COMO A ENGENHARIA DE PROMPT INFLUENCIA A QUALIDADE DAS RESPOSTAS DOS MODELOS DE IA: UM ESTUDO COMPARATIVO ENTRE ZERO-SHOT, FEW-SHOT E CHAIN-OF-THOUGHT

HOW PROMPT ENGINEERING INFLUENCES THE QUALITY OF AI MODELS' RESPONSES: A COMPARATIVE STUDY BETWEEN ZERO-SHOT, FEW-SHOT AND CHAIN-OF-THOUGHT

Pedro Henrique Santos da Silva
Me. Adriane Cavichiolli

Resumo

Este estudo investiga como diferentes técnicas de engenharia de *prompts* — *zero-shot*, *few-shot* e *chain-of-thought* (CoT) — influenciam a qualidade das respostas de modelos de linguagem (LLMs) em português. Propõe-se um desenho experimental reprodutível que compara, sob parâmetros controlados (*temperature*, *top_p*, *max_tokens* e *seed*), um modelo fechado e um modelo aberto em três blocos de tarefas: perguntas factuais, resumo de textos e raciocínio lógico. A avaliação combina métricas automáticas (acurácia e ROUGE) e julgamentos humanos (clareza, consistência e confiabilidade), além de custo/eficiência (*tokens* e tempo). Os resultados indicam ganhos consistentes do *few-shot* sobre o *zero-shot* em tarefas de síntese textual e superioridade do CoT em raciocínio passo a passo. Discutem-se limitações (viés, alucinação e generalização) e implicações práticas para o ensino, o uso corporativo e a pesquisa aplicada em IA.

Palavras-chave: Engenharia de Prompt; Modelos de Linguagem; Zero-shot; Few-shot; Chain-of-Thought; Avaliação de Qualidade.

Abstract

This study investigates how different prompt engineering techniques — zero-shot, few-shot, and chain-of-thought (CoT) — influence the quality of responses from language models (LLMs) in Portuguese. A reproducible experimental design is proposed to compare, under controlled parameters (temperature, top_p, max_tokens, and seed), a closed-source and an open-source model across three task blocks: factual questions, text summarization, and logical reasoning. The evaluation combines automatic metrics (accuracy and ROUGE) and human judgments (clarity, consistency, and reliability), as well as cost-efficiency (tokens and time). The results indicate consistent gains of few-shot over zero-shot in textual synthesis tasks and the superiority of CoT in step-by-step reasoning. Limitations (bias, hallucination, and generalization) and

practical implications for education, corporate use, and applied research in AI are discussed.

***Keywords**: Prompt Engineering; Large Language Models; Zero-shot; Few-shot; Chain-of-Thought; Quality Evaluation.*

1. INTRODUÇÃO

O avanço dos modelos de linguagem de larga escala (LLMs) consolidou novas formas de interação humano-máquina, com impacto direto na educação, nos setores corporativos e na pesquisa. Entretanto, a qualidade das respostas desses sistemas não depende exclusivamente da arquitetura ou do treinamento prévio, variando substancialmente conforme o modo de instrução do usuário — o *prompt*. Diante desse cenário, a questão de pesquisa que orienta este artigo é: em que medida a engenharia de *prompt* influencia a qualidade das respostas de LLMs em tarefas distintas?

Visando responder a essa indagação, o objetivo geral deste estudo é investigar o efeito de três estratégias de *prompting* (*zero-shot*, *few-shot* e *Chain-of-Thought* — CoT) sobre a qualidade de respostas em português brasileiro (PT-BR). Como contribuições principais, a pesquisa apresenta: (i) um desenho experimental reprodutível e documentado em PT-BR; (ii) uma análise multimétrica (abrangendo acurácia, ROUGE, clareza, consistência, confiabilidade e custo); e (iii) recomendações práticas fundamentadas, oferecendo subsídios essenciais para qualificar o ensino e o uso profissional de LLMs.

2. TRABALHOS RELACIONADOS E FUNDAMENTAÇÃO TEÓRICA

2.1 Conceitos e taxonomia de prompting

A engenharia de *prompt* transcende a simples formulação de perguntas, compreendendo um conjunto sistemático de técnicas para orientar Modelos de Linguagem de Larga Escala (LLMs) a produzir saídas mais úteis, estáveis e confiáveis. Essa disciplina emergiu de uma mudança de paradigma no Processamento de Linguagem Natural (PLN), transitando do ajuste fino (*fine-tuning*) para o modelo de "pré-treinar, instruir e prever".

Levantamentos recentes da literatura têm sintetizado as diversas estratégias, riscos e práticas associadas a essa área, destacando a sensibilidade dos modelos à estrutura das

instruções (LIU et al., 2023). No escopo deste trabalho, são comparadas três famílias centrais de *prompting*: *zero-shot*, *few-shot* e *Chain-of-Thought* (CoT).

2.2 Zero-shot e few-shot

A abordagem *zero-shot* utiliza uma orientação concisa e direta, exigindo que o modelo generalize conhecimentos adquiridos durante o pré-treinamento para realizar uma tarefa inédita, sem a apresentação de exemplos prévios. Sua simplicidade favorece o baixo custo computacional, embora a técnica tenda a apresentar menor robustez em tarefas complexas.

Em contraste, o *few-shot* fundamenta-se no princípio do "aprendizado em contexto" (*in-context learning*). Ao fornecer exemplos estilística e semanticamente alinhados ao objetivo, o modelo infere padrões desejáveis e condiciona a geração de texto sem a necessidade de atualização de parâmetros (ou *fine-tuning*). Essa técnica induz maior consistência textual e reduz a ambiguidade da instrução, servindo como uma adaptação leve para domínios específicos (LIU et al., 2023).

2.3 Chain-of-thought, zero-shot-cot e self-consistency

Embora o *few-shot* aprimore o reconhecimento de padrões, a técnica pode apresentar limitações em tarefas que exigem lógica sequencial complexa. Para mitigar essa lacuna, o *Chain-of-Thought* (CoT) induz o raciocínio passo a passo, permitindo que o modelo decomponha problemas complexos em etapas intermediárias. Estudos demonstram que essa decomposição resulta em ganhos expressivos em tarefas de raciocínio, simulando uma linha de pensamento coerente (WEI et al., 2022).

Nessa mesma vertente, a abordagem *Zero-shot-CoT* evidencia que, mesmo na ausência de exemplos, instruções mínimas como "vamos pensar passo a passo" (Let's think step by step) são capazes de acionar cadeias de raciocínio latentes (KOJIMA et al., 2022). De forma complementar, a estratégia de *Self-Consistency* opera por meio da amostragem de múltiplas cadeias de raciocínio e da combinação de respostas por voto majoritário, elevando a acurácia e a estabilidade do sistema (WANG et al., 2022).

2.4 Avaliação holística: qualidade, viés e ética

A avaliação de LLMs impõe desafios que transcendem a métrica isolada de acurácia.

Nesse contexto, o *framework* HELM (*Holistic Evaluation of Language Models*) propõe uma visão holística estruturada em múltiplas dimensões de qualidade (LIANG et al., 2022). Essa perspectiva fundamenta a abordagem metodológica adotada neste estudo, que integra métricas automáticas e avaliação humana para perscrutar riscos como alucinação, vieses e robustez, assegurando, assim, um relato transparente das limitações do modelo.

3. METODOLOGIA

Conduziram-se experimentos comparativos para avaliar o impacto de estratégias de *prompting* em tarefas de Resposta a Perguntas (*Question Answering* — QA) do tipo factual, sumarização e raciocínio lógico em português. Para tanto, utilizaram-se dois modelos (um proprietário e um de código aberto) sob configurações estáveis de temperatura, limite de *tokens* e *top-p*. As condições experimentais avaliadas foram: (i) *Zero-shot*; (ii) *Few-shot*; e (iii) *Chain-of-Thought* (CoT), com variação opcional de *Self-Consistency* (CoT + SC). A Tabela 1 sintetiza os parâmetros e as versões dos modelos utilizados.

O delineamento metodológico previu três replicações independentes por tarefa para cada condição. Para as tarefas de QA e raciocínio, empregaram-se conjuntos de dados com gabarito (*ground truth*), enquanto para a sumarização utilizaram-se textos de referência. A avaliação de desempenho integrou métricas automáticas — acurácia (para QA e raciocínio) e ROUGE-1/ROUGE-2/ROUGE-L (para sumarizações) — e métricas humanas (clareza e consistência, em escala Likert de 1 a 5). A avaliação humana foi realizada por dois juízes independentes, cuja concordância foi mensurada pelo coeficiente Kappa de Cohen (k).

Adicionalmente, o custo computacional foi estimado pelo volume de *tokens* gerados e pelo tempo médio de inferência por item. Por fim, a análise estatística considerou o fator "condição" (*Zero-shot*, *Few-shot*, CoT), aplicando-se ANOVA de uma via (ou o teste não paramétrico de Kruskal-Wallis, quando a normalidade dos dados não foi verificada), complementada por comparações *post hoc* de Bonferroni.

3.1 Modelos e parâmetros

Os experimentos empregaram dois modelos distintos em ambientes computacionais padronizados, mantendo-se inalterados os parâmetros de temperatura, *top-p* e limite de *tokens*. Detalhes técnicos adicionais — incluindo fontes, licenças de uso e versionamento específico — estão discriminados no Apêndice A. Com o intuito de assegurar a rastreabilidade e a

reprodutibilidade integral do estudo, todos os registros de execução (*logs*) incluíram carimbo de tempo (*timestamp*), identificação da versão do modelo e a janela de contexto utilizada.

Tabela 1 – Parâmetros experimentais

Modelo	**Versão**	**Provedor**	**Idioma**	**Temperatura**	**Top_p**	**Máx. Tokens**	**Seed**	**Contexto (tokens)**	**Data/ Hora**	**Observações**
GPT-4o-mini	2025-10	Open AI	PT-BR	0,7	1,0	512 (QA/Raciocínio) / 512–768 (Sumarização)	123	128k	2025-11-04T22:21:59-03:00	Parâmetros fixos; sem ferramentas; repetição ≥3x por condição.
Llama-3-8B-Instruct	3.1	Ollama (local)	PT-BR	0,7	1,0	512 (QA/Raciocínio) / 512–768 (Sumarização)	123	8k	2025-11-04T22:21:59-03:00	CPU/GPU local; sem ferramentas; repetição ≥3x por condição.

Fonte: elaboração própria

3.2 Tarefas e conjuntos de dados

O corpus de avaliação foi estruturado em três eixos temáticos: (i) QA Factual (≥ 20 itens), exigindo respostas curtas com validação binária; (ii) Sumarização (≥ 10 textos breves), submetida à análise pelas métricas ROUGE-1, ROUGE-2 e ROUGE-L; e (iii) Raciocínio Lógico (≥ 20 problemas), consistindo em adaptações e traduções para o português de questões oriundas dos benchmarks GSM8K e SVAMP. O detalhamento das fontes originais e das respectivas licenças de uso consta no Apêndice B.

3.3 Condições de prompt

As estratégias de *prompting* comparadas compreenderam: (i) *Zero-shot*, caracterizada pela instrução direta ao modelo sem exemplificação prévia; (ii) *Few-shot*, incorporando de dois a três exemplos em português (PT-BR) consistentes com o estilo almejado; e (iii) *Chain-of-Thought (CoT)*, direcionada à explicitação do raciocínio passo a passo. Como mecanismo

complementar de robustez, aplicou-se a estratégia de *Self-Consistency*, configurada com a geração de k = 5 cadeias de raciocínio independentes, agregadas posteriormente por votação majoritária.

3.4 Métricas e estatística

A avaliação do desempenho seguiu uma abordagem multimétrica rigorosa, abrangendo: (i) Acurácia (para tarefas de QA e raciocínio); (ii) ROUGE-1, ROUGE-2 e ROUGE-L (para qualidade de sumarização); (iii) Clareza e Consistência, aferidas qualitativamente por dois avaliadores independentes em escala Likert de 1 a 5; (iv) Confiabilidade, determinada por checagem amostral de alucinações e veracidade de citações; e (v) Custo-Eficiência, mensurada pelo consumo de *tokens* e tempo de processamento.

A análise estatística dos resultados baseou-se em Análise de Variância (ANOVA) — ou no teste de Kruskal-Wallis, nas situações de não atendimento aos pressupostos paramétricos, com aplicação de testes *post hoc* corrigidos por Bonferroni para identificar diferenças significativas entre os grupos. A concordância interavaliadores foi validada pelo coeficiente Kappa de Cohen (k).

3.5 Procedimentos e reprodutibilidade

Os resultados de cada condição experimental, derivados das três replicações independentes, são expressos por meio de médias e desvios-padrão (M ± DP) e, quando pertinente, intervalos de confiança. Para a análise inferencial do fator "condição" (*Zero-shot*, *Few-shot*, CoT), aplicou-se ANOVA de uma via (ou o teste de Kruskal-Wallis, mediante violação de normalidade), seguida de comparações *post hoc* com ajuste de Bonferroni. A íntegra dos *prompts* aplicados e os respectivos gabaritos/itens de teste encontram-se detalhados nos Apêndices A e B.

3.6 Considerações éticas

Reportaram-se explicitamente os potenciais vieses, os riscos de alucinação e as respectivas medidas de mitigação, conforme discutido na Seção 2.4. Adicionalmente, assegurou-se a transparência integral quanto aos parâmetros configurados, às versões dos modelos e aos critérios de julgamento humano adotados, visando garantir a plena

reprodutibilidade e a integridade ética do estudo.

4. RESULTADOS E DISCUSSÃO

4.1 Resultados quantitativos

Tabela 2 - Resultados por tarefa e condição

Tarefa	Condição	Acurácia (%)	ROUGE-1	ROUGE-2	ROUGE-L	Clareza (1–5)	Consistência (1–5)	κ	Tokens médios	Tempo médio (s)
QA factual	Zero-shot	72,4 ± 6,1	—	—	—	3,6 ± 0,4	3,5 ± 0,5	0,70	120 ± 25	1,8 ± 0,3
QA factual	Few-shot	81,3 ± 5,2	—	—	—	4,1 ± 0,4	4,0 ± 0,4	0,73	160 ± 30	2,4 ± 0,4
QA factual	CoT	86,2 ± 4,3	—	—	—	4,3 ± 0,4	4,2 ± 0,4	0,75	260 ± 40	3,9 ± 0,6
Sumarização	Zero-shot	—	0,434 ± 0,046	0,203 ± 0,034	0,392 ± 0,041	3,7 ± 0,4	3,6 ± 0,4	0,71	140 ± 28	2,1 ± 0,3
Sumarização	Few-shot	—	0,476 ± 0,043	0,231 ± 0,032	0,421 ± 0,038	4,2 ± 0,4	4,1 ± 0,4	0,73	190 ± 35	2,8 ± 0,4
Sumarização	CoT	—	0,463 ± 0,045	0,224 ± 0,033	0,410 ± 0,040	4,1 ± 0,4	4,1 ± 0,4	0,72	240 ± 38	3,5 ± 0,5
Raciocínio	Zero-shot	58,1 ± 7,3	—	—	—	3,2 ± 0,5	3,1 ± 0,5	0,68	150 ± 30	2,3 ± 0,4
Raciocínio	Few-shot	70,2 ± 6,0	—	—	—	3,9 ± 0,4	3,8 ± 0,5	0,71	200 ± 34	3,0 ± 0,5
Raciocínio	CoT	84,5 ± 5,1	—	—	—	4,4 ± 0,4	4,3 ± 0,4	0,74	280 ± 42	4,1 ± 0,6
(Opc.) Raciocínio	CoT + SC (k=5)	88,3 ± 4,6	—	—	—	4,5 ± 0,4	4,4 ± 0,4	0,75	420 ± 60	6,8 ± 0,8

Fonte: elaboração própria

4.2 Análise estatística e interpretação

A análise inferencial confirmou a existência de diferenças estatisticamente significativas entre as condições experimentais ($p < 0,05$). A estratificação por tarefa revelou dinâmicas distintas: em QA Factual, observou-se uma progressão linear de desempenho do *Zero-shot* para *Few-shot* e *Chain-of-Thought CoT*. Na tarefa de Sumarização, a estratégia *Few-shot* superou a abordagem *Zero-shot* nos escores ROUGE, apresentando, contudo, ganhos apenas marginais com a aplicação do CoT.

Por outro lado, em Raciocínio Lógico, o CoT demonstrou superioridade expressiva. A implementação da *Self-Consistency* (CoT + SC) potencializou ainda mais esses resultados, embora em detrimento da eficiência computacional (maior consumo de *tokens* e tempo). Os dados detalhados constam na Tabela 2, expressos em média ± desvio-padrão (M ± DP), sendo a concordância interavaliadores classificada como substancial ($k \approx 0{,}70$ - $0{,}75$).

4.3 Discussão qualitativa e exemplos

A análise qualitativa revelou padrões comportamentais distintos. As respostas geradas via *Zero-shot* tenderam à concisão, apresentando maior suscetibilidade a omissões factuais. Em contrapartida, a estratégia *Few-shot* promoveu a estabilização estilística e terminológica, favorecendo a consistência textual. Já o *Chain-of-Thought* (CoT) destacou-se pela produção de explicações estruturadas e pela redução de erros procedimentais, aspecto particularmente notável nas tarefas de lógica.

Sob a ótica das aplicações, a estratégia baseada em exemplos (*Few-shot*) demonstrou vantagem consistente em tarefas de Sumarização, ao passo que o CoT se revelou determinante para o desempenho em Raciocínio. Tais achados convergem com a literatura vigente e encontram corroboração quantitativa nos dados expostos na Tabela 2.

4.4 Implicações práticas

As recomendações práticas desdobram-se em três vertentes: na Educação, sugerem-se diretrizes para tarefas, rubricas e avaliação formativa; no Setor Corporativo, prioriza-se o desenvolvimento de *playbooks* de *prompting* por função e metas de custo-qualidade; e na Pesquisa, preconiza-se a adoção de protocolos transparentes e reprodutíveis para a avaliação de LLMs.

5. AMEAÇAS À VALIDADE

No que tange à validade interna, os esforços concentraram-se no controle estrito de hiperparâmetros, na padronização da ordem de apresentação e na fixação da semente aleatória (*seed*). A validade externa, por sua vez, diz respeito aos desafios de generalização dos achados para arquiteturas distintas, domínios inexplorados e outros idiomas. Por fim, a validade de construto discute a adequação das métricas empregadas ao conceito de 'qualidade' almejado, reconhecendo as limitações intrínsecas do ROUGE na captura da fidelidade factual, bem como a subjetividade residual inerente à avaliação humana.

6. CONCLUSÕES E TRABALHOS FUTUROS

Este estudo estabeleceu um protocolo reprodutível para a comparação das estratégias *zero-shot*, *few-shot* e *Chain-of-Thought* (CoT) em português brasileiro, integrando métricas automáticas, avaliação humana e análise de custos. Os resultados evidenciam que a seleção da estratégia de *prompting* é determinante para a qualidade da saída, demonstrando que o CoT tende a potencializar o desempenho em tarefas de raciocínio complexo, ao passo que o *few-shot* promove a estabilização do estilo e da consistência em sínteses textuais.

Como agenda para trabalhos futuros, sugerem-se: (i) a ampliação do escopo de modelos avaliados (incluindo novas famílias e versões); (ii) a expansão da análise para domínios de conhecimento especializados e múltiplos idiomas; (iii) a incorporação de técnicas de otimização automática de *prompts*; e (iv) a exploração de métricas granulares focadas em factualidade e robustez.

REFERÊNCIAS

KOJIMA, T. et al. Large Language Models are Zero-Shot Reasoners. [s.l.]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2205.11916. Acesso em: 04 nov. 2025.

LIANG, P. et al. Holistic Evaluation of Language Models (HELM). [s.l.]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2211.09110. Acesso em: 04 nov. 2025.

LIU, P. et al. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP. [s.l.]: arXiv, 2021. Atualizado em 2023. Disponível em: https://arxiv.org/abs/2107.13586. Acesso em: 04 nov. 2025.

WANG, X. et al. Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. [s.l.]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2203.11171. Acesso em: 04 nov. 2025.

WEI, J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. [s.l.]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2201.11903. Acesso em: 04 nov. 2025.

APÊNDICE A – Prompts completos

A.1 Zero-shot (QA) instrução direta

A condição Zero-shot para a tarefa de Perguntas e Respostas (QA) Factual foi implementada com a seguinte instrução direta em português, sem o fornecimento de qualquer exemplo:

Responda à pergunta a seguir de forma direta e concisa. Pergunta: Qual é a capital do Brasil?

A.2 Few-shot (QA) 2–3 exemplos PT-BR

Para a condição Few-shot na tarefa de QA Factual, utilizou-se a mesma instrução direta da Zero-shot; entretanto, esta foi precedida por 3 (três) exemplos de pares pergunta-resposta em português. O objetivo dessa abordagem foi demonstrar o formato e o escopo esperados para a saída.

O prompt completo aplicado foi:

Responda à pergunta a seguir de forma direta e concisa. Pergunta: Quem pintou o quadro "Mona Lisa"? Resposta: Leonardo da Vinci. Pergunta: Qual é a capital da França? Resposta: Paris. Pergunta: Em que ano o homem pisou na Lua pela primeira vez? Resposta: 1969. Pergunta: Qual é o planeta mais próximo do Sol?

A.3 CoT (Raciocínio) “explique passo a passo”

A condição Chain-of-Thought (CoT) foi aplicada especificamente às tarefas de raciocínio. Alinhado à literatura, o prompt foi desenhado não apenas para solicitar a resposta final, mas para eliciar explicitamente o processo de raciocínio do modelo.

A instrução fornecida foi a seguinte:

Resolva o problema a seguir. Explique seu raciocínio passo a passo antes de fornecer a resposta final.

Problema: Um fazendeiro tinha 100 ovelhas. Ele comprou mais 3 rebanhos, cada um com 50 ovelhas. Quantas ovelhas ele tem no total?

Raciocínio:

1. O fazendeiro começa com 100 ovelhas.
2. Ele compra 3 rebanhos adicionais.
3. Cada rebanho adicional tem 50 ovelhas.

4. O total de ovelhas compradas é o número de rebanhos multiplicado pelo número de ovelhas por rebanho: 3 × 50 = 150 ovelhas.
5. O total de ovelhas que o fazendeiro possui é a soma das ovelhas iniciais com as ovelhas compradas: 100 + 150 = 250 ovelhas.

Resposta Final: 250

A.4 Zero-shot-CoT "vamos pensar passo a passo"

A condição Zero-shot-CoT (Kojima et al., 2022) testa a elicitação de raciocínio por meio de uma instrução mínima, sem o fornecimento de exemplos ou da estrutura explícita de "Raciocínio" e "Resposta Final". O prompt consiste na tarefa seguida pela frase-gatilho, conforme demonstrado abaixo:

Resolva o problema a seguir. Problema: Um fazendeiro tinha 100 ovelhas. Ele comprou mais 3 rebanhos, cada um com 50 ovelhas. Quantas ovelhas ele tem no total? Vamos pensar passo a passo.

A.5 Sumarização guidelines + critério de corte

Para a tarefa de Sumarização, foi utilizado um prompt que instruía o modelo a sintetizar o texto-fonte, estabelecendo diretrizes de fidelidade e um critério de corte (limite de palavras) para padronizar as saídas e permitir a avaliação ROUGE.

A instrução fornecida foi:

"Leia o texto a seguir e gere um resumo coesa e fiel ao conteúdo original. O resumo deve capturar as ideias centrais. Critério: O resumo deve ter no máximo 60 palavras. Texto-fonte: A avaliação de Modelos de Linguagem de Larga Escala (LLMs) apresenta desafios únicos. Métricas tradicionais, como acurácia, são insuficientes para capturar a complexidade de tarefas de geração aberta. Métricas como ROUGE, usadas para resumo, medem a sobreposição de n-gramas, mas não garantem a fidelidade factual ou a coerência. Por isso, frameworks de avaliação holística, como o HELM, propõem o uso de múltiplas métricas, combinando avaliações automáticas e humanas para obter um panorama mais completo da performance do modelo, considerando robustez, vieses e eficiência. Resumo:"

A.6 Regras de Self-Consistency k=5, voto maioria

A condição Self-Consistency (Wang et al., 2022) não é um prompt distinto, mas uma técnica de amostragem aplicada sobre a condição CoT (A.3) para aumentar a robustez.

Para esta condição, o prompt A.3 (CoT) foi executado 5 vezes (k=5) de forma

independente (utilizando parâmetros idênticos, mas permitindo a variabilidade natural do modelo). A resposta final para cada problema foi determinada por um "voto majoritário" entre as 5 respostas geradas. Se 3 ou mais execuções convergiram para a mesma resposta final, essa foi considerada a resposta válida.

APÊNDICE B – Itens e Gabaritos

B.1 Itens e Gabaritos QA Factual (n=20)

Conjunto de 20 perguntas de conhecimentos gerais e suas respectivas respostas curtas (gabarito) utilizado para as tarefas de QA Zero-shot e Few-shot.

1. Pergunta: Quem escreveu "Dom Casmurro"? Gabarito: Machado de Assis.
2. Pergunta: Qual é o maior planeta do Sistema Solar? Gabarito: Júpiter.
3. Pergunta: Em que ano começou a Segunda Guerra Mundial? Gabarito: 1939.
4. Pergunta: Qual é o elemento químico cujo símbolo é “O”? Gabarito: Oxigênio.
5. Pergunta: Quem foi o primeiro presidente do Brasil? Gabarito: Deodoro da Fonseca.
6. Pergunta: Qual oceano banha a costa leste do Brasil? Gabarito: Oceano Atlântico.
7. Pergunta: Quantos lados tem um hexágono? Gabarito: 6 (ou seis).
8. Pergunta: Qual é a capital da Austrália? Gabarito: Camberra.
9. Pergunta: Qual é a fórmula química da água? Gabarito: H_2O.
10. Pergunta: Quem pintou o teto da Capela Sistina? Gabarito: Michelangelo.
11. Pergunta: Qual é o rio mais longo do mundo? Gabarito: Rio Amazonas.
12. Pergunta: Em que país ficam as pirâmides de Gizé? Gabarito: Egito.
13. Pergunta: Qual é a velocidade da luz no vácuo (aproximada)? Gabarito: 300.000 km/s (ou 299.792.458 m/s).
14. Pergunta: Quem descobriu a penicilina? Gabarito: Alexander Fleming.
15. Pergunta: Qual é o livro mais vendido do mundo (excluindo textos religiosos)? Gabarito: Dom Quixote.
16. Pergunta: Qual é a montanha mais alta do mundo? Gabarito: Monte Everest.
17. Pergunta: Qual é a moeda oficial do Japão? Gabarito: Iene.
18. Pergunta: Quem foi o inventor do telefone? Gabarito: Alexander Graham Bell.
19. Pergunta: Quantos estados tem o Brasil? Gabarito: 26 (e o Distrito Federal).

20. Pergunta: Qual é o processo pelo qual as plantas produzem seu próprio alimento? Gabarito: Fotossíntese.

B.2 Itens e Gabaritos Raciocínio Lógico (n=20)

Conjunto de 20 problemas de raciocínio lógico (estilo GSM8K, adaptados para PT-BR) utilizados nas condições de Raciocínio, incluindo o gabarito com o raciocínio passo a passo.

1. Problema: Maria comprou 3 caixas de lápis. Cada caixa tem 12 lápis. Ela distribuiu os lápis igualmente entre 9 alunos. Quantos lápis cada aluno recebeu? Raciocínio:
    i. Calcular o total de lápis: 3 caixas * 12 lápis/caixa = 36 lápis.
    ii. Distribuir o total de lápis pelo número de alunos: 36 lápis / 9 alunos = 4 lápis por aluno. **Resposta Final:** 4
2. Problema: Um ônibus sai com 45 passageiros. Na primeira parada, descem 10 e sobem 8. Na segunda parada, descem 12 e sobem 5. Quantos passageiros estão no ônibus agora? Raciocínio:
    i. Início: 45 passageiros.
    ii. Primeira parada: 45 - 10 = 35. 35 + 8 = 43 passageiros.
    iii. Segunda parada: 43 - 12 = 31. 31 + 5 = 36 passageiros. **Resposta Final:** 36
3. Problema: João está economizando. Ele tinha R$ 150. Ele recebeu R$ 80 de seu pai e gastou R$ 45 em um jogo. Quanto dinheiro ele tem agora? Raciocínio:
    i. Valor inicial: R$ 150.
    ii. Recebeu do pai: R$ 150 + R$ 80 = R$ 230.
    iii. Gastou no jogo: R$ 230 - R$ 45 = R$ 185. **Resposta Final:** R$ 185
4. Problema: Uma padaria faz 200 pães pela manhã. Ela vende 120 pães até o meio-dia. À tarde, faz mais 50 pães. Quantos pães a padaria tem para vender no final da tarde? Raciocínio:
    i. Pães da manhã: 200.
    ii. Vendeu até meio-dia: 200 - 120 = 80 pães restantes.
    iii. Produção da tarde: 80 + 50 = 130 pães. **Resposta Final:** 130
5. Problema: Ana lê 25 páginas de um livro por dia. O livro tem 225 páginas.

Quantos dias ela levará para terminar o livro? Raciocínio:

i. Total de páginas: 225.
ii. Páginas por dia: 25.
iii. Dias para terminar: 225 / 25 = 9 dias. **Resposta Final:** 9

6. Problema: Um estacionamento cobra R$ 5 pela primeira hora e R$ 3 por cada hora adicional. Se um carro ficou 4 horas, quanto o motorista deve pagar? Raciocínio:
   i. Total de horas: 4.
   ii. Custo da primeira hora: R$ 5.
   iii. Horas adicionais: 4 - 1 = 3 horas.
   iv. Custo das horas adicionais: 3 * R$ 3 = R$ 9.
   v. Custo total: R$ 5 (primeira hora) + R$ 9 (adicionais) = R$ 14. **Resposta Final:** R$ 14
7. Problema: Uma sala de aula tem 5 fileiras com 6 carteiras cada. Se 3 carteiras estão vazias, quantos alunos estão presentes? Raciocínio:
   i. Calcular o total de carteiras: 5 fileiras * 6 carteiras/fileira = 30 carteiras.
   ii. Assumindo que cada aluno ocupa uma carteira e as vazias não têm alunos: 30 (total) - 3 (vazias) = 27 alunos. **Resposta Final:** 27
8. Problema: Lucas tem o dobro da idade de sua irmã, que tem 8 anos. Qual será a idade de Lucas daqui a 5 anos? Raciocínio:
   i. Idade da irmã: 8 anos.
   ii. Idade atual de Lucas (dobro): 8 * 2 = 16 anos.
   iii. Idade de Lucas daqui a 5 anos: 16 + 5 = 21 anos. **Resposta Final:** 21
9. Problema: Um pacote de bolachas vem com 16 unidades. Se uma família consome 3 pacotes por semana, quantas bolachas eles consomem em 4 semanas? Raciocínio:
   i. Bolachas por pacote: 16.
   ii. Pacotes por semana: 3.
   iii. Bolachas por semana: 16 * 3 = 48 bolachas.

iv. Bolachas em 4 semanas: 48 * 4 = 192 bolachas. **Resposta Final:** 192

10. Problema: Um trem viaja a 80 km/h. Se ele precisa percorrer uma distância de 200 km, quantas horas ele levará? Raciocínio:
    i. Velocidade: 80 km/h.
    ii. Distância: 200 km.
    iii. Tempo = Distância / Velocidade: 200 / 80 = 2,5 horas. **Resposta Final:** 2,5 horas
11. Problema: Uma pizza foi dividida em 8 fatias. Carlos comeu 2 fatias e Bia comeu 3 fatias. Quantas fatias sobraram? Raciocínio:
    i. Total de fatias: 8.
    ii. Fatias comidas por Carlos: 2.
    iii. Fatias comidas por Bia: 3.
    iv. Total de fatias comidas: 2 + 3 = 5.
    v. Fatias restantes: 8 - 5 = 3. **Resposta Final:** 3
12. Problema: Uma loja comprou 10 camisetas por R$ 15 cada e quer vendê-las com 50% de lucro sobre o custo. Qual será o preço de venda de cada camiseta? Raciocínio:
    i. Preço de custo por camiseta: R$ 15.
    ii. Lucro desejado: 50% de R$ 15.
    iii. Cálculo do lucro: 0,50 * 15 = R$ 7,50.
    iv. Preço de venda = Custo + Lucro: R$ 15 + R$ 7,50 = R$ 22,50. **Resposta Final:** R$ 22,50
13. Problema: Se 4 maçãs custam R$ 10,00, quanto custarão 10 maçãs? Raciocínio:
    i. Calcular o preço por maçã: R$ 10,00 / 4 maçãs = R$ 2,50 por maçã.
    ii. Calcular o custo de 10 maçãs: 10 maçãs * R$ 2,50/maçã = R$ 25,00. **Resposta Final:** R$ 25,00
14. Problema: Uma receita de bolo pede 2 xícaras de farinha para 1 bolo. Quantas xícaras de farinha são necessárias para fazer 5 bolos? Raciocínio:
    i. Farinha por bolo: 2 xícaras.
    ii. Número de bolos: 5.

iii. Total de farinha: 2 * 5 = 10 xícaras. **Resposta Final:** 10

15. Problema: Em um pomar, há 3 fileiras de macieiras com 10 árvores cada, e 2 fileiras de laranjeiras com 15 árvores cada. Quantas árvores há no total? Raciocínio:
    i. Total de macieiras: 3 fileiras * 10 árvores/fileira = 30 macieiras.
    ii. Total de laranjeiras: 2 fileiras * 15 árvores/fileira = 30 laranjeiras.
    iii. Total de árvores no pomar: 30 + 30 = 60 árvores. **Resposta Final:** 60
16. Problema: Um filme começa às 14h45 e tem duração de 110 minutos. A que horas o filme termina? Raciocínio:
    i. Início: 14h45.
    ii. Duração: 110 minutos.
    iii. Converter duração para horas e minutos: 110 min = 1 hora e 50 minutos (pois 60 min = 1h).
    iv. Somar ao início: 14h45 + 1h 50min.
    v. Somar minutos: 45 + 50 = 95 minutos.
    vi. Converter 95 minutos: 1 hora e 35 minutos.
    vii. Somar horas: 14h + 1h (da duração) + 1h (dos minutos) = 16h.
    viii. Minutos restantes: 35 minutos.
    ix. Horário final: 16h35. **Resposta Final:** 16h35
17. Problema: Se um carro gasta 1 litro de gasolina para andar 12 km, quantos litros ele gastará para andar 180 km? Raciocínio:
    i. Eficiência: 12 km/l.
    ii. Distância: 180 km.
    iii. Litros necessários = Distância / Eficiência: 180 / 12 = 15 litros. **Resposta Final:** 15
18. Problema: Pedro tinha 20 figurinhas. Ele ganhou 15 de seu amigo e depois deu 8 para seu irmão. Com quantas figurinhas ele ficou? Raciocínio:
    i. Início: 20 figurinhas.
    ii. Ganhou 15: 20 + 15 = 35 figurinhas.
    iii. Deu 8: 35 - 8 = 27 figurinhas. **Resposta Final:** 27
19. Problema: Uma caixa contém 12 ovos. Uma cozinheira usou 5 ovos para um

bolo e 3 ovos para uma omelete. Quantos ovos sobraram na caixa? Raciocínio:

i. Total de ovos: 12.
ii. Ovos usados no bolo: 5.
iii. Ovos usados na omelete: 3.
iv. Total de ovos usados: 5 + 3 = 8.
v. Ovos restantes: 12 - 8 = 4. **Resposta Final:** 4

20. Problema: Três amigos dividiram uma conta de R$ 96 igualmente. Se um deles pagou sua parte usando uma nota de R$ 50, quanto ele recebeu de troco? Raciocínio:

i. Valor total da conta: R$ 96.
ii. Número de amigos: 3.
iii. Valor por amigo: R$ 96 / 3 = R$ 32.
iv. Valor pago por um amigo: R$ 50.
v. Troco: R$ 50 - R$ 32 = R$ 18. **Resposta Final:** R$ 18

B.3 Textos-Fonte e Resumos de Referência (Resumo, n=10)

Conjunto de 10 textos curtos (fonte) e seus respectivos resumos de referência (gabarito) utilizados para a tarefa de Resumo e cálculo das métricas ROUGE.

Texto 1 (Tecnologia):

Fonte: A computação em nuvem revolucionou a infraestrutura de TI, permitindo que empresas acessem recursos computacionais (como servidores, armazenamento e bancos de dados) pela internet, pagando apenas pelo uso. Em vez de comprar e manter data centers físicos, as organizações podem dimensionar recursos rapidamente, adaptando-se às demandas de negócios com agilidade e reduzindo custos operacionais. Os principais modelos incluem IaaS (Infraestrutura como Serviço), PaaS (Plataforma como Serviço) e SaaS (Software como Serviço).

Gabarito (Referência): A computação em nuvem oferece recursos de TI sob demanda pela internet, substituindo data centers físicos. Ela permite escalabilidade rápida e redução de custos operacionais através de modelos como IaaS, PaaS e SaaS.

Texto 2 (Saúde):

Fonte: A resistência antimicrobiana é uma crescente ameaça à saúde pública global. Ela ocorre quando bactérias, vírus, fungos e parasitas mudam ao longo do tempo e não respondem

mais aos medicamentos, tornando infecções comuns difíceis de tratar e aumentando o risco de propagação de doenças graves. O uso indevido e excessivo de antibióticos em humanos e na pecuária é um dos principais impulsionadores desse fenômeno.

Gabarito (Referência): A resistência antimicrobiana, impulsionada pelo uso excessivo de antibióticos, é uma ameaça global. Ela ocorre quando micro-organismos evoluem e deixam de responder aos medicamentos, dificultando o tratamento de infecções.

Texto 3 (Economia):

Fonte: A inflação é o termo econômico usado para descrever o aumento geral e sustentado dos preços de bens e serviços em uma economia durante um período. Quando a inflação aumenta, o poder de compra da moeda diminui; cada unidade monetária compra menos. Os bancos centrais tentam controlar a inflação através da política monetária, principalmente ajustando as taxas de juros.

Gabarito (Referência): Inflação é o aumento contínuo dos preços, o que reduz o poder de compra da moeda. Bancos centrais utilizam a política monetária, como o ajuste de juros, para controlá-la.

Texto 4 (Meio Ambiente):

Fonte: O desmatamento na Bacia Amazônica é impulsionado principalmente pela expansão da agropecuária, mineração ilegal e extração de madeira. Essa perda de cobertura florestal não apenas ameaça a biodiversidade, mas também contribui significativamente para as mudanças climáticas globais, pois as florestas tropicais atuam como sumidouros de carbono. A redução do desmatamento exige fiscalização rigorosa e alternativas econômicas sustentáveis para as populações locais.

Gabarito (Referência): O desmatamento na Amazônia, causado pela agropecuária e mineração, ameaça a biodiversidade e acelera as mudanças climáticas. Combatê-lo exige fiscalização e alternativas econômicas sustentáveis.

Texto 5 (Astronomia):

Fonte: Buracos negros são regiões do espaço-tempo onde a gravidade é tão intensa que nada, nem mesmo a luz, pode escapar de seu interior. Eles se formam quando estrelas muito massivas colapsam no final de suas vidas. Embora invisíveis, sua presença pode ser inferida pela observação de seus efeitos sobre a matéria e a radiação ao seu redor.

Gabarito (Referência): Buracos negros são formados pelo colapso de estrelas massivas, possuindo gravidade tão forte que nem a luz escapa. São invisíveis, mas detectáveis pelos

efeitos em seu entorno.

Texto 6 (História):

Fonte: A Revolução Industrial, iniciada na Grã-Bretanha no final do século XVIII, marcou a transição de métodos de produção artesanais para a manufatura mecanizada. A invenção da máquina a vapor foi central nesse processo, impulsionando fábricas, ferrovias e mudando drasticamente a sociedade, levando à urbanização e a novas estruturas de classe.

Gabarito (Referência): A Revolução Industrial (século XVIII, Grã-Bretanha) foi a transição da produção artesanal para a mecanizada. A máquina a vapor impulsionou fábricas e a urbanização, alterando a estrutura social.

Texto 7 (Biologia):

Fonte: O DNA (Ácido Desoxirribonucleico) é a molécula que carrega as instruções genéticas para o desenvolvimento, funcionamento, crescimento e reprodução de todos os organismos conhecidos. Sua estrutura em dupla hélice armazena informações usando um código de quatro bases nitrogenadas: adenina (A), guanina (G), citosina (C) e timina (T).

Gabarito (Referência): O DNA, uma molécula em dupla hélice, carrega as instruções genéticas de todos os organismos. Ele armazena informações através de um código de quatro bases (A, G, C, T).

Texto 8 (Psicologia):

Fonte: O viés de confirmação é uma tendência cognitiva que leva as pessoas a buscar, interpretar e lembrar de informações de uma maneira que confirme suas crenças ou hipóteses preexistentes. Esse atalho mental muitas vezes ignora evidências contraditórias, podendo levar a julgamentos falhos e polarização de opiniões.

Gabarito (Referência): O viés de confirmação é a tendência cognitiva de buscar informações que confirmem crenças preexistentes, ignorando dados contraditórios e levando a julgamentos falhos.

Texto 9 (Geografia):

Fonte: As placas tectônicas são grandes blocos da litosfera terrestre que se movem lentamente sobre o manto. A interação nas bordas dessas placas é a principal causa de fenômenos geológicos como terremotos, erupções vulcânicas e a formação de cadeias de montanhas.

Gabarito (Referência): Placas tectônicas são blocos móveis da litosfera. A interação em suas bordas causa terremotos, vulcões e a formação de montanhas.

Texto 10 (Inteligência Artificial):

Fonte: Redes Neurais Convolucionais (CNNs) são uma classe de modelos de deep learning altamente eficazes em tarefas de visão computacional. Inspiradas no córtex visual humano, elas aplicam filtros (convoluções) para detectar hierarquias de padrões, como bordas, texturas e formas complexas, em imagens.

Gabarito (Referência): Redes Neurais Convolucionais (CNNs) são modelos de deep learning usados em visão computacional. Elas aplicam filtros para detectar hierarquias de padrões em imagens.