Nome: Fernando Buligon Antunes **Data:**28/03/2025

Introdução aos LLM

Large Language Models (LLM), como o nome já sugere, são modelos treinados em grandes quantidades de dados de texto, geralmente retirados de publicações públicas de toda a internet como blogs, artigos, posts em redes sociais, comentários e etc. Para ter noção do quão grande é a base de dados usada por esses modelos, um arquivo de 1GB é capaz de armazenar em média 178 milhões de palavras, o que é bastante coisa, mas ainda não chega nem perto do tamanho dos modelos usados, que podem chegar a ter petabytes, dentro de um petabyte, tem 1 milhão gigabytes.

LLMs são uns dos maiores modelos na questão de contagem de parâmetros, parâmetros são valores que o modelo consegue mudar sozinho conforme vai aprendendo, e quanto maior a quantia de parâmetros, possivelmente mais complexo é. Como por exemplo o famoso modelo GPT-3, usa 175 bilhões de parâmetros, sendo pré treinado com 45 terabytes de dados.

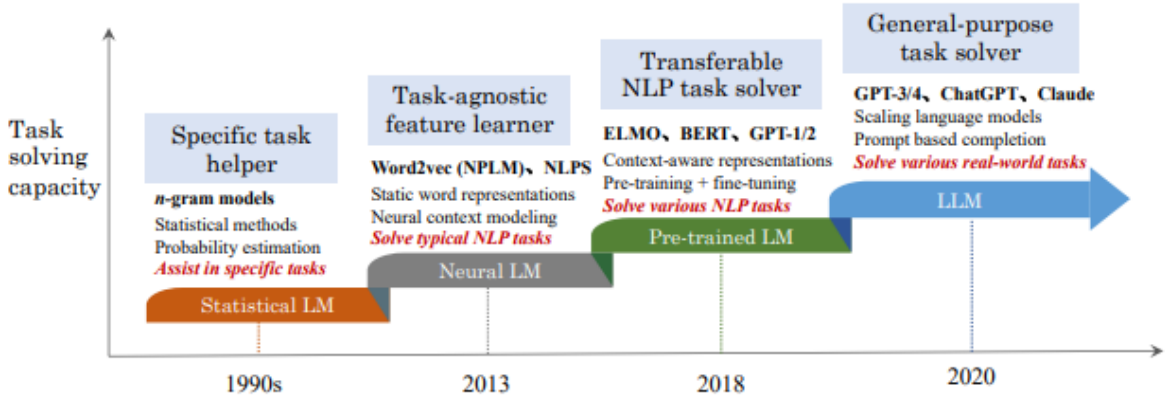
LLM pode ser dividido em três componentes principais: dados, arquitetura e treinamento. A parte de dados já foi falada, e em relação a arquitetura, são usados os transformers, que permitem analisar sentenças inteiras, linhas, palavras, contexto das palavras em relação às outras, depois essa arquitetura é treinada com os dados, e no treinamento o modelo aprende a prever qual é a próxima palavra de uma frase, e a cada iteração o modelo vai ajustando seus parâmetros para cada vez ficar mais preciso.

O modelo pode ser ajustado em uma base de dados menor específica através do fine tuning, tendo dados específicos para uma tarefa, o modelo é capaz de se aperfeiçoar melhor para a tarefa designada

Large Language Models (Tópicos 1 e 2)

O início é dado definindo linguagem como uma habilidade proeminente dos seres humanos de se comunicarem e expressarem, desenvolvendo desde o início da infância até a hora final. Por outro lado, máquinas não possuem a habilidade de desenvolver naturalmente esse processo. Esse é um dos desafios mais antigos dentro do campo da inteligência artificial, permitir com que as máquinas sejam capazes de entender nossa linguagem para que consigam ler e escrever, e é aí que LM (Language Modeling) entra, sendo uma das tentativas mais promissoras nesse campo.

Depois é mostrado um pouco da evolução dos modelos ao decorrer dos tempos, colocando em perspectiva as gerações e a capacidade de solução de tarefas, como é possível ver na imagem abaixo.



O desenvolvimento dos estudos de LM podem ser separados nesses quatro pontos destacados na imagem acima.

Statistical LM: A ideia principal é construir um modelo de predição de palavras baseado na hipótese de Markov, prevendo a próxima palavra em relação ao contexto mais recente, o principal problema era em relação a precisão em modelos de alta ordem.

Neural LM: Usavam redes neurais, como RNN ou MLP para fazer a predição de sequências de palavras, foi aqui que introduziram representação distribuída de palavras, mais tarde, word2vec foi proposto para construir uma rede neural mais simples.

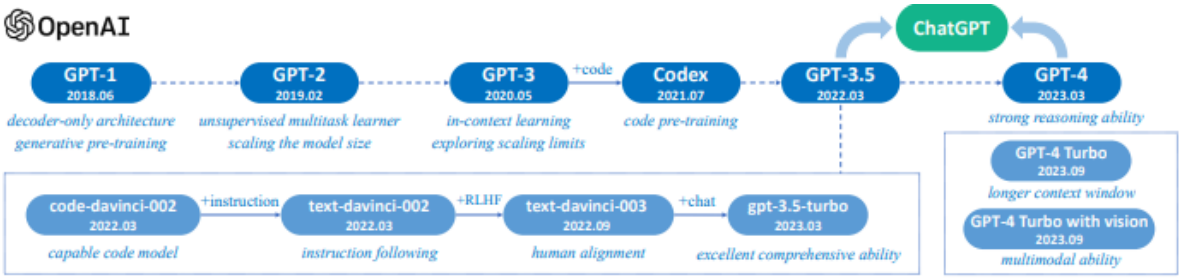
Pre-trained LM: Faziam a captura da representação de contexto das palavras, o ELMO foi um dos primeiros usando um rede pré treinada com fine tuning, depois o BERT com arquitetura transformer, treinado com tarefas de pré treinamento em grandes corpus não rotulados. Essas representações de palavras pré-treinadas e sensíveis ao contexto são muito eficazes em propósitos gerais.

LLM: Pesquisas mostraram que aumentando a escala de PLM, contendo base de dados maiores, a capacidade do modelo também aumentava na solução de tarefas genéricas e específicas, um exemplo famoso é o ChatGPT.

Na seção foi feita uma revisão sobre como LLMs funcionam e sobre a evolução da série de modelos GPT. Falando que tipicamente, LLMs se referem a transformers que contém inúmeros parâmetros, passando da casa do bilhão, que são treinados em uma base de dados textual gigantesca, permitindo com que sua capacidade em compreender a linguagem natural e de resolver tarefas por meio da geração de textos seja muito forte.

Depois foi falado da família dos modelos GPT, que ficaram muito famosos com o ChatGPT, que graças a sua alta capacidade em resolver tarefas através de um simples prompt, animou bastante a comunidade da IA desde seu lançamento.

Na imagem abaixo é possível ver um resuminho sobre cada modelo da OpenAI



Large Language Models (Tópicos 8 e 9)

Na seção oito, foi abordado sobre progressos nas aplicações de LLM em dois pontos, comunidade de pesquisa e domínios específicos.

Comunidade de pesquisa:

- Cenários clássicos:
  - Tarefas de NLP clássicas
    - Processamento de palavras e sentenças;
    - Segmentação de sequência;
    - Extração de informações;
    - Geração de texto.
  - IR (Information Retrieval
    - LLM como modelo de IR
    - LLM como modelo de IR melhorado
  - Recomendação
    - LLM como modelo de recomendação
    - LLm como modelo de recomendação melhorado
    - LLM como simulador de recomendação
- Capacidades melhoradas:
  - LLM Multimodal
    - Pré-treinamento de alinhamento visão-linguagem
    - Ajuste de instruções visuais
    - Avaliação de LLM multimodal
  - LLM aprimorado por KG
    - LLM com recuperação aumentada
    - LLM sinergicamente aumentado
- Novos cenários:
  - Agente baseado em LLM
    - Componentes: Memória/Planejamento/Execução
    - Aplicação baseada em agente único/múltiplo
  - LLM para avaliação
    - Avaliação baseada em pontuação/linguagem
    - Design de instruções, múltiplos feedbacks, agente de debate
    - Meta-avaliação

Domínios específicos:

- Saúde
- Finanças
- Pesquisa científica
- Leis
- Educação

Na seção nove, foi feita uma conclusão sobre o assunto abordado ao longo do artigo, falando de alguns pontos como **princípios básicos**, que fala sobre como LLMs aprendem por pré-treinamento não supervisionado em base de dados gigantescas, **arquitetura do modelo**, transformers são a arquitetura padrão para LLM devido a sua eficiência, **treinamento do modelo**, o pré treinamento do modelo exige uma infraestrutura

de dados limpa e organizada, **utilização do modelo**, baseada na interpretação de linguagem natural, a utilização de prompts é a principal forma de usar LLMs para resolver problemas, **segurança e alinhamento**, o principal risco de LLMs são as alucinações (quando elas inventam coisa), e uso malicioso (dependendo do prompt, o usuário pode extrair informações que não deveria saber), **Aplicações e ecossistema**, as LLMs demonstram capacidades altas em solucionar inúmeras tarefas, por isso podem ser aplicadas em uma grande área de problemas do mundo real, o mais famoso hoje em dia é o ChatGPT.