PROMPT ENGINEERING · 2026

Prompts que Economizam Tokens.

Modelos de prompt prontos para cada finalidade — pensando no que cada palavra custa em Haiku, Sonnet, Opus e Fable

Adriano Mota
Fundador & CEO · Omni8 Soluções
@adrianoafmota · Julho 2026

PROVOCAÇÃO

"O prompt errado não erra a resposta. Ele erra o custo."

Adriano Mota · Omni8 Soluções · Julho 2026

Neste carrossel:

→ O que realmente consome token dentro de um prompt

→ Um modelo de prompt pronto para cada finalidade e cada modelo da Claude

→ Como o cache de prompt corta até 90% do custo repetido

→ Os erros que multiplicam sua conta sem você perceber

I — O QUE VOCÊ ESTÁ PAGANDO

Todo prompt tem 4 partes que custam.

Token é a unidade que a Claude usa pra "ler" e "escrever" texto — e cada uma dessas partes do prompt é cobrada:

- → **Instruções do sistema** — o "papel" e as regras que você (ou o app) definem
- → **Histórico da conversa** — tudo que já foi dito antes, reenviado a cada nova mensagem
- → **O que você escreve agora** — sua pergunta e qualquer arquivo colado
- → **A resposta gerada** — geralmente custa mais caro por token do que a entrada

> → *Saída custa mais que entrada em todos os modelos atuais — pedir "seja direto" já é economia na prática.*

II — ANTES DE TROCAR DE MODELO

3 hábitos que cortam custo em qualquer prompt.

1

Diga o formato e o tamanho da resposta

"Responda em até 5 linhas" ou "só uma tabela, sem explicação" evita que o modelo gere um texto de saída bem maior — e mais caro — do que você precisava.

2

Não recole o mesmo documento a cada pergunta

Se você já anexou o contexto, continue perguntando na mesma conversa. Reenviar tudo de novo dobra o custo de entrada sem necessidade.

3

Combine o modelo certo com o prompt certo

Um prompt econômico no modelo errado ainda sai caro. O modelo certo é a primeira decisão — o prompt é a segunda.

III — HAIKU 4.5

Prompt para volume: classificar e extrair.

FINALIDADE

Classificação, extração de dados, respostas curtas e tarefas repetitivas em massa — onde velocidade importa mais que profundidade.

MODELO DE PROMPT

```
"Classifique cada item da lista abaixo em [categorias].
Responda apenas com [tabela / JSON], sem explicações."
```

💰 **NOTA DE TOKENS**

É o modelo mais barato por token da família — $1 de entrada e $5 de saída por milhão. Prompt curto + saída objetiva = o par ideal para rodar em volume alto sem estourar o orçamento.

IV — SONNET 5

Prompt para o dia a dia: escrever e revisar.

FINALIDADE

Escrita, revisão, código e resumo — a maior parte do trabalho comum, com boa qualidade e custo controlado.

MODELO DE PROMPT

```
"Você é um [papel]. [Revise/escreva] a partir do texto abaixo.
Limite: [X palavras]. Não repita o texto original — entregue só
o resultado final."
```

💰 **NOTA DE TOKENS**

Preço introdutório de $2 de entrada e $10 de saída por milhão (até 31/ago/2026, depois $3/$15). Pedir "não repita o original" evita pagar de novo por um texto que você já tem.

V — OPUS 4.8

Prompt para análise: carregue o contexto 1 vez.

FINALIDADE

Raciocínio em várias etapas, planejamento e revisão de documentos longos — onde errar sai mais caro que pensar um pouco mais.

MODELO DE PROMPT

```
"Aqui está todo o contexto necessário: [contexto completo]. A
partir de agora, nas próximas perguntas desta conversa, não
repita esse contexto — responda direto com base nele."
```

💰 **NOTA DE TOKENS**

$5 de entrada e $25 de saída por milhão. Carregar o contexto pesado uma única vez e continuar na mesma conversa aproveita o cache de prompt — releia o contexto na próxima pergunta custa bem menos do que reenviá-lo.

VI — FABLE 5

Prompt para o mais difícil: peça só a conclusão.

FINALIDADE

Pesquisa profunda, agentes de longa duração e decisões com muitas variáveis — reservado para o problema mais difícil da semana.

MODELO DE PROMPT

```
"Objetivo: [meta ampla]. Pense internamente nos trade-offs
entre as opções e me entregue apenas a conclusão final e o
plano de ação — sem narrar o raciocínio."
```

💰 NOTA DE TOKENS

É o modelo mais caro da família — $10 de entrada e $50 de saída por milhão, o dobro do Opus. Pedir só a conclusão evita pagar por um raciocínio inteiro narrado como texto de saída.

O QUE CADA PROMPT CUSTA

Preço por milhão de tokens, lado a lado.

MODELO	IDEAL PARA PROMPTS DE...	ENTRADA / SAÍDA
Haiku 4.5	Classificação, extração, respostas curtas	$1 / $5
Sonnet 5	Escrita, revisão e código do dia a dia	$2 / $10*
Opus 4.8	Análise e planejamento com várias etapas	$5 / $25
Fable 5	O problema mais difícil, do início ao fim	$10 / $50

*Preço introdutório do Sonnet 5 até 31/ago/2026; depois $3/$15. Fonte: Claude Platform Docs — Pricing, jul/2026.

O TRUQUE QUE POUCO MUNDO USA

Perguntar a mesma coisa várias vezes tem desconto.

Quando você mantém o mesmo documento ou instrução no início da conversa e só troca a pergunta, a Claude pode reaproveitar essa parte em vez de reprocessá-la do zero a cada mensagem — isso se chama cache de prompt.

até 90%

de redução no custo da parte reaproveitada da entrada, em comparação com reprocessá-la do zero a cada mensagem

→ Na prática: cole o documento uma vez, depois só faça novas perguntas na mesma conversa — em vez de recolar tudo de novo a cada pergunta.

O QUE MULTIPLICA SUA CONTA

4 erros que inflam o consumo de tokens.

✗ **Colar o mesmo documento gigante em cada nova pergunta**
Conserto: mantenha a conversa e só pergunte de novo — sem recolar o contexto.

✗ **Pedir "explique passo a passo" em tarefas simples**
Conserto: peça o raciocínio só quando for realmente usar — ele vira saída paga.

✗ **Não limitar o tamanho da resposta**
Conserto: diga o formato e o teto de linhas ou palavras logo no prompt.

✗ **Usar Opus ou Fable para o que o Haiku resolveria**
Conserto: combine finalidade e modelo antes de escrever o prompt.

FONTES DESTE GUIA

Dados oficiais, para você conferir.

Claude Platform Docs — Pricing
platform.claude.com/docs/en/about-claude/pricing

Claude Platform Docs — Prompt caching
platform.claude.com/docs/en/build-with-claude/prompt-caching

Claude Platform Docs — Prompt engineering overview
platform.claude.com/docs/en/build-with-claude/prompt-engineering/overview

Claude Platform Docs — Models overview
platform.claude.com/docs/en/about-claude/models/overview

@adrianoafmota

Cada prompt que você escreve tem um preço. Agora você sabe qual.

Salve esse carrossel e use os modelos de prompt na próxima vez que abrir a Claude.

Conheça a Omni8 → omni8.com.br

#omni8 #claude #promptengineering #IA #tokens #haiku #sonnet #opus #fable