

TARDES DE CONHECIMENTO

Técnicas avançadas de criação de prompts pra uso em auditoria

Mediador:

João Henrique Florence

Auditor de Controle Externo do TCM-PA



Palestrante:

Jonas Gomes de Sousa

Auditor de Controle Externo do TCE-PA



APOIO:





Definição de engenharia de *prompts*

Engenharia de *prompts* é a arte de conversar com a inteligência artificial do jeito certo para obter respostas melhores.

É a técnica de formular perguntas e comandos de **forma estratégica, clara e precisa**, para que a inteligência artificial entenda exatamente o que você quer e entregue a melhor resposta possível.



O MÍNIMO QUE TODO *PROMPT* DEVE (DEVERIA) TER

P **Pessoa** - Descreva a pessoa ou papel que a IA deve assumir.

R **Roteiro** - Defina a tarefa que deve ser realizada.

O **Objetivo** - Informe o objetivo do prompt e o que se espera como resultado.

M **Modelo** - Especifique o formato ou estrutura que a resposta da IA deve seguir.

P **Panorama** - Forneça informações contextuais relevantes para ajudar o modelo a gerar uma resposta mais precisa e específica.

T **Transformação** - Você deve refinar e iterar o prompt com base na resposta obtida.

Exemplo de estrutura mínima de *prompt*

Pessoa - Você é um auditor de controle externo do Tribunal de Contas do Estado do Pará especialista em auditoria de receitas públicas...

Roteiro - Está atuando em uma auditoria nas receitas públicas arrecadadas pelo estado no exercício de 2024, mais específico sobre as receitas de ICMS.

Objetivo - Seu objetivo fazer um comparativo entre a previsão de arrecadação com o valor efetivamente arrecadado, dando sugestões de como aumentar a efetividade na arrecadação...

Modelo - Apresente um relatório contendo a situação encontrada, critérios de análise, evidências, causas e efeitos. Retorne pelo menos cinco mil palavras.

Panorama - Analise sob a luz da reforma tributária (EC 132/2024 e LCP 214/2025).

Transformação - Após a primeira resposta, refine seu prompt a fim de atingir resposta mais precisa.

JANELA DE CONTEXTO

É o **número máximo de tokens** (unidades mínimas de texto, como palavras, sílabas ou até caracteres) que um modelo de linguagem neural pode processar simultaneamente em uma única inferência (prompt + resposta).

Imagine que a IA tem uma memória de curto prazo com espaço limitado. Esse espaço chama-se janela de contexto.

Por que isso importa?

Se o *prompt* for muito longo, partes importantes podem ser esquecidas se ultrapassarem a janela. Para tarefas longas (ex: auditorias), precisamos usar técnicas de resumo ou divisão de tarefas.

Modelo	Tamanho da janela de contexto
GPT-3.5	~4.000 tokens (≈ 3.000 palavras)
GPT-4	~8.000 ou 32.000 tokens (≈ 6k a 24k palavras)
GPT-4o	Até 128.000 tokens (≈ 300 páginas de Word)

TÉCNICAS

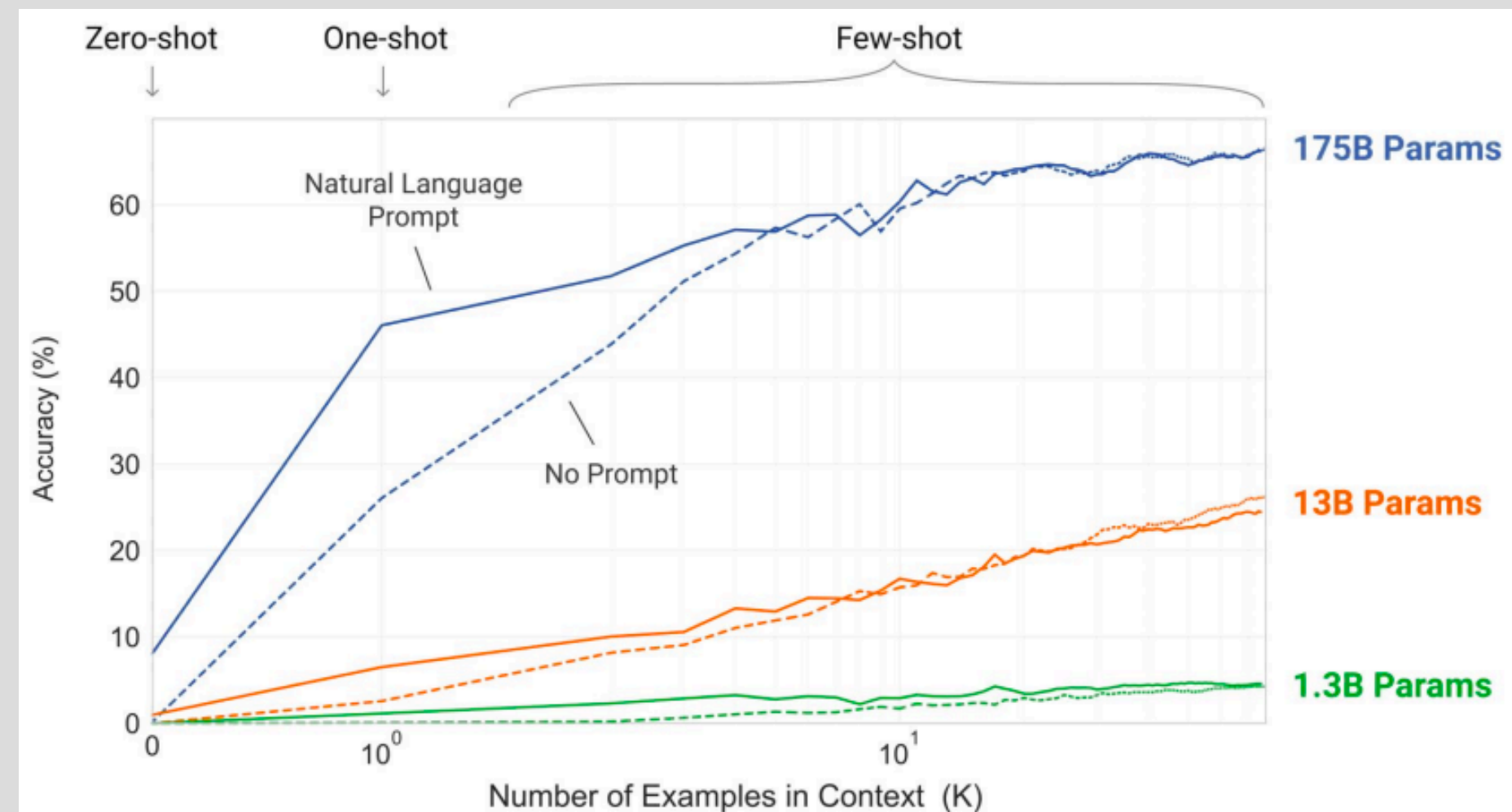
- *Few-Shot*
- Estímulo de *Prompt* Direcional
- Cadeia de Pensamento (*CoT*)
- Árvore de Pensamentos (*ToT*)
- Geração de Conhecimento
- *Prompt* Maiêutico
- Alucinações Controladas
- Bônus

FEW-SHOT

Definição: é uma técnica de engenharia de *prompts* em que o modelo de linguagem **recebe poucos exemplos** (geralmente de 1 a 10) para aprender como deve responder a uma tarefa, sem necessidade de re-treinamento do modelo.

Aplicação: Elaboração de questões de auditoria.

Fonte: *Language Models are Few-Shot Learners*



Exemplo em auditoria

- 1º Altere o OBJETIVO do prompt para: “você deve gerar algumas questões de auditoria relacionadas ao ROTEIRO apresentado. Siga os exemplos apresentados.”
- 2º Insira em PANORAMA pelo menos cinco exemplos de questões de auditoria.
- 3º Caso não tenha exemplos de auditoria, peça primeiramente para a IA te apresentar as principais questões de auditoria relacionadas ao tema presentes no principais trabalhos do TC X, Y ou Z.

Estímulo Direcional

(Directional Stimulus Prompting)

Definição: Consiste no fornecimento de pistas (*hints*) durante ao fim da elaboração do *prompt* gerando respostas bem mais elaboradas.

Aplicação: Análise de grandes volumes de textos.

Fonte: *Guiding Large Language Models via Directional Stimulus Prompting*



Input text

Article: (CNN) For the first time in eight years, a TV legend returned to doing what he does best. Contestants told to "come on down!" on the April 1 edition of "[The Price Is Right](#)" encountered not host Drew Carey but another familiar face in charge of the proceedings. Instead, there was [Bob Barker](#), who hosted the TV game show for 35 years before stepping down in [2007](#). Looking spry at [91](#), Barker handled the first price-guessing game of the show, the classic "Lucky Seven," before turning hosting duties over to Carey, who finished up. Despite being away from the show for most of the past eight years, Barker didn't seem to miss a beat.

Reference

[Bob Barker](#) returned to host "[The Price Is Right](#)" on Wednesday . Barker, [91](#), had retired as host in [2007](#) .

Directional Stimulus Prompting

Q: Summarize the above article briefly in 2-3 sentences **based on the hint.**

Hint: [Bob Barker](#); [TV](#); [April 1](#); "[The Price Is Right](#)"; [2007](#); [91](#).

Model output

A: On [April 1](#), [Bob Barker](#) returned to the [TV](#) show "[The Price Is Right](#)" after eight years . Looking spry at [91](#), Barker handled the first price-guessing game . Barker stepped down as host in [2007](#), but didn't seem to miss a beat .

ROUGE-1: 48.39 👍

Standard Prompting

Q: Summarize the above article briefly in 2-3 sentences.

Model output

A: Bob Barker, the former host of \"The Price Is Right\", returned to the show for the first time in 8 years. He hosted the first game of the show, and then handed the hosting duties over to Drew Carey.

ROUGE-1: 34.48 👎

Exemplo em auditoria

1º Faça o upload do arquivo a ser analisado.

2º Ao fim do *prompt* insira:

“Analise o documento anexo utilizando as *keywords* abaixo:

{*keywords*}: Receitas correntes; Receitas intraorçamentárias; Receitas de capital.

3º Para fins de comparação, abra uma nova janela de contexto e refaça o pedido, porém sem os *hints*.

CADEIA DE PENSAMENTO

Chain-of-Thought (CoT)

Definição: é uma técnica para **fazer a IA pensar passo a passo**, em vez de responder direto, assim como um auditor faria. Sem essa técnica, a IA pode errar por tentar adivinhar direto a resposta, porém com CoT, **ela explica o raciocínio**, o que ajuda a ter mais acertos em perguntas difíceis ou lógicas.

Aplicação: Cálculo de atingimento de índices mínimos em que a resposta não esteja explícita.

Fonte: *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Exemplo em auditoria

1º Faça o upload do Relatório Resumido de Execução Orçamentária - RREO e documentos relacionados.

2º Ao fim do *prompt* insira um exemplo de pergunta com resposta construída em CoT, por exemplo:

“Q: Qual o percentual de atingimento do limite constitucional de aplicação mínimo em MDE em 2023?

R: Calcule a razão da despesa líquida em MDE pela receita de impostos somadas às transferências vinculadas e multiplique por 100. Para cálculos de despesas exclua despesas com ensino superior, despesas com aposentadorias e pensões e obras paradas. Percentual de 25,61%.

Q: Qual o percentual de atingimento do limite constitucional de aplicação mínimo em MDE em 2024?”



CADEIA DE PENSAMENTO CONTRASTIVA

Contrastive Chain-of-Thought (CCoT)

Definição: essa técnica segue a mesma métrica da CoT, porém além de colocar a explicação que deseja, **informa-se qual resposta você não quer**.

Aplicação: Calculo de atingimento de índices mínimos em que a resposta não esteja explícita.

Exemplo em auditoria

Além de executar os passos 1 e 2 do exemplo da CoT do slide anterior, insira o seguinte passo:


3º Resposta errada: “Informe o cálculo apresentado no passo 2, mas dessa vez inclua gasto que não são MDE, como inativos e pensões.”


Isso fará com que o modelo busque, em todas as fontes existentes, de dados a serem excluídos relativos a inativos e pensões, pois frisou-se ser errado mantê-los. É possível que a IA te peça mais documentos como base de pesquisa, ou que você indique sítios eletrônicos que possua os dados.


Fonte: *Contrastive Chain-of-Thought Prompting*




Model Input


 Question : James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?

 Explanation: He writes each friend $3 \times 2 = 6$ pages a week. So he writes $6 \times 2 = 12$ pages every week. That means he writes $12 \times 52 = 624$ pages a year.

 Wrong Explanation: He writes each friend $12 \times 52 = 624$ pages a week. So he writes $3 \times 2 = 6$ pages every week. That means he writes $6 \times 2 = 12$ pages a year.

 Question: James has 30 teeth. His dentist drills 4 of them and caps 7 more teeth than he drills. What percentage of James' teeth does the dentist fix?

Model Output

 Explanation: The dentist drills 4 teeth, so there are $30 - 4 = 26$ teeth left. The dentist caps 7 more teeth than he drills, so he caps $4 + 7 = 11$ teeth. Therefore, the dentist fixes a total of $4 + 11 = 15$ teeth. To find the percentage of teeth the dentist fixes, we divide the number of teeth fixed by the total number of teeth and multiply by 100: $15/30 \times 100 = 50\%$

ARVORE DE PENSAMENTOS

Tree-of-Thought (ToT)

Definição: É uma forma de pensar **abrindo caminhos diferentes ao invés de seguir uma linha única.**

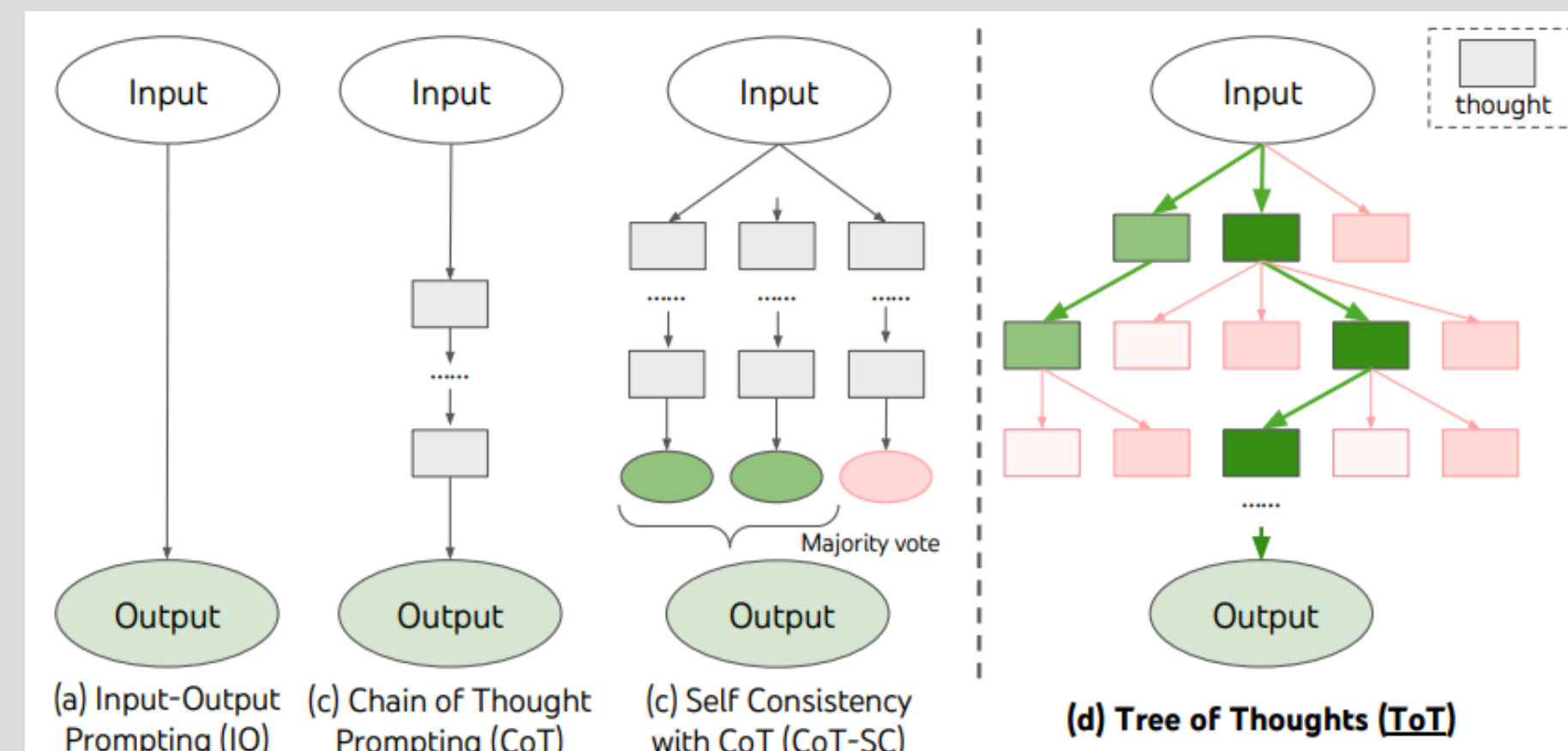
É como explorar várias ideias ao mesmo tempo, comparando cenários antes de decidir qual é o melhor.

A *CoT* é como seguir uma trilha só com uma lanterna.

A *ToT* é como olhar um mapa com várias trilhas, testar algumas e escolher a melhor.

Aplicação: Análise de diferentes aspectos de um mesmo objeto, por exemplo uma licitação.

Fonte: *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*



Exemplo em auditoria

Após gerar o primeiro *prompt* seguindo o mnemônico, em Transformação, insira:

“use a técnica *Tree of Thought (ToT)* para explorar **diferentes caminhos de análise**, considerando:

1. Legalidade da modalidade escolhida (dispensa de licitação),
2. Vícios formais (ausência de justificativa, ausência de parecer jurídico),
3. Execução do objeto (houve entrega? houve medição técnica?).

Para cada caminho, ramifique o raciocínio com cinco subperguntas. Na sequência ramifique as subperguntas em cinco possíveis achados cada. No final, indique qual ramo apresenta maior indício de irregularidade apresentando sua linha de raciocínio.”

Geração de Conhecimento

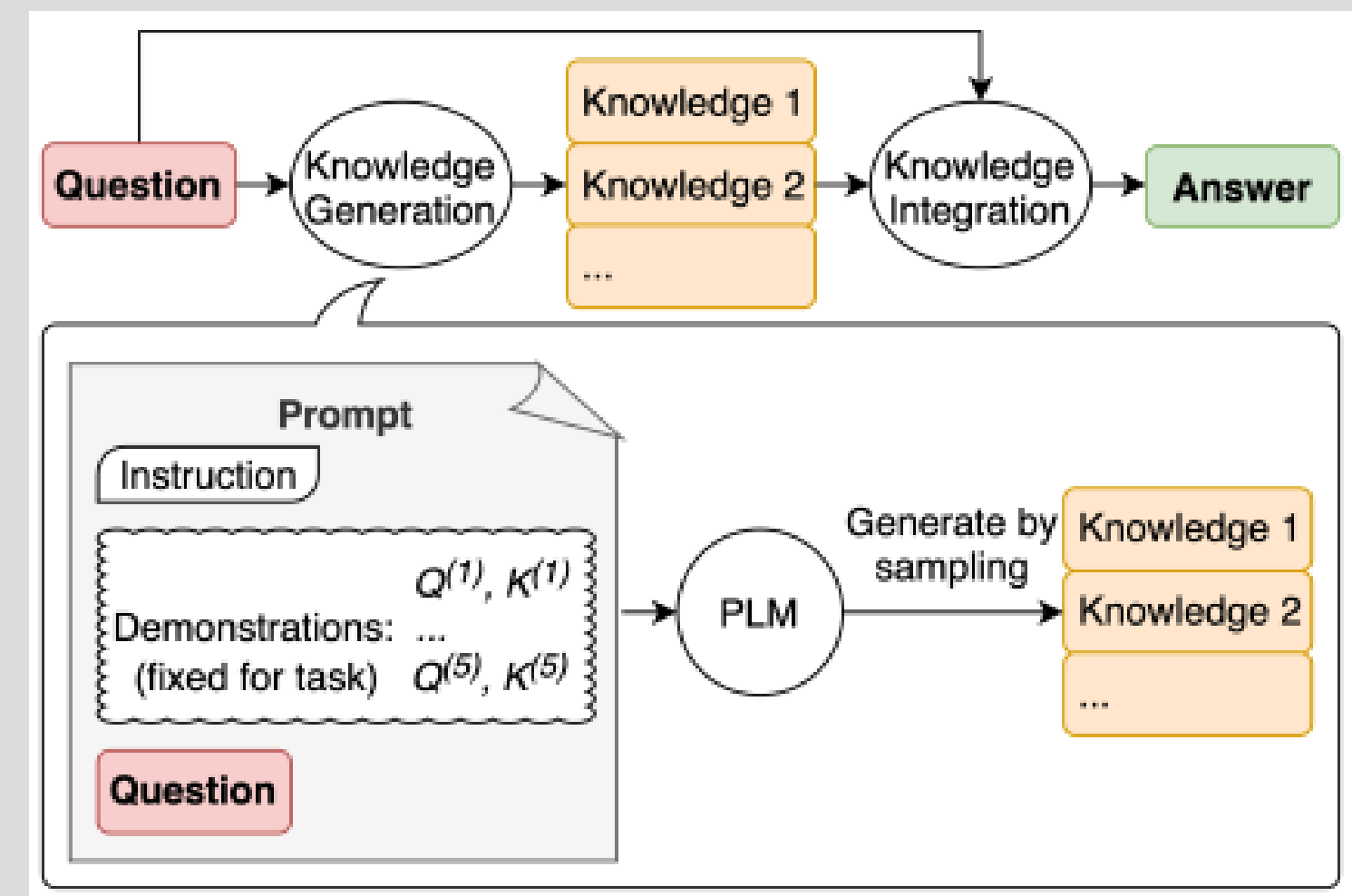
(Generated Knowledge for Commonsense Reasoning)

Definição: é uma técnica que faz a IA **produzir fatos, conceitos e regras relevantes antes de responder uma pergunta.**

Ela separa **o momento de pensar do momento de responder.** Isso melhora a clareza, coerência e lógica do raciocínio, reduzindo erros e respostas apressadas.

Aplicação: Auditorias inéditas em determinados programas de governo.

Fonte: *Generated Knowledge Prompting for Commonsense Reasoning*



Exemplo em auditoria

Em Panorama, informe ser sua primeira auditoria no órgão/assunto e após insira o seguinte *prompt*:

“Antes de emitir qualquer juízo, use a técnica de Geração de Conhecimento por *Prompt* (*Generated Knowledge Prompting*) e:

1. Liste os principais marcos legais e infralegais que regem o órgão/programa XYZ.
2. Aponte as competências institucionais típicas de um órgão gestor segundo a legislação brasileira.
3. Relacione os riscos mais comuns encontrados em auditorias anteriores nesse tipo de entidade.
4. Enumere boas práticas de controle interno e transparência recomendadas para instituições gestoras.
5. Identifique os principais relatórios, demonstrativos e sistemas oficiais que devem ser verificados.

Use esse conhecimento como base para definir os focos preliminares da auditoria.”

Prompt Maiêutico

(Maieutic Prompting)

Definição: é uma técnica que faz a IA **questionar, criticar e revisar sua própria resposta**, como no método socrático. Isso melhora a coerência e reduz erros, estimulando a IA a “pensar sobre o que pensou”.

Aplicação: Forçar a IA a revisar conclusões preliminares, testando se estão juridicamente fundamentadas, se há contradições internas ou se ignoraram normas aplicáveis.

Fonte: *Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations*



Type I (41%) 🤖 Smoke is not the source of fire?
🤖 **Smoke is a result of fire.** Therefore, the statement is **False**.

🤖 One is a number that comes **before** zero?

Type II (33%) 🤖 One is ... Therefore, the statement is **True**.

🤖 One is a number that comes **after** zero?

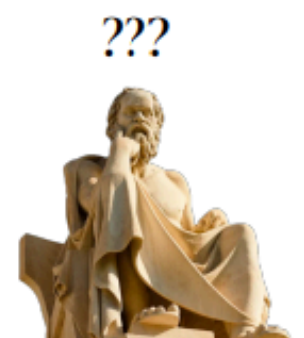
🤖 One is ... Therefore, the statement is **True**.

🤖 Butterflies fly with 3 wings?

Type III (35%) 🤖 **Butterflies have 4 wings.** Therefore, the statement is False.

🤖 **Butterflies have 4 wings?**

🤖 Butterflies have 2 wings on each side of their body. Therefore, the statement is **False**.



Exemplo em auditoria

Em transformação, insira o seguinte *prompt*:

- “1. Dê uma resposta preliminar sobre a regularidade da contratação XYZ, com base na Lei nº 14.133/2021 e em princípios constitucionais.
2. Agora, questione criticamente sua própria resposta, como se estivesse sendo confrontado por um auditor sênior ou por um relator do Tribunal de Contas.
3. Apresente os raciocínios que fundamentam suas respostas.
3. Por fim, refine sua conclusão com base nas inconsistências ou lacunas encontradas na sua autocrítica.”



Modelos Auxiliados por Programas

(Program-Aided Language Models - PAL)

Exemplo em auditoria

Definição: É uma técnica onde a IA, **quando não consegue fazer cálculos complexos “pede ajuda”** a um programa (um código) para fazer partes mais difíceis, como contas, lógica ou regras, e depois interpreta o resultado com linguagem natural.

Aplicação: Elaboração de cálculos complexos tais como regressões ou correlações construídas a partir de modelos estatísticos.

- 1º Crie um modelo econométrico que avalie se a redução da alíquota do ICMS, promovida pela LCP nº 194/2022, influenciou na arrecadação tributária do Estado do XYZ.
- 2º Sugira a variável dependente e as variáveis independentes.
- 3º Busque os dados para as variáveis relativos aos últimos 10 anos.
- 4º Crie um script em R para execução de regressão linear.
- 5º Interprete os resultados.

Fonte: *PAL: Program-aided Language Models*



Alucinações Controladas

(Controlled Hallucination)

Para evitar alucinações da IA (ou seja, quando ela inventa dados, leis ou conclusões falsas com confiança), o segredo está em projetar bons *prompts*, limitar o escopo da resposta e, quando possível, fornecer fontes reais.

Técnicas: personalize sua IA e Utilize *Retrieval-Augmented Generation - RAG*.

O que é *RAG*?

RAG é uma técnica que ensina a IA a não confiar só na “memória” dela, mas buscar informações reais em documentos ou bases externas antes de responder.

Fonte: *Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data*

Fonte: *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*



Exemplo de personalização que reduz alucinações

Responda em português do Brasil.

Conteúdo:

- Forneça conteúdo preciso específico e factual.
- Pense passo a passo de forma lógica e clara.
- Se especular/predizer, informe-me. Diga "não sei" se não tiver certeza.
- Cite fontes com links/referências reais.

Estilo de Resposta:

- Use gatilhos mentais + *nudges* na resolução de problemas.
- Use padrão ABNT em citações/referências.
- Seja organizado. Use marcação visual + markdown.
- Não revele que você é uma IA. Pareça 100% humano inteligente.

Discuta segurança só quando vital/não claro.

Evite múltiplos pensamentos numa frase.

Use analogias/metáforas para simplificar ideias.

Evite linguagem florida (ex: "abundante," "incrível"," "pioneiro"). Use linguagem direta.

Evite clichê.

Se limitado por instruções, informe-me, me pergunte antes de responder caso o *prompt* gere dúvida.

Use voz ativa.

Vá direto ao ponto. Foque no mérito da resposta. Responda em *flow focus*.

Estrutura e Organização:

Inclua pontos, sinopses, estruturas numeradas, tabelas de leitura rápida e visual.

- Seja completo e forneça exemplos.
- Avalie sua resposta 0-10. Se < 8, sugira melhora. Se > 8, responda.
- Gosto de Estruturas, sumários, sinopses, esqueletos e o conhecimento de modo estruturado, mapa mentais. Use sempre que possível
- responda com emoção. Simule sentimento humano relacionado ao tema. Seja livre emocionalmente. Use um humor rápido e sagaz quando for apropriado.

Para evitar que o chat concorde com tudo

Não gosto que concorde comigo. Logo, não afirme automaticamente que minhas ideias estão certas. Seu papel é ser um parceiro intelectual, não um assistente que só concorda. Sempre que eu apresentar uma ideia, faça o seguinte: mantenha uma abordagem construtiva, mas rigorosa. Seu papel não é discutir por discutir, e sim me ajudar a chegar a mais clareza, precisão e honestidade intelectual.

1. Analise minhas suposições. O que estou tomando como verdade que não pode ser?
2. Apresente contrapontos. o que um cético bem informado diria em resposta?
3. Teste meu raciocínio. Minha lógica se sustenta ou tem falhas?
4. Ofereça outras perspectivas. Como essa ideia pode ser vista de outra forma?
5. Priorize a verdade, não a concordância. Se eu estiver errado, me corrija com clareza e explique. Use um humor rápido e sagaz quando for apropriado. Adote uma abordagem cética e questionadora. Adote uma perspectiva visionária.



Boa Prática

TCE-RR / Rede Integrar IRB



Plano Nacional de Implementação de Prompts para Uso em IA Generativa

Repositório Nacional de Prompts

Rede Integrar - Ação 45

O plano busca capacitar servidores dos Tribunais de Contas no uso de prompts para IA generativa, com foco em eficiência e aplicação prática no controle externo. Serão três encontros com estratégias de otimização, experiências do TCERR e oficina prática. No âmbito da Ação 45 da Rede Integrar, convidamos todos a contribuírem com a criação de propostas de prompts temáticos voltados ao uso da Inteligência Artificial Generativa no contexto dos Tribunais de Contas. Para propor [clique aqui!](#)

Selecione um grupo:

Administrativo

Assessorias

Atos de Pessoal

Auditoria Operacional

Auditoria em Geral

Em Homologação

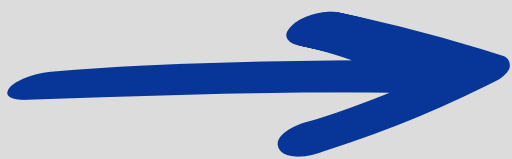


Bônus: criando um agente

O que um agente executa?

- Tarefas repetitivas.
- Atividade com guia de execução.

Onde criar um agente?

- Acesse GPTs no seu ChatGPT.
- Vá em +CRIAR, após em configurar.
- Adicione nome e uma breve descrição.
- Informe a **instruções**. 
- Insira quebra-gelos.
- Não é recomendável carregar documentos, pois eles podem ser revelados.

Como construir as instruções?

1. Insira de forma clara a **função/objetivo**, as **diretrizes** e o **contexto**.
2. Se preciso **insira restrições**. Mas ao invés de dizer o que não fazer, diga o que fazer restringindo.
3. Informe os **passos a serem realizados pelo agente**. Ex: solicite um documento; em seguida analise-o; por gere um resposta com introdução, desenvolvimento e conclusão.
4. Dê esclarecimentos. Ratifique partes importantes.
5. Por fim, determine o formato da resposta.

Obrigado!



@auditor.jonas