

## JANELA DE CONTEXTO E TOKENS

A **janela de contexto** é a **memória de trabalho temporária** utilizada pelos **modelos de Inteligência Artificial (IA)** para processar informações durante uma interação. Este componente técnico **define a quantidade de dados que o sistema consegue analisar simultaneamente**.

Assim como um **operador do Direito** analisa um processo, ele o faz dentro de um **"contexto"**: a petição inicial, a contestação e as provas daquele caso específico. **A IA faz o mesmo**.

A janela de contexto define a **quantidade máxima de dados que o sistema consegue "lembrar", correlacionar e analisar simultaneamente** dentro de um mesmo diálogo.

Essa capacidade é **medida em tokens** (unidades de texto, como palavras, fragmentos ou símbolos). O **volume total de tokens no prompt** (instrução) e na **resposta** determina a **"memória ativa" do sistema**. Quanto maior a janela, mais informações o modelo considera de forma integrada, **mantendo a coerência em diálogos ou análises de documentos extensos**.

**Tokens** são as **unidades mínimas de texto** que a IA processa. Um token pode ser **uma palavra, um fragmento de palavra, um número ou um símbolo** (como uma vírgula ou um parêntese). Em média, na língua portuguesa, **100 tokens correspondem a cerca de 75 palavras**, ou seja,  $\frac{3}{4}$  de uma palavra. Uma **página com espaçamento simples e fonte 12** pode conter cerca de **500 palavras**, mas isso pode variar de acordo com o conteúdo.

Tudo o que acontece na interação **consome tokens e ocupa espaço na janela de contexto**:

- **As suas instruções (prompts);**
- **O conteúdo integral dos arquivos que você anexa (PDFs, TXTs);**
- **As respostas geradas pela própria IA.**

Quando a janela de contexto é sobrecarregada com dados em excesso, ocorre **perda de funcionalidade**.

Nessas situações, **partes importantes do histórico** ou dos documentos podem ser **desconsideradas** (a IA "esquece" o que veio primeiro). As respostas tendem a ficar **genéricas, incompletas** ou podem **ignorar fatos cruciais** que estavam no início de um arquivo anexado.

Um dos mecanismos para otimização da memória é a engenharia de prompt: a elaboração de instruções lógicas, claras e enxutas. Essa prática evita o desperdício de **tokens** e maximiza a eficiência da **janela de contexto**, resultando em respostas mais precisas.

Desta forma, quando a **janela de contexto** é **sobrecarregada com dados irrelevantes ou em excesso**, há **perda de funcionalidade**. Nesses casos, partes importantes do histórico podem ser desconsideradas (**truncamento**) e a **coerência do modelo** tende a se reduzir.

Para garantir o melhor desempenho, recomenda-se **utilizar entre 30% e 60% do limite total da janela de contexto do modelo**. Quando esse limite é ultrapassado, há maior risco de **lentidão no processamento (latência)**, **corte de informações (truncamento)** e **redução na qualidade das respostas**, assim **filtrar os arquivos juntados**, se aumenta a eficiência.

O **gerenciamento estratégico de arquivos e da própria conversa** é, portanto, **crucial para a efetividade**. A IA processa todos os dados inseridos, sejam arquivos ou histórico de diálogo. Se forem anexados **documentos irrelevantes ou redundantes** (ex: cópias, anexos desnecessários), eles **consumirão tokens preciosos**, diluindo o foco da análise. Da mesma forma, **conversas excessivamente longas** esgotam a janela de contexto, fazendo com que a IA "esqueça" as instruções ou fatos mencionados no início.

Para **tarefas complexas**, é mais eficaz "**limpar a mesa**": **inicie uma nova conversa e anexe apenas os documentos estritamente necessários** para aquela análise específica.

**Exemplos de janelas de contexto por modelo**

- [Google - NotebookLM](#): cada notebook pode ter 50 fontes, com até 500.000 palavras cada.
- [Google - Gemini 2.5 Pro](#): 1.048.576 tokens de entrada e 65.536 de saída.
- [OpenAI - GPT-5](#): 400k de contexto e 128K tokens de saída.
- [Anthropic - Claude \(modelos\)](#): 200k tokens (500 páginas ou 100 imagens) usuários padrões.
- [Meta - Llama 4 Maverick](#): 1 milhão de tokens.
- [Meta - Llama 4 Scout](#): 10 milhões de tokens.

| Características          | Utilidade          | Perda de Funcionamento     |
|--------------------------|--------------------|----------------------------|
| Capacidade limitada      | Prompts claros     | Excesso de tokens          |
| Processamento sequencial | Contexto relevante | Prompts desorganizados     |
| Flexibilidade de entrada | Divisão em blocos  | Informações contraditórias |

**Ministério Público do Estado de Mato Grosso**  
**Centro de Apoio Operacional de Defesa de Dados Pessoais**  
**e Inteligência Artificial**

**Equipe do Centro de Apoio Operacional de Defesa de Dados Pessoais e Inteligência Artificial**

**Membro Coordenador do Centro de Apoio Operacional de Defesa de Dados Pessoais e Inteligência Artificial**

Adalberto Ferreira de Souza Junior – Promotor de Justiça do Ministério Público do Estado de Mato Grosso

**Membro Coordenador Adjunto do Centro de Apoio Operacional de Defesa de Dados Pessoais e Inteligência Artificial**

Fabício Miranda Mereb – Promotor de Justiça do Ministério Público do Estado de Mato Grosso

**Membro Colaborador do Centro de Apoio Operacional de Defesa de Dados Pessoais e Inteligência Artificial**

Adalberto Biazotto Junior – Promotor de Justiça do Ministério Público do Estado de Mato Grosso

**Membro Colaborador do Centro de Apoio Operacional de Defesa de Dados Pessoais e Inteligência Artificial**

Leoni Carvalho Neto – Promotor de Justiça do Ministério Público do Estado de Mato Grosso

**Servidores**

Maria Cristina Alves Ormond - Auxiliar Ministerial

Pedro Carlos Nogueira Felix - Residente

Vitor Hugo Cabral Araujo - Voluntário

**Elaboração do Material :**

Adalberto Ferreira de Souza Junior - Promotor de Justiça e Coordenador

Fabício Miranda Mereb - Promotor de Justiça e Coordenador Adjunto

Adalberto Biazotto Junior - Promotor de Justiça e Colaborador

Leoni Carvalho Neto - Promotor de Justiça e Colaborador

Maria Cristina Alves Ormond - Auxiliar Ministerial

Pedro Carlos Nogueira Felix - Residente

Vitor Hugo Cabral Araujo - Voluntário



**MPMT**  
Ministério Público  
DO ESTADO DE MATO GROSSO



**CAO/MPMT**  
DEFESA DE DADOS PESSOAIS  
E INTELIGÊNCIA ARTIFICIAL