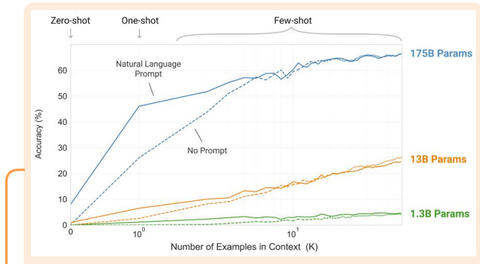
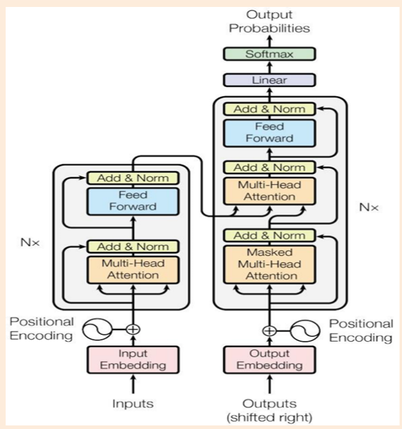
IA em Cenários Reais

1. ChatGPT

- Lançado em 30/11/2022. Significado do acrônimo: *Chat Generative Pre-Trained Transformer*.
- É um chatbot baseado em um modelo largo de linguagem (LLM) desenvolvido pela OpenAI.
- É uma aplicação web na forma de chatbot

2. Modelo GPT

- É o modelo de IA utilizado para prover o serviço do ChatGPT
- O modelo GPT é dividido em versões
  - **GPT-1:** lançado em 11/06/2018, possuia uma estrutura de 117 milhões de hiperparâmetros, treinado como conjunto de dados BookCorpus (4,5GB de texto)
    - O treinamento da rede levou 30 dias utilizando 8 GPUs do modelo P600 que são capazes de processar aproximadamente 1 PetaFLOP/s (1.000.000.000.000)
      - FLOP, *Floating Point Operation* ou *Operações com ponto flutuante* são uma medida de quantas operações matemáticas os processadores são capazes de realizar por segundo - **Peta** é um sufixo que indica 10 elevado a décima segunda potência ou 1 seguido de doze zeros
  - **GPT-2:** Aumento o número de hiperparâmetros para 1,5 bilhões, e foi treinado utilizando 40BG de texto, sendo lançado em 14/02/2019.
    - O treinamento teve um custo estimado de 1,5 ZettaFLOPs (15 seguido de 20 zeros, 1.500.000.000.000.000.000.000
      - **Zetta** - Sufixo que indica 1 seguido de 21 zeros, ou 10^{21}
  - **GPT-3:** Lançado em maio de 2020, foi o modelo utilizado inicialmente no ChatGPT em seu lançamento.
    - Este modelo elevou o número de hiperparâmetros para 175 bilhões e o conjunto de dados para treinamento foi composto pelo CommonCrawl (Aproximadamente 570GB), WebText, Wikipedia e dois livros cujo conteúdo é ofuscado.
    - O treinamento do modelo usou um total de 310 ZettaFLOPs (310.000.000.000.000.000.000.000 - 310 x 10^{21})
    - O GPT possui uma publicação aberta no arxiv.org, um artigo explicando o funcionamento dos modelos de linguagem
      - 

      - No trabalho dos pesquisadores, foi avaliado a acurácia dos modelos com diferentes dimensões (quantidade de hiperparâmetros) como pode ser observado na figura
      - É possível observar o número de hiperparâmetros tem impacto significativo na acurácia
- Embora a dimensão do modelo de IA tenha impacto direto no sucesso da ferramenta, a tecnologia GPT não é uma criação antiga
  - Os geradores pré-treinados *(generative pre-trained)* são modelos de deep learning desenvolvidos nos anos 2012.
  - Já os transformadores (transformers) são uma topologia de deep learning proposta por pesquisadores do Google em 2017 para substituir redes neurais artificias complexas recorrentes e redes neurais convolucionais no processamento de linguagem natural. Os transformadores são compostos por:
    - **Tokenizers:** Responsáveis por converter o texto em tokens
    - **Camada de Embutir:** Responsável por converter tokens e as posições dos tokens em representação vetorial
    - **Camadas de Transformação:** Executam repetidas transformações nas representações vetoriais, extraindo cada vez mais informação linguística. São compostas por duas subcamadas - Subcamada de atenção e subcamada feedforward
    - **Camada de Desembutir (opcional):** Responsável por converter as representações finais em distribuição de probabilidades nos diferentes tokens.
  - 

    A Figura 2 apresenta a organização de um transformador como apresentado no artigo original do Google: os elementos cinza são as camadas de transformação com as subcamadas de atenção (alaranjado) e feedforward (azul). A camada de embutir está representada na cor rosa e a camada de desembutir é omitida no modelo.