Gustavo Rocha

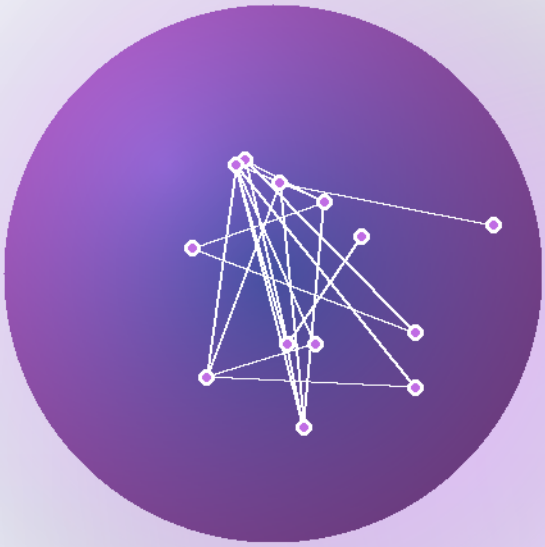
gustavorocha.com gustavo@gustavorocha.com

E-BOOK 2026

Claude 4.8 na advocacia

Como usar esforço, pensamento e modelos no escritório, com exemplos práticos

Modelos · Esforço · Pensamento · Chat · Cowork

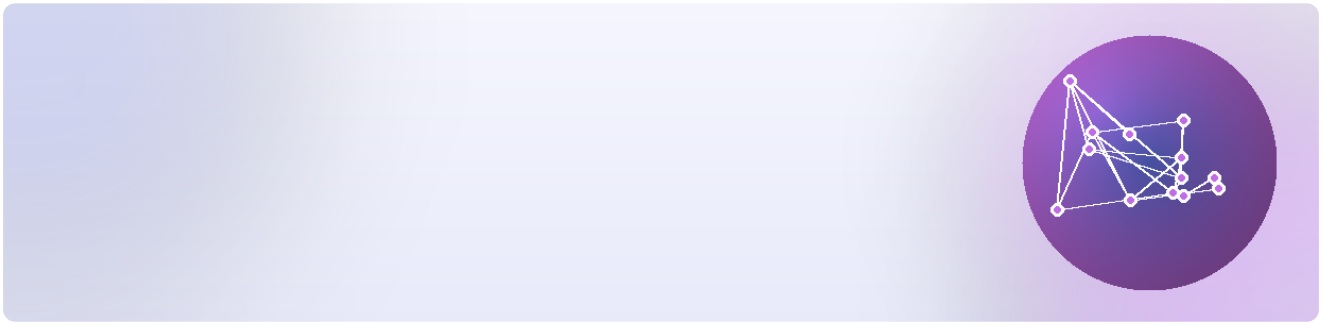


Gustavo Rocha · Consultoria em gestão, tecnologia e marketing jurídico
gustavorocha.com

SUMÁRIO

O que você vai encontrar aqui

Esse material nasceu de uma pergunta que ouço muito de escritório: "Tudo bem, instalei o Claude, e agora, qual botão eu aperto?". A resposta tem três partes, e quem entende as três para de brigar com a ferramenta. Vou te levar por elas com calma, sempre com exemplo de quem vive a rotina de um escritório brasileiro.



CAPÍTULO 01

O que mudou no Claude 4.8

Saiu em 28 de maio de 2026 o Claude Opus 4.8, a versão mais avançada da Anthropic até agora. Como sempre acontece quando aparece um modelo novo, vem junto uma chuva de post dizendo que tudo mudou e que o advogado que não usar vai ficar para trás. Eu prefiro outro caminho. Quero te mostrar o que de fato muda no seu dia a dia e, principalmente, três controles que quase ninguém usa direito.

A parte mais útil do 4.8 não está nos números de benchmark. Está em como você passa a controlar a ferramenta. Essa versão trouxe um raciocínio que se ajusta sozinho conforme a dificuldade da tarefa e um controle de esforço que você regula na mão. Some isso à escolha de qual versão do Claude usar e você tem três alavancas para mexer. A maioria das pessoas só conhece uma.

> Antes de tudo, um pé no chão
> A própria Anthropic divulgou que o 4.8 alcançou a maior pontuação que eles já registraram no benchmark jurídico interno, e ainda assim foi o primeiro modelo a passar de 10% no critério mais rígido, aquele que só conta a tarefa como acertada quando ela passa em tudo. Leia de novo: o melhor modelo do mercado para tarefas jurídicas acerta tudo em cerca de 10% das vezes nesse teste duro. Isso não é motivo para não usar. É motivo para usar do jeito certo, com você no comando e conferindo o resultado.

Guarde esse número. Ele volta no fim, quando falarmos dos cuidados. Por enquanto, vamos entender as tais três alavancas.

CAPÍTULO 02

As três alavancas que ninguém usa direito

Quando você usa o Claude, três decisões mudam completamente o resultado que volta para você. A maioria das pessoas só conhece a primeira, e olhe lá. Domine as três e você sai do uso amador para o uso profissional.

1. Modelo

Qual Claude

A versão que faz o trabalho: Haiku, Sonnet ou Opus. Como escolher entre um estagiário, um advogado de equipe e o sócio sênior.

2. Esforço

Quanto ele se dedica

De baixo a máximo. Regula o tempo e a energia que o Claude gasta na tarefa, e com isso o seu consumo e a sua espera.

3. Pensamento

Quando ele raciocina

O modelo decide sozinho se para para pensar antes de responder. Você influencia isso pelo nível de esforço.

Nas próximas páginas eu abro cada uma, com exemplos da rotina. No fim, junto tudo numa régua simples que você vai usar no automático depois de alguns dias.

> A ideia central
> É a mesma lógica de alocação de equipe que você já faz no escritório. Você não coloca o sócio mais caro para carimbar protocolo, nem manda o estagiário do primeiro mês redigir a tese central de um caso difícil. Com o Claude é igual: a arte está em casar a ferramenta certa com a tarefa certa.

CAPÍTULO 03

Modelos: Opus, Sonnet e Haiku

O Claude vem em três tamanhos, e essa é a primeira escolha que aparece na tela. Pensa neles como três profissionais diferentes que você poderia colocar na mesma tarefa.

Haiku

O estagiário rápido

Faz coisa simples num piscar de olhos: organiza lista, padroniza formatação, classifica e-mails, transcreve, converte. Gasta pouco, responde rápido. Não peça a ele a tese central da defesa.

Sonnet

O advogado de equipe

Resolve a maior parte do trabalho do escritório com qualidade e sem te fazer esperar. Para 80% a 90% do que você precisa no dia, ele dá conta. É por onde eu recomendo começar.

Opus 4.8

O sócio sênior

Você chama quando o caso é cabeludo. Raciocina mais fundo, conecta mais pontos, segura uma análise longa sem se perder. Em troca, é mais lento e consome mais do seu limite.

Como isso aparece para você

No Claude.ai ou no aplicativo, você clica no nome do modelo ali perto do botão de enviar e troca. No plano gratuito você tem Haiku e Sonnet. Para chegar no Opus, precisa do Pro ou superior. Existe ainda uma opção "Auto", que escolhe o modelo sozinho conforme a tarefa. É cômoda, mas você perde a previsão de quanto vai consumir do seu limite. Eu prefiro escolher na mão e saber onde estou pisando.

Na prática, no escritório

Haiku: classificar trezentos documentos de uma due diligence por tipo e data. É volume e padrão, não pede profundidade.

Sonnet: redigir a primeira versão de uma notificação extrajudicial, revisar uma cláusula, responder uma dúvida de procedimento. O feijão com arroz do escritório.

Opus 4.8: montar a estratégia de uma contestação trabalhista complexa, analisar um acórdão para ver se cabe recurso especial, cruzar várias teses e antecipar os contra-argumentos do outro lado.

Esquece os números de versão

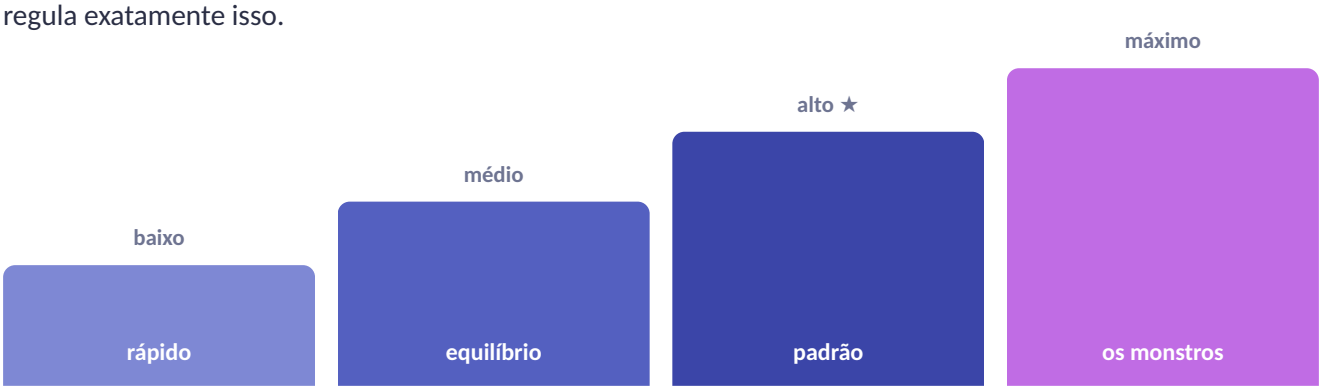
Se você usa o Claude no navegador, não precisa decorar 4.8, 4.6 e afins. O site já usa sempre a versão mais recente de cada modelo. Você só guarda três nomes: Opus, Sonnet e Haiku. Os números importam para quem programa via API e precisa fixar uma versão específica.

CAPÍTULO 04

Esforço: quanto o Claude se dedica

Essa é a alavanca nova que pouca gente está usando, e talvez seja a mais útil para o seu bolso e a sua paciência. No Claude.ai e no Cowork você escolhe o nível de esforço que o Claude aplica na tarefa. São quatro níveis, do mais econômico ao mais caprichado. O 4.8 vem no "alto" como padrão, que a Anthropic considera o melhor equilíbrio entre qualidade e velocidade.

Pensa no esforço como o tanto de tempo e energia que você daria a uma tarefa. Tem coisa que você responde de cabeça, andando pelo corredor. Tem coisa que você senta, fecha a porta e pensa por uma hora. O esforço regula exatamente isso.



★ nível padrão do Claude 4.8

O que cabe em cada nível

Baixo é para o que é rápido e direto: classificação simples, conversão de formato, uma pergunta de "como faço tal coisa no sistema". **Médio** é o meio-termo confortável, trabalho sólido sem torrar tempo. **Alto**, o padrão, serve para a maior parte do que exige raciocínio de verdade: análise de documento, redação que precisa ficar boa, comparação de alternativas. **Máximo** é para os monstros: a análise que toca dezenas de documentos, a auditoria de um contrato gigante, a tarefa que você delegaria a uma equipe inteira.

Na prática, no escritório

Esforço baixo: padronizar a formatação de dez minutas que vieram bagunçadas de estagiários diferentes. Não há o que pensar, é execução.

Esforço alto: análise crítica de uma cláusula de não concorrência, apontando riscos e sugerindo ajustes.

Esforço máximo: jogar uma reclamação trabalhista de quarenta páginas e pedir a estratégia completa de defesa, ponto a ponto, com risco quantificado em três cenários.

A disciplina que corta sua conta

Rodar esforço baixo nas tarefas simples e guardar o máximo para as difíceis reduz bastante o seu consumo mensal sem tocar na qualidade do que importa. É o equivalente a não pagar hora de sócio para tarefa de estagiário.

CAPÍTULO 05

Pensamento: quando ele para para raciocinar

Aqui mora uma das mudanças mais bacanas do 4.8, e ela funciona meio nos bastidores. O modelo agora tem o que a Anthropic chama de pensamento adaptativo. Na prática, ele lê a sua pergunta e decide sozinho se aquilo merece uma parada para raciocinar antes de responder ou se pode ir direto.

Pergunta boba, ele responde na hora. Problema complicado, ele para, pensa, organiza o raciocínio e só então te entrega a resposta. É parecido com o que um bom advogado faz por instinto. Ninguém reflete profundamente antes de responder que a audiência é às 14h. Mas todo mundo deveria parar e pensar antes de cravar uma tese numa peça que define o caso.

> O detalhe que liga as duas coisas
>
> Esse pensamento conversa direto com o nível de esforço. Quando você sobe o esforço para alto ou máximo, o Claude tende a pensar mais e por mais tempo nos problemas difíceis. Quando baixa, ele pensa menos e responde mais rápido. Você não configura nada complicado para isso. Mexendo no esforço, você já está mexendo no quanto ele pensa.

Como isso te ajuda na prática

Nas tarefas em que o raciocínio importa, contestação, parecer, análise de viabilidade recursal, você quer o esforço lá em cima justamente para forçar o modelo a pensar antes de cuspir uma resposta. É nessas horas que a diferença entre uma resposta apressada e uma resposta pensada vira diferença de qualidade no seu trabalho. Nas tarefas mecânicas, você deixa ele responder rápido, sem perder tempo pensando o que não precisa de pensamento.

CAPÍTULO 06

Onde usar: chat ou Cowork

Falta combinar um detalhe que muda bastante o resultado: o lugar onde você aperta esses botões. O Claude aparece para você em dois ambientes bem diferentes, e muita gente usa o errado para a tarefa errada. A separação mais honesta que conheço é essa: o chat pensa com você, o Cowork trabalha por você.

Claude Chat

pensa com você

- É o claude.ai no navegador, no celular e no app de computador.
- Conversacional, uma mensagem de cada vez.
- Não enxerga os arquivos do seu computador. Só vê o que você cola, digita ou sobe.
- Disponível inclusive no plano gratuito.
- Para: rascunhar, resumir o que você colou, debater uma tese, tirar dúvida, escrever um post.

Claude Cowork

trabalha por você

- Roda só dentro do app de computador instalado. Não tem versão no navegador.
- A máquina precisa ficar ligada enquanto a tarefa acontece.
- Agêntico: abre suas pastas, cria e edita arquivos, conecta a ferramentas externas.
- Monta um plano, você aprova, e só então ele age. Roda partes em paralelo e aceita tarefas agendadas.
- Está nos planos pagos. Histórico fica guardado no seu aparelho.

Traduzindo para o escritório

No chat: "resume esse acórdão", "me dá três linhas para essa contestação", "revisa essa cláusula", "escreve esse e-mail para o cliente". Cola, conversa, recebe.

No Cowork: "pega essa pasta com quarenta contratos e me diz quais têm foro de eleição fora do RS", "organiza a pasta desse cliente por tipo e data", "monta o relatório mensal a partir dessas planilhas". O relatório de fim de mês que se repete é o feijão com arroz do Cowork.

Dois avisos que poupam dor de cabeça

Eles não compartilham memória. O que você falou no chat não está no Cowork, e o contrário também. Para passar contexto de um para o outro você cola na mão, sobe os mesmos arquivos ou usa os projetos do Cowork. O erro clássico é achar que o Cowork lembra do que você conversou no navegador. Não lembra.

Mais poder, mais risco. O Cowork acessa pastas com dado de cliente e mexe em arquivo de verdade. A etapa de aprovar o plano antes de ele agir não é burocracia, é o seu freio de mão. Não aprove plano que você não entendeu, do mesmo jeito que não assina petição que não leu.

Régua de bolso: dúvida rápida, rascunho ou brainstorm, abre o chat. Tarefa de várias etapas, com várias pastas e arquivos seus, ou que você quer deixar agendada, aí é Cowork. E o controle de modelo e esforço vale nos dois, com uma diferença: é no Cowork, nas tarefas longas, que subir o esforço rende mais.

CAPÍTULO 07

A régua de bolso do escritório

Na hora de usar, você não vai ficar pensando em teoria. Vai bater o olho na tarefa e decidir. Para facilitar, deixo a régua que uso e ensino. Repare que o mesmo escritório, no mesmo dia, transita entre os três tipos de tarefa.

Tipo de tarefa	Modelo	Esforço	Onde
Mecânica e de volume formatar, classificar, converter, transcrever	Haiku ou Sonnet	baixo	Chat ou Cowork
Dia a dia que precisa ficar bom redigir notificação, revisar cláusula, resumir processo	Sonnet	alto	Chat
Complexa e estratégica montar defesa, analisar acórdão, parecer ao cliente	Opus 4.8	alto ou máximo	Chat
Em lote, multi-arquivo, recorrente varrer pasta de contratos, relatório mensal	Opus ou Sonnet	alto ou máximo	Cowork

> O erro dos dois extremos
>
> Quem coloca tudo no Opus com esforço máximo só queima limite à toa. Quem coloca tudo no Haiku para economizar entrega trabalho raso nos casos que mais importam. O jogo está no meio, casando ferramenta e tarefa.

CAPÍTULO 08

Os cuidados que ninguém pode pular

Agora a parte que separa o uso profissional do uso amador, e que como consultor eu me recuso a deixar de fora.

1. Sigilo

Você lida com dado de cliente, dado sensível, segredo de Justiça. Antes de jogar qualquer coisa em qualquer ferramenta de IA, entenda como aquele provedor trata seus dados, o que faz parte da conta corporativa, o que fica retido e o que não fica. Isso não é detalhe técnico, é dever profissional. A LGPD não tira férias porque você descobriu uma ferramenta nova.

2. Conferência humana, sempre

Volto naquele número do começo. O melhor modelo jurídico do mercado, no teste mais duro, acerta tudo em torno de 10% das vezes. Traduzindo: ele erra, inventa jurisprudência que não existe, cita artigo de lei trocado com uma confiança de dar inveja. A responsabilidade pela peça é sua, com o seu número de OAB nela. O Claude é o estagiário mais rápido e estudioso que você já teve, mas você não assina nada sem ler. Nunca.

3. Não terceirize o raciocínio jurídico

A IA acelera a parte braçal e organiza a parte pensante, mas a estratégia, a leitura de qual juiz é aquele, a empatia com o cliente, o faro do que vai colar e do que não vai, isso continua sendo seu. No dia em que você delegar isso para a máquina, deixou de ser advogado e virou revisor de máquina. E revisor ganha menos.

> Resumo dos cuidados
>
> Trate o Claude como um auxiliar brilhante e veloz, não como um substituto seu. Sigilo em primeiro lugar, conferência humana em tudo que sai com o seu nome, e o raciocínio estratégico permanece com você. Esses três pontos não são opcionais.

CAPÍTULO 09

Roteiro de 15 dias para começar

Se você está chegando agora, não tente dominar as três alavancas de uma vez. Vai pela ordem abaixo e em duas semanas isso vira automático, do mesmo jeito que você regula a água do chuveiro sem pensar.

Período	O que fazer
Dias 1 a 7	Use só o **Sonnet**, no esforço padrão, no chat, para tudo. O objetivo é pegar intimidade com a ferramenta sem se preocupar com configuração.
Dias 8 a 12	Comece a **baixar o esforço** nas tarefas bobas e a **subir para o Opus** nos casos difíceis. Repare na diferença de profundidade das respostas.
Dias 13 a 15	Instale o app de computador e teste o **Cowork** numa tarefa de lote que se repete no seu escritório, como o relatório mensal ou uma varredura de pasta. Aprove o plano com atenção.

O Claude 4.8 é a melhor versão que já saiu para trabalho jurídico, e ainda assim é uma ferramenta. Ferramenta boa na mão de quem sabe usar multiplica resultado. Na mão de quem não sabe, gera retrabalho e uma falsa sensação de produtividade. A diferença entre os dois grupos não está no modelo. Está em quem aprendeu a usar o modelo. Espero que esse material te coloque do lado certo.

> Quer ajuda para aplicar isso no seu escritório?
> É exatamente o tipo de coisa que faço na consultoria: desenhar com a equipe quais tarefas vão para qual ferramenta, com qual nível de esforço, e como manter sigilo e conferência no processo. Me chama.

GUSTAVO ROCHA

Consultoria em gestão, tecnologia e marketing jurídico

Mais de 12 anos ajudando escritórios e departamentos jurídicos no Brasil e em Portugal a usar tecnologia, inteligência artificial e boa gestão a favor do resultado.

site

gustavorocha.com

e-mail

gustavo@gustavorocha.com

Gustavo Rocha