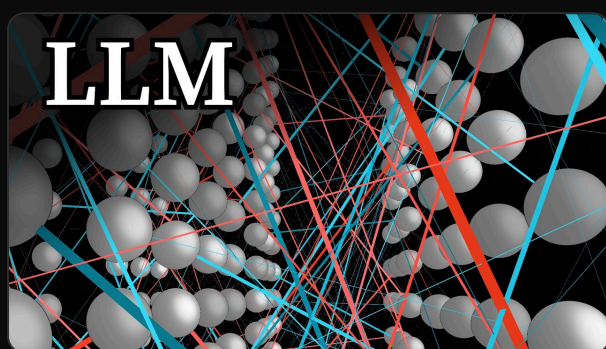


10 VÍDEOS PARA ENTENDER IA DE VERDADE.

Gratuitos no YouTube, em inglês (**ative a legenda em português**).
Assista na ordem: do básico de LLMs até os agentes.

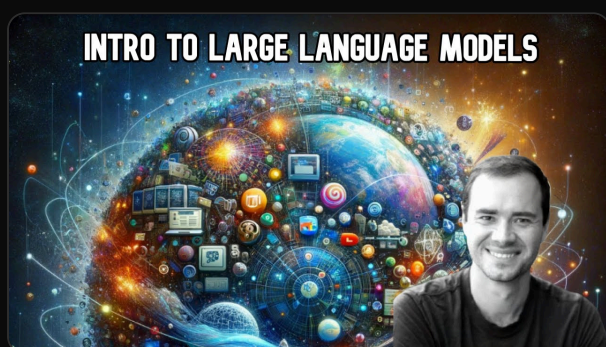


01 LLMS EXPLICADOS

3Blue1Brown · 7 min · 2024 · em inglês

A ideia central: o modelo só prevê a próxima palavra (token), uma de cada vez, a partir de bilhões de parâmetros ajustados no treino.

► youtube.com/watch?v=LPZh9BOjkQs

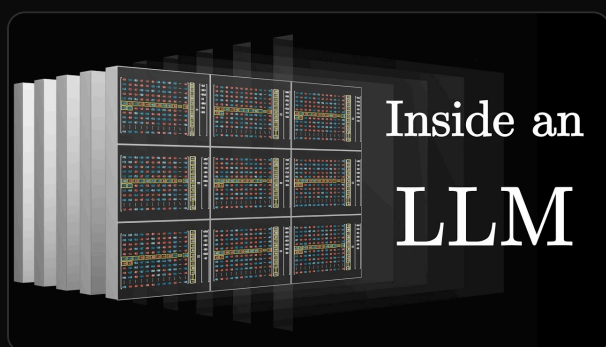


02 INTRODUÇÃO AOS LLMS

Andrej Karpathy · 59 min · 2023 · em inglês

As duas fases (pré-treino e ajuste fino), por que o modelo alucina com confiança e os riscos de segurança, como prompt injection e jailbreak.

► youtube.com/watch?v=zjkBMFhNj_g

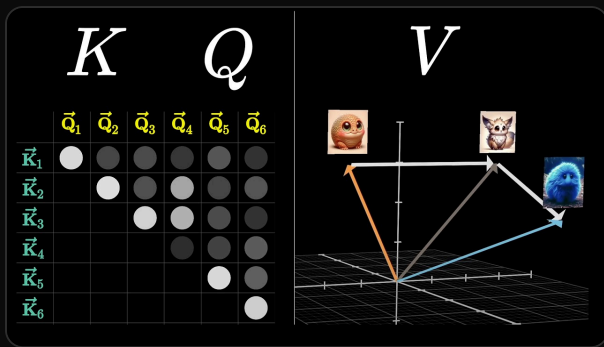


03 TRANSFORMERS

3Blue1Brown · 27 min · 2024 · em inglês

O caminho de um texto dentro do GPT: tokens, embeddings, camadas de atenção e a probabilidade da próxima palavra no final.

► youtube.com/watch?v=wjZofJX0v4M



04 ATENÇÃO NOS TRANSFORMERS

3Blue1Brown · 26 min · 2024 · em inglês

Query, key e value: como cada token decide em quais outros prestar atenção, e por que o custo cresce com o tamanho do contexto.

► youtube.com/watch?v=eMlx5fFNoYc



05 UM GPT DO ZERO, EM CÓDIGO

Andrej Karpathy · 1h56 · 2023 · em inglês

Não precisa acompanhar o código todo: preste atenção em como texto vira tokens, como o treino funciona e onde entra a atenção.

► youtube.com/watch?v=kCc8FmEb1nY



06 COMO O CHATGPT É TREINADO

Andrej Karpathy · 3h31 · 2025 · em inglês

O pipeline completo (pré-treino, ajuste fino e reforço) e a parte de “psicologia” dos LLMs: alucinação, tokens e por que ele erra conta e ortografia.

► youtube.com/watch?v=7xTGNNLPyMI

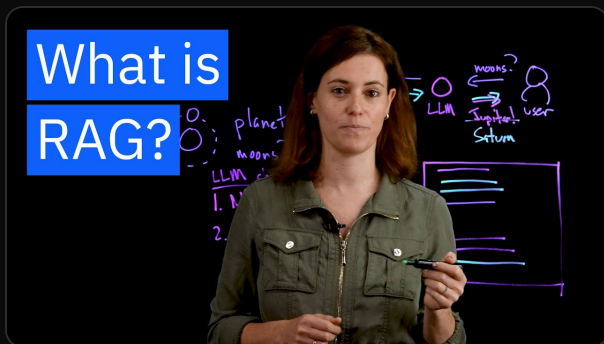


07 EMBEDDINGS E BANCOS VETORIAIS

Fireship · 3 min · 2023 · em inglês

O que é um embedding (texto virando vetor de números) e como um banco vetorial encontra conteúdo parecido por significado, e não por palavra exata.

► youtube.com/watch?v=kITvEwg3oJ4

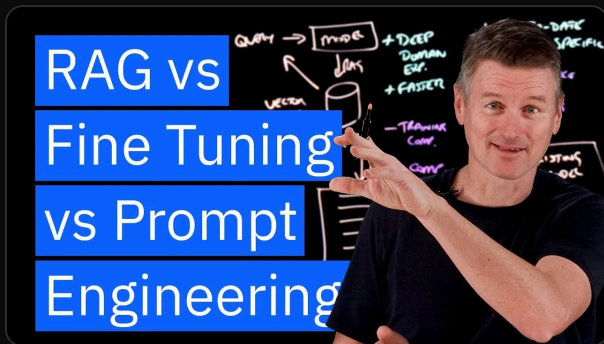


08 O QUE É RAG

IBM Technology · 6 min · 2023 · em inglês

O fluxo: pergunta → busca nos seus dados → contexto no prompt → resposta com fonte. Repare em como isso reduz alucinação e evita treinar o modelo.

► youtube.com/watch?v=T-D1OfcDW1M

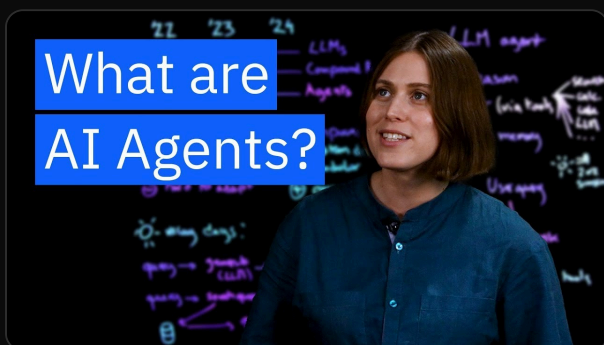


09 RAG, FINE-TUNING OU PROMPT

IBM Technology · 13 min · 2025 · em inglês

Os critérios de escolha: dados que mudam pedem RAG, comportamento específico pede fine-tuning, e muita coisa se resolve só com um bom prompt.

► youtube.com/watch?v=zYGDpG-pTho



10 AGENTES DE IA

IBM Technology · 12 min · 2024 · em inglês

A diferença entre um LLM que só responde e um agente que planeja, usa ferramentas e memória, e os trade-offs de dar autonomia ao modelo.

► youtube.com/watch?v=F8NKhkZZWI

Depois de cada vídeo, responda uma pergunta: **onde isso entra no produto em que eu trabalho hoje?** É assim que você para de só usar IA e passa a construir com ela.

Segue @lucasadvice

Conteúdo sobre carreira dev, todo dia.