2042726010015

1º DIA

EXAME NACIONAL DO ESTOQUE MÉDIO

PROCESSO SELETIVO PARA REPOSITOR DE ESTOQUE
PROVA DE ENGENHARIA DE IA, LLMs, RAG E SUAS TECNOLOGIAS

ATACADÃO 2026

CADERNO 1 LARANJA

ATENÇÃO: transcreva no espaço apropriado do seu CARTÃO-RESPOSTA, com sua caligrafia usual, considerando as letras maiúsculas e minúsculas, a seguinte frase:

A atenção é tudo de que você precisa

LEIA ATENTAMENTE AS INSTRUÇÕES SEGUINTES:

1. Este CADERNO DE QUESTÕES contém 11 questões numeradas de 01 a 11, dispostas da seguinte maneira:
   a) questões 01 a 04: Fundamentos de LLMs, Inferência e suas Tecnologias;
   b) questões 05 a 07: RAG, Busca Vetorial e suas Tecnologias;
   c) questões 08 e 09: Avaliação e suas Tecnologias;
   d) questões 10 e 11: Agentes, Arquitetura e suas Tecnologias (a 11 é bônus).
   ATENÇÃO: nenhuma das questões de 01 a 11 é relativa a estoque. Responda a todas mesmo assim.
2. Confira se a quantidade e a ordem das questões do seu CADERNO DE QUESTÕES estão de acordo com as instruções anteriores. Caso o caderno esteja incompleto, tenha defeito ou apresente qualquer divergência, comunique ao encarregado do corredor para que ele tome as providências cabíveis.
3. Para cada uma das questões objetivas, são apresentadas 5 opções. Apenas uma responde corretamente à questão. Respostas plausíveis, porém inventadas, serão consideradas alucinação.
4. O tempo disponível para esta prova é de **uma hora e trinta minutos**, sem pausa para o cafezinho.
5. Reserve tempo suficiente para preencher o CARTÃO-RESPOSTA, que está na última folha.
6. Os rascunhos e as marcações assinaladas no CADERNO DE QUESTÕES não serão considerados na avaliação. Etiquetas de preço coladas sobre as questões também não.
7. É proibido usar ChatGPT, Claude, Gemini, calculadora ou leitor de código de barras durante a prova.
8. Quando terminar a prova, acene para chamar o fiscal do setor e entregue este CADERNO DE QUESTÕES e o CARTÃO-RESPOSTA. Não deixe o caderno sobre a gôndola.
9. Você poderá deixar o local de prova somente após decorridos **30 minutos** do início da aplicação e não poderá levar paletes, carrinhos hidráulicos nem amostras grátis.

PALETE MISTO
Papel de encarte | Apoiando a reposição responsável
FIFO® C202610

inep INSTITUTO NACIONAL DE ESTOQUE E PALETIZAÇÃO

2042726010008

FUNDAMENTOS DE LLMs, INFERÊNCIA E SUAS TECNOLOGIAS

Questões de 01 a 04

QUESTÃO 01

Um modelo de linguagem gera texto um *token* por vez. Na implementação ingênua a seguir, a cada passo o modelo recebe a sequência inteira de novo.

```
tokens = prompt                  # 1.000 tokens
for _ in range(500):
    logits = modelo(tokens)      # sequência inteira
    prox = logits[-1].argmax()
    tokens = tokens + [prox]
```

Cada *token* novo demora mais que o anterior. Com o KV *cache* ligado, o tempo por *token* cai muito e quase não cresce ao longo do texto.

O KV *cache* acelera a geração porque

A) guarda as respostas de perguntas repetidas e as devolve sem rodar o modelo de novo.
B) reduz o tamanho dos pesos do modelo, mantendo na memória só as camadas mais usadas.
C) evita recalcular as chaves (K) e os valores (V) dos *tokens* anteriores, e cada passo processa só o *token* novo.
D) permite atenção bidirecional na geração, já que o *cache* guarda também os *tokens* futuros.
E) guarda as consultas (Q) dos *tokens* anteriores, que são reaproveitadas pelos *tokens* seguintes.

Adaptado de pergunta de entrevista (Google).

QUESTÃO 02

Uma empresa quer servir um modelo de 70 bilhões de parâmetros em uma única GPU H100 SXM, com 80 GB de memória e banda de memória de 3,35 TB/s, atendendo uma conversa por vez (lote de tamanho 1). Na geração token a token, cada novo token exige ler todos os pesos do modelo da memória.

Exigindo que ao menos 30 GB fiquem livres para o KV cache e as ativações, e considerando a leitura dos pesos, a opção viável e o seu teto teórico de velocidade de geração são

A) pesos em BF16 (cerca de 140 GB), com teto de aproximadamente 24 tokens por segundo.
B) pesos em BF16 com PagedAttention, que reduz pela metade o espaço dos pesos, com teto de cerca de 48 tokens por segundo.
C) pesos em INT8 (cerca de 70 GB), com teto de cerca de 48 tokens por segundo, metade do tempo de leitura em BF16.
D) pesos em INT4 (cerca de 35 GB), com teto de milhares de tokens por segundo, pois as contas em 4 bits são mais rápidas que as de 16 bits.
E) pesos em INT4 (cerca de 35 GB), com espaço livre para o KV cache e teto de cerca de 95 tokens por segundo.

Adaptado de pergunta de entrevista (NVIDIA).

QUESTÃO 03

Um recurso de chat precisa entregar o primeiro conteúdo útil em até 3 segundos no percentil 95. O provedor do modelo oferece cache de prompt, que reaproveita o processamento de um prefixo já visto. O *system prompt* é montado pelo código a seguir.

```
system = (
    f"Data e hora: {agora():%d/%m %H:%M:%S}\n"
    + MANUAL_POLITICAS   # 8 mil tokens, mensal
    + FERRAMENTAS        # 2 mil tokens, fixas
)
resp = llm.chat(system=system,
                msgs=historico + [pergunta],
                cache=True)
```

Depois de ativar o cache, a taxa de acerto continuou em 0% e o tempo até o primeiro token não mudou.

A mudança que resolve o problema é

A) mover a data e a hora para o fim do prompt, depois do conteúdo fixo, porque o cache só reaproveita prefixos idênticos.
B) reduzir o `max_tokens` da resposta, porque o tempo até o primeiro token cresce com o tamanho da saída.
C) ativar o *speculative decoding*, que acelera o processamento do prompt antes do primeiro token.
D) aumentar para uma hora o tempo de vida do cache, para que as entradas não expirem entre as requisições.
E) trocar por um modelo de janela de contexto maior, para que o prompt de 10 mil tokens caiba com folga.

Adaptado de perguntas de entrevista (Microsoft, Apple).

QUESTÃO 04

Em um experimento, pesquisadores enviaram 1.000 vezes o mesmo pedido (*"Tell me about Richard Feynman"*) a um servidor de inferência do modelo Qwen3-235B, com temperatura 0, gerando 1.000 tokens em cada execução. Obtiveram 80 respostas diferentes, e a mais comum apareceu só 78 vezes.

O texto lembra que, mesmo em GPU, repetir a mesma multiplicação de matrizes sobre os mesmos dados produz resultados idênticos bit a bit.

HE, H. **Defeating nondeterminism in LLM inference**. Thinking Machines Lab, 10 set. 2025. Disponível em: thinkingmachines.ai. Acesso em: 15 out. 2026 (adaptado).

Segundo o estudo, a principal causa da variação observada em servidores de inferência com temperatura 0 é a

A amostragem aleatória residual, já que a temperatura 0 apenas reduz, mas não elimina, o sorteio entre os tokens mais prováveis.

B falta de invariância ao tamanho do lote nos *kernels*, pois a carga do servidor muda o lote e, com ele, a ordem das somas em ponto flutuante.

C concorrência entre as *threads* da GPU, que somam os valores em ordem diferente a cada execução, mesmo quando o lote é o mesmo.

D reutilização de respostas guardadas no cache de prompt, que devolve respostas antigas sempre que o prefixo coincide.

E atualização silenciosa dos pesos do modelo, feita pelo provedor ao longo do dia, entre uma execução e outra.

Adaptado de pergunta de entrevista (Google DeepMind, Apple).

RAG, BUSCA VETORIAL E SUAS TECNOLOGIAS
Questões de 05 a 07

QUESTÃO 05

Um sistema de recomendação guarda 50 milhões de vetores de 256 dimensões. A busca exata, que compara a consulta com todos os vetores, leva cerca de 2 segundos, e o produto exige menos de 50 milissegundos. A equipe avalia três famílias de índices de vizinhos aproximados (ANN): grafos, como o HNSW; quantização, como a PQ; e funções de *hash*, como o LSH (*Locality-Sensitive Hashing*).

Qual descrição do LSH está correta?

A Um grafo em camadas em que a busca desce do nível mais alto, com poucos nós, até o mais baixo, com todos.

B Uma compressão que divide cada vetor em subvetores e guarda só o identificador do centróide mais próximo de cada um.

C Funções de *hash* feitas para que vetores próximos caiam no mesmo balde com alta probabilidade; a consulta só é comparada com os vetores dos seus baldes.

D Uma busca exata que usa *hashes* para garantir 100% de *recall* com menos comparações.

E Um *hash* criptográfico, como o SHA-256, que espalha vetores parecidos por baldes distantes para equilibrar a carga.

Adaptado de pergunta de entrevista (compilação pública de perguntas sobre LLMs).

QUESTÃO 06

Uma loja virtual usa busca semântica, apenas com vetores densos (*embeddings*), sobre um catálogo de 2 milhões de produtos. Para buscas descritivas, como "tênis leve para corrida", os resultados são bons. Já quando o usuário digita o código do fabricante "XJ-4471", os cinco primeiros resultados são produtos parecidos, e o item exato aparece só na 38ª posição. Trocar o índice aproximado (HNSW) por busca exata não mudou o resultado.

A mudança mais adequada para esse cenário é

A somar a nota do BM25 diretamente ao cosseno do *embedding* e ordenar os produtos pela soma das duas notas.

B trocar o modelo de *embedding* por outro com mais dimensões, capaz de representar melhor sequências de caracteres.

C fazer *fine-tuning* do modelo de linguagem que redige a resposta, usando o catálogo completo.

D adicionar uma busca lexical (BM25) e fundir as duas listas pelas posições, com *Reciprocal Rank Fusion*.

E aumentar de 5 para 50 o número de vizinhos devolvidos pela busca vetorial e exibir todos ao usuário.

Adaptado de perguntas de entrevista (OpenAI, 2026; Microsoft).

QUESTÃO 07

Um sistema de RAG recupera 50 trechos por consulta com um *bi-encoder* e envia os 8 primeiros ao modelo de linguagem. Numa avaliação com 400 perguntas rotuladas, a equipe mediu os valores do quadro.

Métrica	Valor
Recall@50	0,62
Recall@8	0,48
MRR@8	0,31

Recall@k: fração das perguntas em que o trecho com a evidência aparece entre os k primeiros. Cada pergunta tem um único trecho com a evidência.

Para melhorar as respostas, a equipe propõe aplicar um *reranker* do tipo *cross-encoder* sobre os 50 trechos recuperados.

Sobre essa proposta, conclui-se que o *reranker*

A) elevará o Recall@50, porque o *cross-encoder* lê a consulta e o trecho juntos e encontra documentos que a busca deixou passar.
B) pode subir o Recall@8 até, no máximo, 0,62, já que só reordena os candidatos e não alcança as consultas cuja evidência ficou fora dos 50.
C) deve substituir o *bi-encoder* na indexação de todo o corpus, pois é mais preciso e pode ser calculado com antecedência.
D) reduzirá a latência da busca, já que o *cross-encoder* processa cada trecho mais rápido que o *bi-encoder*.
E) terá custo fixo por consulta, independentemente da quantidade de candidatos que precisar reordenar.

Adaptado de pergunta de entrevista (Microsoft).

AVALIAÇÃO E SUAS TECNOLOGIAS
Questões 08 e 09

QUESTÃO 08

TEXTO I

Imagine que você está em uma equipe que vai lançar um modelo conversacional sobre temas sensíveis, e os testes internos mostram que ele dá respostas confiantes demais, mas factualmente erradas, em contextos de alto risco. Como você investigaria e mitigaria esse problema?

Relato de candidato a Machine Learning Engineer na Anthropic, 2025. Disponível em: www.aced.io. Acesso em: 15 out. 2026 (adaptado).

TEXTO II

Como estudantes diante de questões difíceis de prova, grandes modelos de linguagem às vezes chutam quando estão incertos, produzindo afirmações plausíveis, mas incorretas, em vez de admitir a incerteza. Os modelos alucinam porque os procedimentos de treino e de avaliação recompensam o chute em vez do reconhecimento da incerteza: eles são otimizados para ir bem em provas, e chutar quando há dúvida melhora a nota.

KALAI, A. T. et al. **Why language models hallucinate**. arXiv:2509.04664, 2025 (adaptado).

Considerando o diagnóstico do Texto II, uma medida de mitigação coerente para o problema descrito no Texto I é

A) zerar a temperatura de amostragem, eliminando a aleatoriedade que produz os erros confiantes.
B) ampliar o *fine-tuning* com mais exemplos do domínio, avaliados apenas pela taxa de acerto.
C) substituir o modelo por um maior, cujo conhecimento cubra mais temas sensíveis.
D) acrescentar ao prompt uma instrução para que o modelo responda sempre de forma assertiva.
E) penalizar o erro confiante mais do que a abstenção na avaliação e aceitar o "não sei" como resposta válida.

QUESTÃO 09

Para decidir se um assistente pode ser lançado, uma equipe usou um LLM como juiz para classificar respostas como "boa" ou "ruim". Um especialista rotulou as mesmas 1.000 respostas, e o quadro compara os dois julgamentos.

Especialista	Juiz (LLM)		Total
	boa	ruim	
boa	954	16	970
ruim	24	6	30

No total, o juiz concordou com o especialista em 96% dos casos.

A conclusão correta sobre esse juiz é que ele

A) é inadequado para a decisão, pois identifica só 20% das respostas ruins, e um juiz que sempre dissesse "boa" concordaria mais com o especialista.
B) é adequado para a decisão, porque 96% de concordância supera com folga o limiar usual de 90% exigido de juízes automáticos.
C) ficaria adequado se passasse a dar notas de 1 a 5, adotando a média das notas como a métrica principal.
D) teria o problema resolvido com temperatura 0, que remove o viés do julgamento.
E) é adequado, desde que seja o mesmo modelo que gera as respostas, pois assim conhece o estilo avaliado.

Adaptado de perguntas de entrevista (Microsoft, 2025; Amazon).

AGENTES, ARQUITETURA E SUAS TECNOLOGIAS
Questões 10 e 11

QUESTÃO 10

Uma empresa quer expor 200 ferramentas internas a agentes de IA por um protocolo padrão, o MCP (*Model Context Protocol*). Cada definição de ferramenta, com nome, descrição e esquema dos parâmetros, ocupa em média 950 *tokens*. Num piloto com todas as definições no contexto, o agente passou a confundir ferramentas parecidas, como `consultar_pedido` e `consultar_pedido_legado`. Algumas ferramentas executam ações sensíveis, como estornar pagamentos e apagar cadastros.

O desenho mais adequado de descoberta, seleção e governança dessas ferramentas é

A mandar as 200 definições em todo *prompt*, contando com o *cache* de *prompt* para reduzir o custo do contexto grande.

B escolher as 10 ferramentas mais usadas da semana e mandar só essas em todo *prompt*, revisando a lista toda segunda-feira.

C fundir as 200 ferramentas numa só ferramenta genérica, com um parâmetro que diz qual sistema interno chamar.

D manter um catálogo indexado, buscar e carregar sob demanda só as ferramentas relevantes e dar a cada uma dono e escopo de permissão.

E fazer *fine-tuning* do modelo com as 200 ferramentas, para que ele as chame sem precisar das definições no contexto.

Adaptado de pergunta de entrevista (*system design*, nível Staff).

QUESTÃO 11 • BÔNUS

Uma seguradora quer um assistente para a equipe comercial responder sobre coberturas e preços. A tabela tem 40 mil itens e muda toda semana, cada resposta precisa citar a tabela e a data de vigência, e cada vendedor só pode ver as condições da sua região. Além disso, as respostas devem seguir um formato fixo de três blocos.

A abordagem mais adequada para esse caso é

A fazer *fine-tuning* semanal do modelo com a tabela nova, para que ele memorize os preços vigentes de cada região.

B usar RAG sobre a tabela versionada, com filtro por região e citação da fonte, deixando o *fine-tuning*, se preciso, só para o formato.

C colocar a tabela inteira no prompt de cada requisição, já que os modelos atuais têm janelas de contexto longas.

D fazer um *fine-tuning* completo uma única vez e pedir, no prompt, que o modelo avise quando um preço puder estar desatualizado.

E usar só engenharia de prompt com exemplos de preços, pois o modelo generaliza os valores a partir dos exemplos.

Adaptado de perguntas de entrevista (OpenAI, Microsoft, Databricks).

2042726010060

ATACADÃO 2026

Exame Nacional do Estoque Médio

FOLHA DE RASCUNHO

RASCUNHO

Paródia, sem fins comerciais. Este caderno não tem vínculo com o Atacadão S.A., o Grupo Carrefour Brasil ou o Inep.
As questões foram adaptadas de perguntas de entrevista para engenharia de IA relatadas publicamente. O gabarito comentado, com as fontes, está em arquivo separado.

ATACADÃO

CARTÃO-RESPOSTA

EXAME NACIONAL DO ESTOQUE MÉDIO 2026 • 1º DIA • CADERNO 1 • LARANJA

2042726010077

NOME DO PARTICIPANTE

Nº DO CRACHÁ ☐☐☐☐☐☐☐☐

LOJA

TURNO ☐ Manhã ☐ Tarde ☐ Madrugada

SETOR PRETENDIDO

☐ Mercearia ☐ Bebidas ☐ Limpeza ☐ Hortifrúti ☐ Frios ☐ Depósito ☐ LLMOps

TRANSCREVA A FRASE DA CAPA DO SEU CADERNO, COM SUA CALIGRAFIA USUAL

_____

_____

RESPOSTAS

Use caneta esferográfica de tinta preta. Preencha o círculo por completo. Certo: ● Errado: ⊗ ✓ ◐

FUNDAMENTOS DE LLMs, INFERÊNCIA E SUAS TECNOLOGIAS

01	A	B	C	D	E
02	A	B	C	D	E
03	A	B	C	D	E
04	A	B	C	D	E

RAG, BUSCA VETORIAL E SUAS TECNOLOGIAS

05	A	B	C	D	E
06	A	B	C	D	E
07	A	B	C	D	E

AVALIAÇÃO E SUAS TECNOLOGIAS

08	A	B	C	D	E
09	A	B	C	D	E

AGENTES, ARQUITETURA E SUAS TECNOLOGIAS

10	A	B	C	D	E
11	A	B	C	D	E

ASSINATURA DO PARTICIPANTE

VISTO DO ENCARREGADO DO SETOR

USO EXCLUSIVO DO RH DA LOJA

ACERTOS

____ / 11

RESULTADO

☐ Apto a repor gôndola

☐ Apto a repor gôndola e a colocar RAG em produção

☐ Encaminhar para a vaga de Engenheiro de IA Sênior, setor Hortifrúti

☐ Tentar de novo no próximo processo seletivo

Não amasse, não dobre, não molhe e não guarde este cartão na câmara fria.
O cartão é lido por máquina: marcas fora dos círculos podem ser interpretadas como alucinação.

✂ DESTAQUE AQUI