

### PARTE III

#### O que são LLMs?

##### O que é a "IA" que tanto dizem

Large Language Model, ou Modelo de Linguagem Grande são modelos de aprendizado profundo baseado em redes neurais.

Treinado com conjuntos massivos de dados textuais para compreender, processar e gerar textos em linguagem natural. Funcionam processando entradas de texto (prompts) e produzindo saídas preditivas, simulando compreensão semântica por meio de padrões aprendidos.

##### Por que Existem Tantos Modelos?

Competição entre empresas. E as vezes para adaptação a contextos específicos.

Modelos são treinados com foco em contextos específicos como código, imagens, raciocínio ético, etc.

Claude é para geração de código, Grok para conversação sem restrições, ChatGPT para contexto geral, etc.



##### Como os Modelos nos Respondem? Mágica?

Respostas advêm de um processo de predição sequencial e estocástico.

Cada palavra (ou token) é gerada com base na probabilidade condicional das anteriores, semelhante a um auto-complete no celular.

Semelhante também a uma Cadeia de Markov, mas ao invés de assumir o estado presente, ele escala para contextos longos.

Processos estocásticos introduzem "criatividade" via ruído aleatório (existindo uma entropia)

##### O que Está por Baixo dos Panos?

Muitos números, muitos vetores, muito cálculo, muito poder de processamento, muito calor, etc.

"Por dentro do capô" os LLMs baseiam-se na arquitetura Transformer, introduzida em 2017 pelo paper "Attention is All You Need" por pesquisadores da Google.

Uma grande revolução ao substituir Redes Neurais Recorrentes (RNNs) por mecanismos de atenção paralelizáveis. RNNs processavam texto de forma sequencial, criando limitações e o chamado 'gargalo sequencial'. Transformers processam sequências inteiras de texto em paralelo usando 'atenção'.

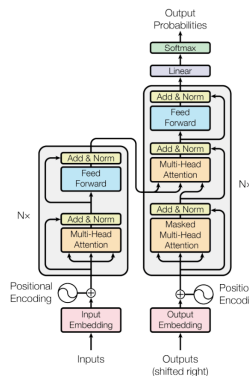


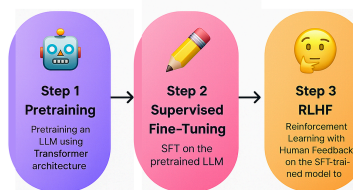
Figure 1: The Transformer - model architecture.

##### Como os Modelos Aprendem?

Aprendem via treinamento supervisionado massivo em três fases principais.

1. Pré-treinamento: Exposição a dados não rotulados para prever tokens mascarados ou próximos, aprendendo gramática, fatos gerais e diversos outros padrões de escrita.
2. Fine-tuning: Ajuste supervisionado em tarefas específicas com dados rotulados.
3. RLHF (Reinforcement Learning from Human Feedback): Otimização via feedback humano para alinhar saídas a preferência de quem o treina.

#### Training Reasoning LLM Pipeline



##### "Modelos de IA" são Inteligentes?

# NÃO.

Quaisquer LLM que você use, todos eles continuam sendo "papagaios estocásticos" ou "geradores de texto glorificados".

Independente do quão 'inteligentes' pareçam, não há nenhum raciocínio, lógica ou inferência, apenas probabilidades.

**MODELOS DE LLM SÃO APENAS GERADORES DE TEXTO PROBABILÍSTICOS GLORIFICADOS, APENAS.**