Fundamentos e Aplicações de Modelos de Linguagem Grandes (LLMs)

18 de outubro de 2025

Resumo

Este documento serve como um guia de estudos consolidado e uma ferramenta de autoavaliação para os conceitos fundamentais e aplicados de LLMs. A estrutura é dividida em duas partes: (1) O material de estudo curado, com resumos dos principais pontos de cada vídeo, e (2) um questionário abrangente para validar o conhecimento adquirido.

Parte 1: Material de Estudo e Resumos

Básicos de LLM e Transformers

- **3Blue1Brown - Large Language Models explained briefly** (Link para o vídeo)
  - *Resumo do usuário:* Cobre a natureza autorregressiva dos LLMs, a ideia geral de treinamento, o papel dos parâmetros/pesos, a natureza da saída de um LLM e o conceito de atenção.
- **3Blue1Brown - Transformers, the tech behind LLMs** (Link para o vídeo)
  - *Resumo do usuário:* Explica tokens, embeddings e seu significado, o cálculo da atenção e sua importância, o impacto da contagem de palavras na geração, o conceito de temperatura e como a saída de um LLM se transforma em palavras.
- **3Blue1Brown - Attention in transformers, step-by-step** (Link para o vídeo)
  - *Resumo do usuário:* Detalha a intuição por trás da atenção, o cálculo matemático e os papéis conceituais de *query*, *key* e *value* em LLMs.

Técnicas e Conhecimentos de LLM no Contexto de Aplicações

- **Should You Use Open Source Large Language Models?** (Link para o vídeo)
  - *Resumo do usuário:* Explica o que são LLMs de código aberto e seus benefícios.
- **What is a Context Window? Unlocking LLM Secrets** (Link para o vídeo)
  - *Resumo do usuário:* Explica a tokenização, o mecanismo de autoatenção e os desafios da janela de contexto.

- **Context Rot: How Increasing Input Tokens Impacts LLM Performance** (Link para o vídeo)
  - *Resumo do usuário:* Explica como os comprimentos da janela de contexto são definidos, o teste da "agulha no palheiro", a perda de desempenho em contextos inchados e exemplos de engenharia de contexto.
- **RAG vs Fine-Tuning vs Prompt Engineering: Optimizing Al Models** (Link para o vídeo)
  - *Resumo do usuário:* Define RAG, fine-tuning e engenharia de prompt, cobrindo seus benefícios, problemas e diferenças.
- **What is a Vector Database? Powering Semantic Search & Al Applications** (Link para o vídeo)
  - *Resumo do usuário:* Explica o que são bancos de dados vetoriais, como a busca semântica é implementada e como o RAG os utiliza.
- **LLM as a Judge: Scaling Al Evaluation Strategies** (Link para o vídeo)
  - *Resumo do usuário:* Explica o conceito de "LLM como Juiz"e seu papel em benchmarks.
- **What are AI Agents?** (Link para o vídeo)
  - *Resumo do usuário:* Define agentes de IA, o que eles fazem, como são implementados e seu propósito.
- **Context Engineering: The Art of Serving LLMs Efficiently** (Link para o vídeo)
  - *Resumo do usuário:* Explica o KV cache e seu impacto nas aplicações, agentes no contexto de apps de LLM e como a engenharia de contexto impacta a qualidade da resposta.
- **1-Bit LLM: The Most Efficient LLM Possible?** (Link para o vídeo)
  - *Resumo do usuário:* Cobre a quantização de parâmetros para inferência, como diferentes níveis de quantização afetam a VRAM e o desempenho, e a proposta de pesquisa de 1 bit (anotado como opcional).
- **The Unreasonable Effectiveness of Prompt "Engineering"** (Link para o vídeo)
  - *Resumo do usuário:* Define engenharia de prompt, como é feita, o que ela muda e como o tipo de modelo é impactado por essas técnicas.
- **How RAG Turns Al Chatbots Into Something Practical** (Link para o vídeo)
  - *Resumo do usuário:* Explica o que é RAG, como é usado, seus benefícios e os perigos de um sistema RAG mal implementado.

Parte 2: Questões para Avaliação

Conceitos Fundamentais

1. Uma simplificação comum é que os LLMs são como os antigos sistemas de auto-completar texto de celulares. Essa afirmação está correta? Por que sim ou por que não?
2. LLMs têm bilhões de parâmetros. O que exatamente é um parâmetro? É o mesmo que um "peso"?
3. Qual é melhor: um LLM com 3B, 8B ou 100B de parâmetros? Se o objetivo fosse um LLM especializado (ex: apenas sobre arte), isso mudaria o número ideal de parâmetros? Por quê?
4. O que é uma "cabeça de atenção"(*attention head*)? Um LLM possui mais de uma?
5. Para treinar um LLM em todo o conhecimento sobre Minecraft, o que o corpus de treinamento provavelmente conteria? Estime o volume de texto necessário. Como o resultado seria diferente se treinado com um volume igual de texto da saga Harry Potter?
6. Um LLM padrão pode ler nativamente imagens/áudio ou gerar diretamente um arquivo Excel (.xlsx)? Por que sim ou por que não? Se você precisasse de um arquivo Excel, qual seria uma abordagem prática?
7. O que ocorre durante o treinamento de um LLM que não ocorre durante a inferência, tornando a inferência computacionalmente muito mais leve?
8. O treinamento de um LLM visa minimizar um valor de "erro". O que é esse "erro"em termos conceituais?
9. Por padrão, um LLM ainda está aprendendo ou sendo treinado quando o usamos para inferência?
10. Por que recebo respostas diferentes quando envio exatamente o mesmo prompt para um modelo como o GPT em duas sessões separadas?
11. Defina e diferencie entre uma "palavra", um "token"e um "embedding".
12. Qual é o propósito de um embedding e por que os tokens precisam ser convertidos em vetores?

Conceitos de Aplicação

13. O que é uma "janela de contexto"? Se um LLM usa execução de código (ex: pandas) para responder a uma pergunta sobre um arquivo, quais são as implicações de todo esse processo ocorrer dentro de uma única janela de contexto?
14. O que é "context rot"(deterioração de contexto) e por que ocorre, considerando como a autoatenção funciona? Além da precisão, como uma janela de contexto inchada afeta a velocidade de inferência?
15. O que é "engenharia de contexto"? Forneça um exemplo prático.

16. Por que um LLM consegue gerar com sucesso um esquema de banco de dados em um formato estruturado como a sintaxe de diagrama Mermaid, mas falha em produzir um diagrama visual diretamente como "texto"?

17. Quais são duas razões principais pelas quais o fine-tuning completo é impraticável para a maioria dos indivíduos e organizações?

18. O que é engenharia de prompt? Defina e forneça exemplos distintos para um prompt de *zero-shot* e um de *3-shot*.

19. O que é RAG (Retrieval-Augmented Generation)?

20. Um usuário busca na wiki interna de uma empresa por "problemas de faturamento". Uma busca tradicional por palavra-chave não encontra nada. Como um sistema de busca semântica provavelmente forneceria uma resposta útil, e quais e como as tecnologias dos vídeos (ex: embeddings, bancos de dados vetoriais) tornam isso possível?

21. Elabore um plano de alto nível para um Agente de IA encarregado de: "Resumir as 3 principais reclamações de clientes do último trimestre relacionadas ao nosso novo sistema de faturamento e redigir um e-mail para a equipe de produto."

- Que papel a busca semântica desempenharia na fase de 'Ação' (*Act*) do agente?
- Que informações o agente precisaria armazenar em sua 'Memória' (*Memory*) para completar a tarefa?

22. Você está liderando um projeto para criar um chatbot de suporte interno para uma empresa que lida com dados de clientes altamente sensíveis. Compare as vantagens e desvantagens de usar um LLM de código aberto hospedado localmente (*on-premise*) versus usar um serviço de API de um grande provedor (como OpenAI ou Google). Faça uma tabela com as vantagens e desvantagens de cada um dos 2 tipos.