOGLE / Serpro	1.063
gpt-4.1	Servidor	1.200
RAGs		
Base PANGEA		
Embeddings		
Vetorização Semântica para busca		
Mitigação Alucinação		
RAGs		
Engenharia de Prompt		
Validação Humana (Art. 32, 34 CNJ 615)		
Filtro de Segurança		

O gráfico mostra a proporção de requisições atendidas por cada modelo no trimestre. O Gemini 2.5 Pro concentrou a maior parte do uso.

Uso por LLM

- gemini-2.5-pro
- gemini-3.1-preview
- gpt-4.1



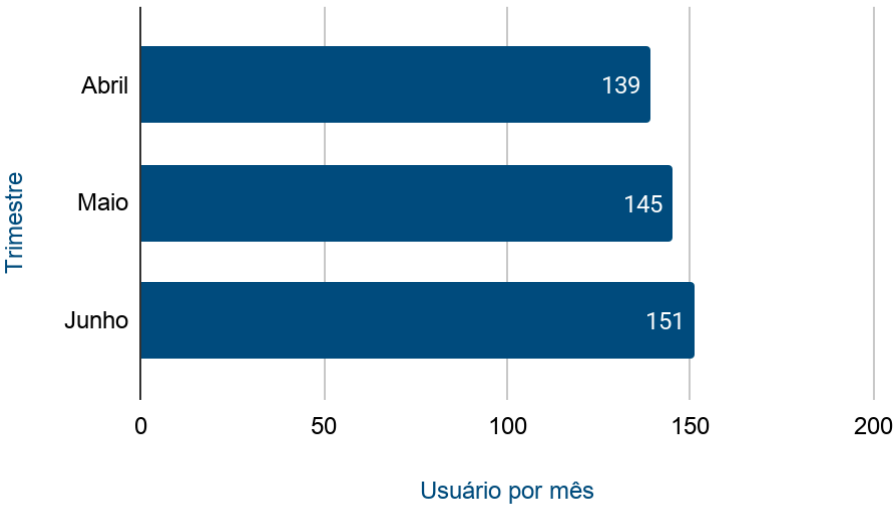
3. Métricas de Usabilidade

A tabela a seguir resume o uso da solução no trimestre. Os gráficos abaixo detalham a evolução mês a mês.

Métrica	Abril a Junho de 2026
Usuários Ativos	204
Qtd. de Requisições	120.939
Consulta de Precedentes	1.644
Total Transcrições	2.898
Total de Tokens	1.073.048.811

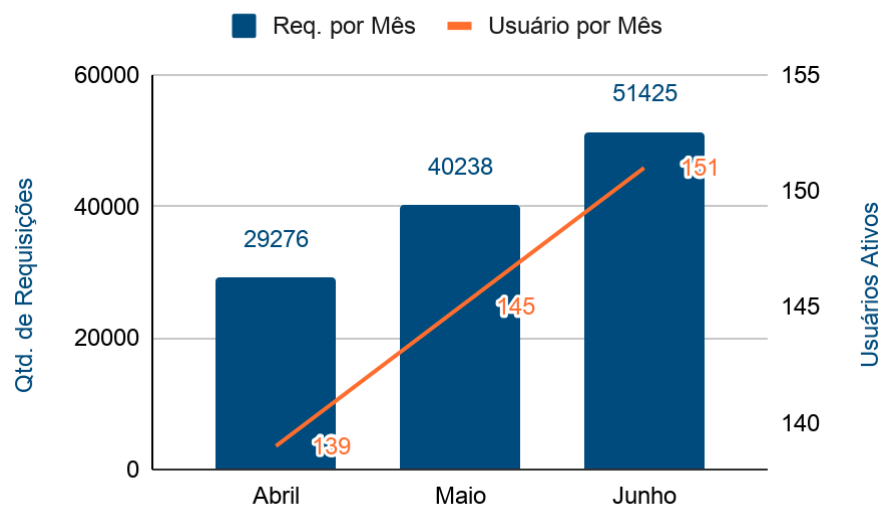
Número de usuários ativos em cada mês do trimestre, com crescimento contínuo de abril a junho.

Usuário por mês no Período



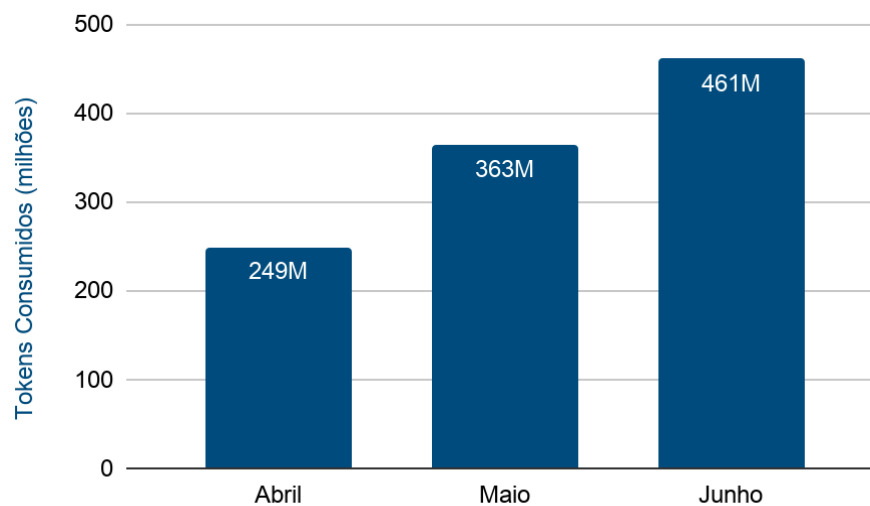
Comparativo entre volume de requisições (barras) e usuários ativos (linha) mês a mês. O crescimento das requisições acompanha o aumento de usuários.

Evolução Mensal de Uso - Trimestre



Volume de tokens processados por mês. O crescimento reflete tanto o aumento de usuários quanto o de requisições por usuário.

Consumo de Tokens por Mês



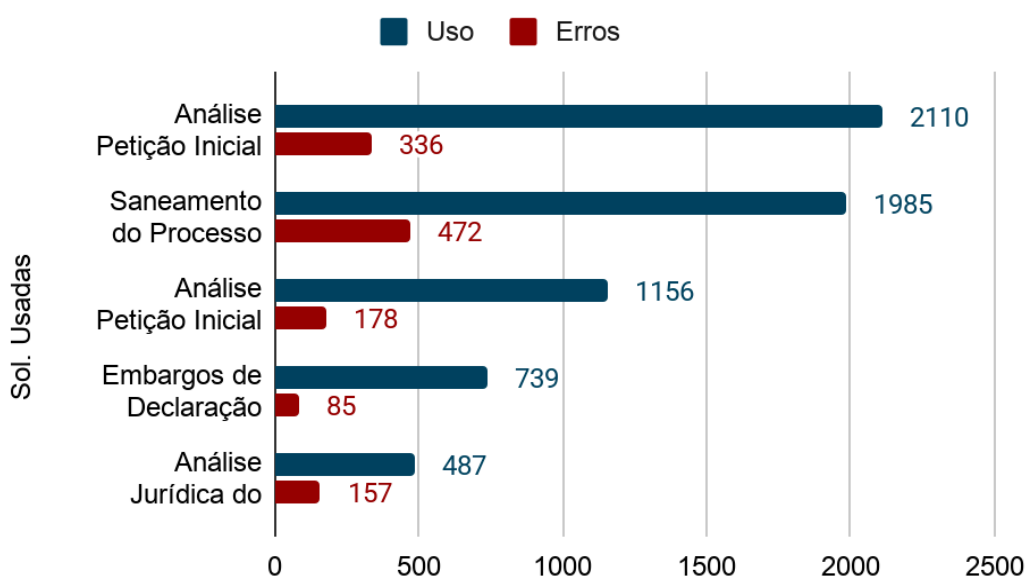
4. Desempenho Técnico

A tabela apresenta o volume de requisições por serviço (os cinco mais utilizados no período) e a quantidade de erros técnicos registrados. Os erros correspondem a falhas de comunicação entre a aplicação e os servidores dos modelos (erros HTTP 500), geralmente associadas a *timeout* e não a erros de conteúdo ou de qualidade das respostas.

Soluções	Requisições	Erros	Taxa de Erro (%)
Análise Petição Inicial Genérica	2.110	336	15,92%
Saneamento do Processo	1.985	472	23,78%
Análise Petição Inicial FIRAC+ (Prof. George Marmelstein)	1.156	178	15,40%
Embargos de Declaração Cível 1º Grau	739	85	11,50%
Análise Jurídica do Recurso de Apelação Criminal	487	157	32,24%

Comparativo entre volume de uso e quantidade de erros técnicos (500/*timeout*) por solução. A leitura conjunta ajuda a identificar quais soluções são mais sensíveis a instabilidades de conexão com os servidores dos modelos.

Desempenho Técnico por Solução



5. Segurança Ofensiva: Prompt Injection & Defesas

Em junho de 2026, foi implementada a funcionalidade de checagem automatizada de tentativas de prompt injection: comandos inseridos em documentos com o objetivo de burlar as regras de funcionamento do sistema.

O sistema realiza uma análise prévia de cada documento antes de liberá-lo para processamento. Caso essa análise identifique indícios de tentativa de injection, o documento recebe uma flag e o usuário visualiza os contextos específicos que motivaram o alerta. Cabe ao usuário revisar o caso: se confirmada a tentativa, o arquivo é denunciado dentro do sistema; caso contrário, é tratado como falso positivo. Todas as ocorrências, confirmadas ou não, são registradas em log e acompanhadas pelo sistema de monitoramento.

Desde a implementação, em junho, foram verificados 5.239 documentos. Desses, 31 (aproximadamente 0,6%) apresentaram pontuação de risco maior que zero e foram sinalizados para revisão do usuário. Todos os 31 casos foram classificados como falsos positivos, não havendo, portanto, tentativas de prompt injection confirmadas no período.

Por se tratar do primeiro mês de operação da funcionalidade, ainda não há série histórica para comparação. A evolução mês a mês dessas métricas, incluindo a taxa de falsos positivos ao longo do tempo, passará a ser apresentada a partir do próximo relatório trimestral.

6. Anonimizador: Desempenho e Conformidade

O módulo de anonimização está ativo e realiza a identificação e o mascaramento de dados pessoais e sensíveis antes do envio das informações aos modelos de linguagem.

Neste primeiro ciclo, o esforço concentrou-se na estabilização da arquitetura do módulo. A contabilização exata do volume de documentos processados ainda está em fase de implementação, com previsão de apresentação dos dados quantitativos no próximo relatório trimestral.

7. Eventos Adversos e Gestão de Riscos (Art. 42 CNJ 615)

Por se tratar do primeiro relatório trimestral de monitoramento, não há eventos adversos a registrar no período. A tabela abaixo será preenchida conforme ocorrências futuras, com data, descrição, causa e medidas adotadas.

Data	Descrição do Evento	Causa	Medidas Adotadas