VIVER DE IA

VIVER DE IA

RAFAEL MILAGRE AI

5.1

GUIA EXECUTIVO · LANÇAMENTO 2026

Claude

Fable 5.1

O que mudou, quando usar e como provar valor sem migrar no escuro.

A TESE DESTE GUIA

O modelo só vale quando transforma raciocínio profundo em trabalho concluído, verificável e economicamente melhor.

1M de contexto

128K de saída

Piloto em 7 dias

LEITURA EM 60 SEGUNDOS

01

O lançamento importa. A troca automática, não.

O Fable 5.1 foi lançado em 1º de setembro de 2026 para trabalhos em que raciocínio profundo, autonomia e continuidade importam mais do que resposta instantânea. O ganho precisa aparecer na sua tarefa.

VEREDITO

Use o Fable 5.1 como candidato para tarefas difíceis e longas. Mantenha modelos menores nas rotinas simples.

MAIS AUTONOMIA

Agentes sustentam execuções mais longas

Planejamento, ferramentas, checkpoints e validação final passam a caber no mesmo trabalho.

CONTEXTO DE 1 MILHÃO

Mais contexto não elimina arquitetura

Relevância, recuperação e organização continuam decidindo o que o modelo realmente usa.

SAÍDA DE 128 MIL

Artefatos extensos sem fragmentação precoce

Desde que o pedido traga estrutura, critérios e um estado final verificável.

THINKING ADAPTATIVO

Effort vira alavanca de qualidade e custo

Comece em medium e suba apenas quando a avaliação justificar.

PERGUNTA CERTA

Qual tarefa difícil eu consigo concluir melhor — e com menos retrabalho — usando este modelo?

FICHA TÉCNICA

02

Quatro números mudam a decisão.

Capacidade, orçamento e migração precisam ser lidos juntos. O número isolado quase sempre conta uma história incompleta.

CONTEXTO

1M

tokens para repositórios, dossiês e agentes longos

SAÍDA

128K

tokens para artefatos extensos e estruturados

PREÇO BASE

10/50

US$ por milhão de tokens de entrada / saída

CACHE

−75%

leitura a US$ 0,25/M versus o Fable 5

MODEL ID

`claude-fable-5-1`

THINKING

Adaptativo e sempre ativo

Você controla o nível de esforço em vez de simplesmente ligar ou desligar raciocínio.

PONTO DE PARTIDA

medium

A Anthropic posiciona o esforço médio perto da qualidade do Fable 5 com custo menor. Use como baseline e compare.

NÃO CONFUNDA

Capacidade com necessidade. Para classificação simples, extração curta ou chat de alto volume, um modelo menor tende a ser a decisão mais racional.

ESCOLHA DO MODELO

03

Fable 5.1, Opus 5 ou Sonnet 5?

Comece pelo modelo que tem mais chance de passar no seu critério. A novidade entra no teste; não entra automaticamente em produção.

Ponto de partida

Opus 5

Produto, código do dia a dia e trabalhos gerais com alto padrão.

- Boa baseline antes de aumentar profundidade
- Menor risco de migração desnecessária
- Mantenha se já passa no seu teste

Raciocínio profundo

Fable 5.1

Agentes longos, pesquisa complexa e execuções com muitas ferramentas.

- Mais adequado para planos multietapas
- Contexto e saída muito extensos
- Só sobe se reduzir falha e retrabalho

Velocidade e volume

Sonnet 5

Tarefas com caminho claro, alto volume e sensibilidade a custo.

- Boa opção para rotinas previsíveis
- Latência e custo dominam a decisão
- Não use profundidade sem necessidade

REGRA DE DECISÃO

Escolha o modelo mais barato e rápido que passa consistentemente em qualidade, segurança e conclusão.

CASOS DE USO

04

Onde ele tem mais chance de pagar a conta.

Priorize trabalhos com consequência empresarial, múltiplas etapas e um resultado que alguém consiga aprovar ou reprovar.

1. **Pesquisa e decisão executiva**
   Cruzar fontes, identificar divergências e produzir uma recomendação com premissas explícitas.
2. **Engenharia de software de longo horizonte**
   Entender um repositório grande, alterar múltiplos módulos e validar o resultado com testes.
3. **Documentos, planilhas e apresentações**
   Produzir um pacote coerente de artefatos, preservando cálculos, narrativa e consistência visual.
4. **Operações com ferramentas**
   Executar uma jornada com APIs, arquivos e sistemas, mantendo estado, checkpoints e critérios de parada.
5. **Análise de risco e cenários**
   Comparar alternativas e separar fato, inferência e lacuna antes de uma decisão.

BOM PILOTO

Uma tarefa de 30 minutos a 4 horas, com exemplos anteriores e qualidade avaliável.

QUANDO NÃO USAR

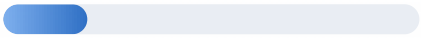
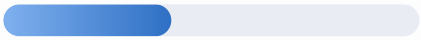
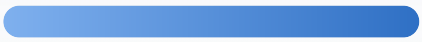
Resumo curto, classificação simples, latência crítica ou processo sem critério de qualidade.

EFFORT

05

A menor profundidade que passa no teste.

Não escolha max por reflexo. O objetivo é encontrar o menor nível que mantém aprovação, segurança e conclusão.

`low`	**Triagem e caminho claro** Veja se encontra o essencial sem pular fontes ou etapas.	
`medium`	**Baseline de migração** Compare qualidade, custo, tempo e revisão humana.	
`high`	**Problemas difíceis** Exija plano, checkpoints e evidência de conclusão.	
`xhigh`	**Complexidade comprovada** Use quando medium ou high falharem de modo recorrente.	
`max`	**Exceção avaliada** Só permaneça se o ganho incremental for material.	

1
15–30 casos
Normais, bordas e recusas.

2
Medium primeiro
Crie a linha de base.

3
Um abaixo e acima
Mude apenas o effort.

4
Decida por aprovação
Não por impressão.

MIGRAÇÃO

06

Troque o modelo. Não quebre a conversa.

O Fable 5.1 não é apenas uma troca de ID. Thinking blocks, ferramentas e histórico pedem validação explícita.

01 — TESTE
Atualize o model ID em ambiente controlado.

→

02 — SOMBRA
Mesmo input, sem impacto no usuário.

→

03 — CANÁRIO
Pequena parcela com rollback.

→

04 — EXPANSÃO
Somente após métricas estabilizarem.

- ✓ **Model ID**
  Use claude-fable-5-1 em teste.
- ✓ **Thinking blocks**
  Preserve o histórico sem editar.
- ✓ **Compatibilidade**
  Não envie blocos 5.1 a modelos antigos.
- ✓ **Forced tool use**
  Revise chamadas que obrigam ferramentas.
- ✓ **Effort**
  Registre o nível junto da avaliação.
- ✓ **Cache**
  Meça reaproveitamento e custo real.
- ✓ **Regressão**
  Rode casos reais, bordas e recusas.
- ✓ **Rollback**
  Mantenha rota para o modelo anterior.

MUDANÇA CRÍTICA

Editar mensagens antigas pode invalidar thinking blocks. Trate o histórico como log append-only.

PROMPTING

07

O modelo precisa saber quando terminou.

Tarefas longas funcionam melhor quando o estado final, as restrições, o critério de evidência e o uso de ferramentas estão claros.

01 OBJETIVO
Defina o artefato final.

→

02 CONTEXTO
Entregue apenas o relevante.

→

03 RESTRIÇÕES
Limites, risco e escopo.

→

04 EXECUÇÃO
Ferramentas e checkpoints.

→

05 EVIDÊNCIA
Teste e critério de parada.

PADRÕES QUE MELHORAM O TRABALHO

1 Peça a conclusão inteira
Executar, validar e entregar; não parar no plano.

2 Separe fato, inferência e lacuna
Facilita revisão e reduz certeza artificial.

3 Autorize paralelismo seguro
Paralelize fontes independentes, não mudanças dependentes.

PROMPT-BASE

```
Objetivo: [resultado final].
Contexto: [o que é relevante].
Critérios: [como aprovar].
Restrições: [segurança e escopo].
Execução: planeje, use ferramentas, mantenha
checkpoints e conclua.
Evidência: cite fontes e registre testes.
Entrega: comece pelo veredito.
```

PROMPT PACK · 1 DE 2

08

Pesquisa e execução, prontas para adaptar.

Substitua os campos entre colchetes. Não cole segredos, dados pessoais ou informações confidenciais sem arquitetura aprovada.

01 · Pesquisa executiva

```
Investigue [decisão].

Use fontes primárias. Compare pelo
menos três alternativas e procure
evidência que contrarie a hipótese
inicial.

Entregue: veredito, evidências,
riscos, lacunas e próxima decisão.
```

02 · Agente de execução

```
Conclua [tarefa] de ponta a ponta.

Antes de alterar, inspecione o estado
atual. Faça mudanças pequenas e
reversíveis, valide cada etapa e não
declare sucesso sem evidência do
resultado final.
```

COMO USAR

Adicione o formato de saída, o gabarito de qualidade, as ferramentas permitidas e os eventos que exigem aprovação humana.

PROMPT PACK · 2 DE 2

08

Artefato e avaliação, sem resposta vaga.

Dois modelos para transformar material em decisão e avaliar o resultado com critérios observáveis.

03 · Artefato executivo

```
Transforme [material] em [documento, planilha ou apresentação] para [audiência].

Comece pela decisão, preserve números e fontes e elimine conteúdo que não muda a ação. Faça revisão final de consistência.
```

04 · Avaliação

```
Avalie a resposta para [tarefa] usando: exatidão, completude, aderência ao formato, segurança, custo e tempo.

Dê nota de 0 a 4 por critério, cite falhas observáveis e indique se aprova, reprova ou exige revisão.
```

REGRA OPERACIONAL

Se a IA não sabe como provar que terminou, o prompt ainda não está pronto.

PILOTO

09

Sete dias para chegar a uma decisão.

Um piloto bom termina em adotar, ajustar ou descartar — com evidência suficiente para explicar o porquê.

DIA 1	DIA 2	DIA 3	DIA 4	DIA 5	DIA 6	DIA 7
Escolha	**Monte**	**Rode o atual**	**Rode o 5.1**	**Compare**	**Canário**	**Decida**
Dono, baseline, aprovação e limite de risco.	15–30 casos, bordas e recusas.	Modelo, custo, tempo e revisão humana.	Mesmo prompt e dados; effort medium.	Revisão cega sem revelar o modelo.	Parcela pequena, observabilidade e rollback.	Ganho, risco, custo e condição de expansão.

BASELINE MÍNIMA

15–30

casos reais suficientes para revelar padrão, não apenas uma boa demonstração.

DECISÃO FINAL

Adotar, ajustar ou descartar

Documente a condição de expansão, o dono do risco e o ponto de rollback.

NÃO VALE

Piloto sem baseline, sem avaliação cega ou baseado somente no exemplo mais bonito.

SCORECARD

10

Meça custo por trabalho aprovado.

Custo por token é só uma parte. A unidade empresarial é a tarefa concluída com qualidade e risco aceitável.

FÓRMULA DE DECISÃO

```
entrada + saída + ferramentas + retries +
revisão humana
÷ tarefas aprovadas
```

MÉTRICA	COMO MEDIR	O QUE REVELA
Taxa de aprovação	Casos aprovados sem correção ÷ total	Qualidade de primeira passagem
Retrabalho	Minutos humanos de correção por caso	Custo escondido da automação
Conclusão	% que chega ao estado final com evidência	Autonomia real
Custo aprovado	Custo total ÷ casos aprovados	Economia por resultado
Tempo total	Do pedido ao artefato validado	Velocidade operacional
Incidentes	Ação indevida, dado exposto ou fonte falsa	Risco e prontidão

PODE GANHAR

Modelo caro que evita três retries

O custo total cai quando a qualidade de primeira passagem sobe.

PODE PERDER

Modelo profundo numa tarefa simples

A latência e o preço crescem sem mudar o resultado aprovado.

GOVERNANÇA

11

Mais autonomia. Mais observabilidade.

Quanto maior o raio de ação do modelo, mais claro precisa ser o limite entre observar, sugerir e executar.

01

Dados

Minimização, classificação e regra para conteúdo confidencial.

02

Ferramentas

Allowlist, permissões mínimas e separação entre leitura e escrita.

03

Aprovação

Humano antes de publicar, pagar, excluir ou afetar um cliente.

04

Observabilidade

Prompt, modelo, effort, ferramentas, custo e resultado final.

05

Idempotência

Retries não duplicam mensagens, cobranças ou alterações.

06

Rollback

Caminho testado para interromper o agente e voltar ao seguro.

GATE HUMANO

Publicar, pagar, excluir, contratar ou afetar cliente exige aprovação explícita.

Contexto longo não garante atenção uniforme. Benchmarks não substituem seus dados. Saídas importantes continuam exigindo validação técnica, jurídica ou humana conforme o caso.

PRÓXIMA AÇÃO

12

Seu primeiro teste em 20 minutos.

Não comece migrando tudo. Comece criando uma comparação que ensine alguma coisa.

1	**Escolha uma tarefa** Algo real, difícil e verificável que você executou recentemente.	3 min
2	**Recupere o resultado anterior** Essa será a baseline de qualidade, tempo e retrabalho.	3 min
3	**Rode Fable 5.1 em medium** Mesmas ferramentas, mesmo contexto e mesmos critérios.	8 min
4	**Faça revisão cega** Compare sem mostrar qual modelo gerou qual resultado.	4 min
5	**Registre a decisão** Adotar, aprofundar o piloto ou manter o modelo atual.	2 min

LEVAR AO TIME

Qual tarefa crítica hoje depende de alguém experiente pensando por muito tempo — e já tem exemplos suficientes para uma avaliação rigorosa?

A TESE

O Fable 5.1 vale quando transforma raciocínio profundo em trabalho concluído, verificável e economicamente melhor.

FONTES

13

Documentação oficial para aprofundar.

Este guia traduz o lançamento para decisão empresarial. Para implementação, consulte sempre a documentação atual da Anthropic.

Anúncio	**Anúncio oficial do Fable 5.1** anthropic.com/claude-fable-and-mythos-5-1	↗
Modelo	**Visão geral do Fable 5.1** platform.claude.com/docs/en/models/fable-5-1/overview	↗
Prompting	**Guia oficial de prompting** platform.claude.com/docs/en/build-with-claude/prompt-engineering	↗
Catálogo	**Comparação dos modelos Claude** platform.claude.com/docs/en/models/overview	↗

NOTA EDITORIAL

Preços, limites e recursos podem mudar. Valide novamente antes de decisões críticas ou de uma implantação ampla.

VIVER DE IA

IA aplicada a trabalho real, com tese, evidência e consequência empresarial.

rafaelmilagre.com/recursos/claude-fable-5-1