Introdução a Generative AI, LLMs e desenvolvimento com LangChain

Implementação prática em Python

O que é IA Generativa?

Definição: A inteligência artificial generativa (IA generativa) é um tipo de IA que pode criar novos conteúdos e ideias, incluindo conversas, histórias, imagens, vídeos e músicas. As tecnologias de IA tentam imitar a inteligência humana em tarefas de computação não tradicionais, como reconhecimento de imagem, processamento de linguagem natural (PLN) e tradução. A IA generativa é a próxima etapa da inteligência artificial. Você pode treiná-la para aprender linguagem humana, linguagens de programação, arte, química, biologia ou qualquer assunto complexo. Ela re-utiliza dados de treinamento para resolver novos problemas.

O que é IA Generativa?

IA generativa refere-se a uma classe de modelos de inteligência artificial que podem gerar novos conteúdos, como texto, imagens, áudio e vídeo, com base nos dados em que foram treinados.

Sua organização pode usar a IA generativa para várias finalidades, como chatbots, criação de mídia e desenvolvimento e design de produtos.

Aplicações comuns da IA generativa

Geração de texto: Os exemplos incluem chatbots, criação de conteúdo e serviços de tradução.

Geração de imagens: criar obras de arte, melhorar fotos ou gerar deepfakes.

Geração de áudio: sintetizar música, criar locuções ou gerar efeitos sonoros.

Geração de vídeos: Produção de animações, vídeos deepfake e edição de vídeos.

Geração de código: A geração de código é uma das aplicações mais promissoras para IA generativa, como exemplo a utilização de assistentes para desenvolvimento de software, e ferramentas de aumento de produtividade do desenvolvedor.

O que são LLMs?

Definição: Em contraste, os LLMs (Large Language Models) constituem uma categoria específica de modelos generativos de IA com foco especializado em dados baseados em texto.

Estes modelos passam por um treinamento rigoroso em grandes volumes de dados de texto, abrangendo fontes como livros, artigos e códigos. Após a conclusão do treinamento, os LLMs estão preparados para tarefas relacionadas a texto, incluindo geração de texto, tradução de idiomas e criação de conteúdo em vários gêneros, além de fornecer respostas informativas a consultas.

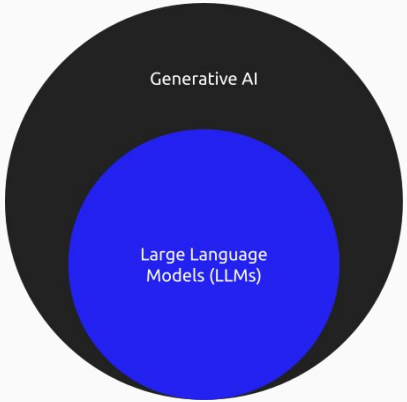
O que são LLMs?

Em outras palavras, LLMs são uma forma de IA que se concentra na compreensão de entradas de texto, usando processamento de linguagem natural (PLN), e na criação de texto semelhante ao humano com base em uma determinada entrada.

Os LLMs são um subconjunto da IA generativa e concentram-se principalmente em tarefas relacionadas à linguagem.

Generative AI vs. Large Language Models (LLMs)

- Cria vários tipos de saídas de conteúdos
- Construído em llms, mas também em outros tipos de modelos de aprendizado de máquina



- criar somente saídas de texto
- construído sobre bilhões de parâmetros para compreender e produzir texto

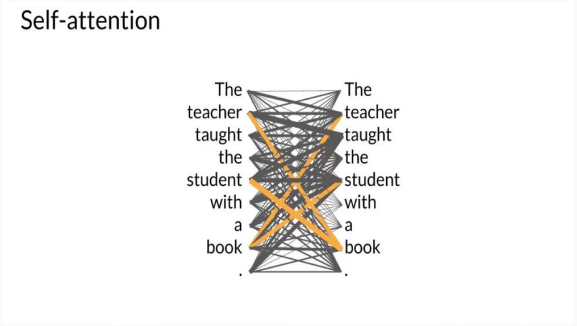
Como funcionam e se aplicam os LLMs

Large Language Models (LLMs) como GPT-3 e GPT-4 da OpenAI (Generative pre-trained transformer) são algoritmos de aprendizado de máquina projetados para compreender e gerar texto semelhante ao humano com base nos dados nos quais foram treinados. Esses modelos são construídos usando redes neurais com milhões ou até bilhões de parâmetros, tornando-os capazes de realizar tarefas complexas como tradução, resumo, resposta a perguntas e até redação criativa.

Treinados em conjuntos de dados diversos e extensos, muitas vezes abrangendo partes da Internet, livros e outros textos, os LLMs analisam os padrões e relações entre palavras e frases para gerar resultados coerentes e contextualmente relevantes. Embora possam realizar uma ampla gama de tarefas linguísticas, não são conscientes e não possuem compreensão ou emoções, apesar de sua capacidade de imitar tais qualidades no texto que geram.

Como funcionam e se aplicam os LLMs

Os LLMs pertencem principalmente a uma categoria de estruturas de aprendizagem profunda (deep learning) conhecidas como transformers network. Um transformer network é um tipo de rede neural que obtém uma compreensão do contexto e do significado ao identificar as conexões entre os elementos de uma sequência, assim como as palavras em uma determinada frase.

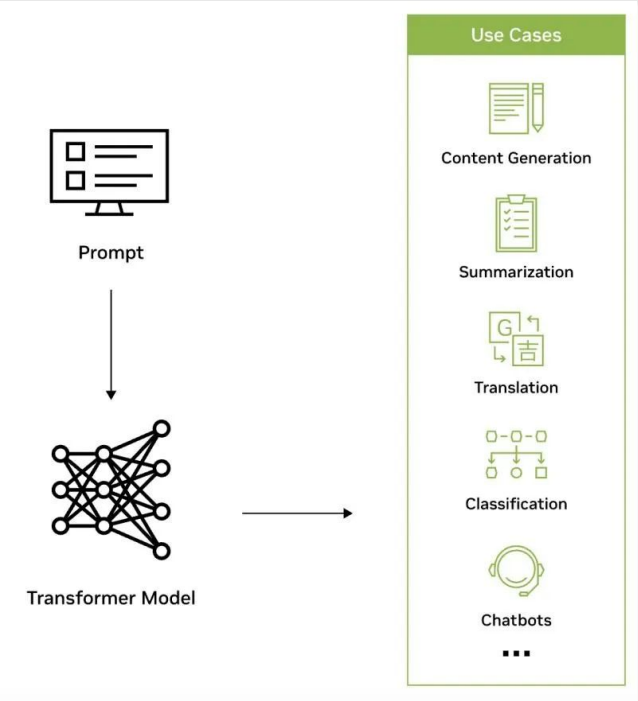


A arquitetura do transformador revolucionou as tarefas de linguagem natural e impulsionou os modelos de linguagem a novos patamares de desempenho. Um de seus principais pontos fortes está na autoatenção (self-attention), que permite ao modelo compreender a relevância e o contexto de cada palavra em uma frase. Ao atribuir pesos de atenção às relações entre as palavras, o modelo ganha uma compreensão abrangente da linguagem. Isso é representado em um mapa de atenção, onde as conexões entre as palavras são destacadas.

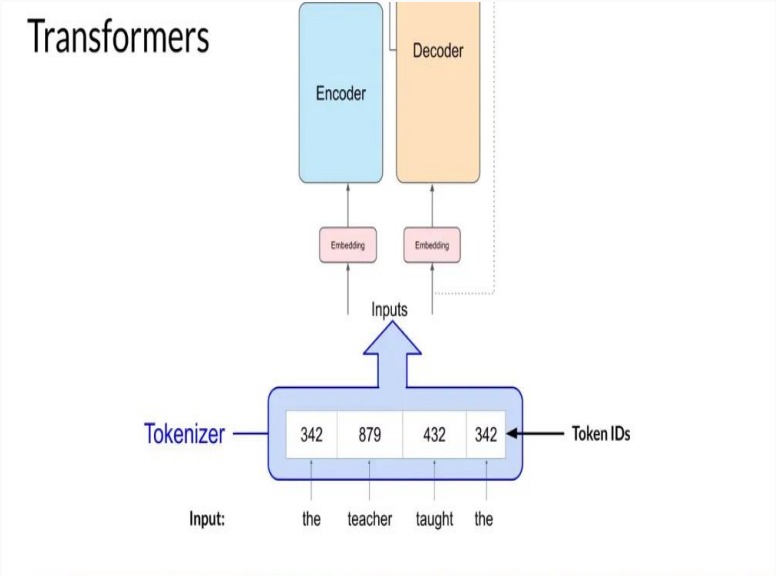
Como funcionam e se aplicam os LLMs

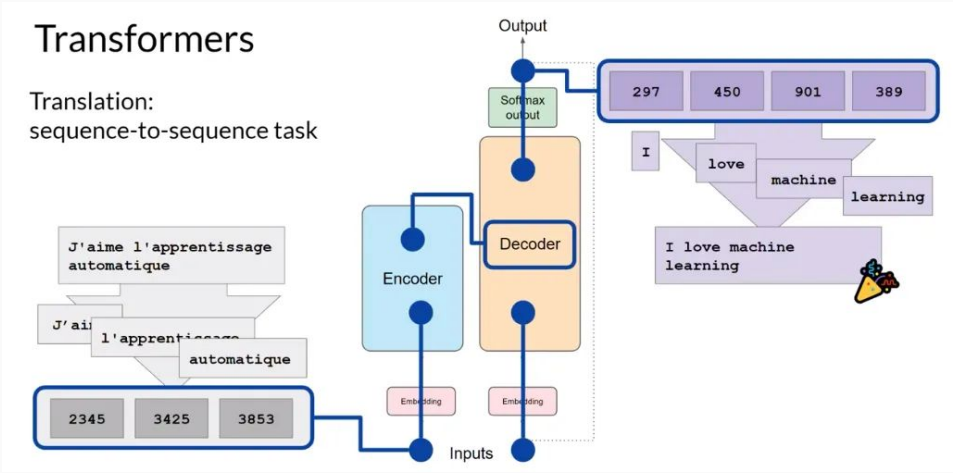
A arquitetura do transformador consiste em dois componentes principais: o codificador e o decodificador, ambos compartilhando semelhanças.

No entanto, antes de inserir texto no modelo, as palavras precisam ser “tokenizadas” e convertidas em representações numéricas usando um tokenizer. Isso permite que o modelo trabalhe com números em vez de palavras. A camada de incorporação mapeia esses IDs de token para vetores de alta dimensão, codificando o significado e o contexto de cada token. Esses vetores ocupam uma localização única no espaço de incorporação, facilitando a compreensão matemática da linguagem.



Como funcionam e se aplicam os LLMs





https://medium.com/@yash9439/introduction-to-llms-and-the-generative-ai-part-1-a946350936fd

O que é LangChain?

LangChain é uma estrutura de orquestração de código aberto para o desenvolvimento de aplicativos integrados com LLMs.

O LangChain atua como uma interface genérica para praticamente qualquer LLM, ou seja, o núcleo do LangChain fornece um ambiente de desenvolvimento que otimiza e facilita a programação de aplicativos de LLM por meio do uso de abstração.

https://www.langchain.com/

O que é LangChain?

O LangChain é essencialmente uma biblioteca de abstrações para Python e Javascript (principais linguagens), representando etapas e conceitos comuns necessários para trabalhar e simplificar o processo de criação de aplicativos baseados em LLM.

Esses componentes modulares, como funções e classes de objetos, servem como blocos de construção de programas de IA generativos.

Podem ser "encadeados" para criar aplicativos, minimizando a quantidade de código e o entendimento refinado necessário para executar tarefas complexas de PLN (processamento de linguagem natural).

Modelos de Linguagem

Quase qualquer LLM pode ser usado no LangChain.

É fácil importar modelos de idioma no LangChain, desde que você tenha uma chave da API. A classe base de interface com as LLMs foi projetada para oferecer uma interface padrão para todos os modelos.

A maioria dos provedores de LLM exigirá que você crie uma conta para receber uma chave de API. Algumas dessas APIs, especialmente aquelas para modelos proprietários de código fechado, como as oferecidas pela OpenAI ou pela Anthropic, podem ter custos associados.

Modelos de Prompt

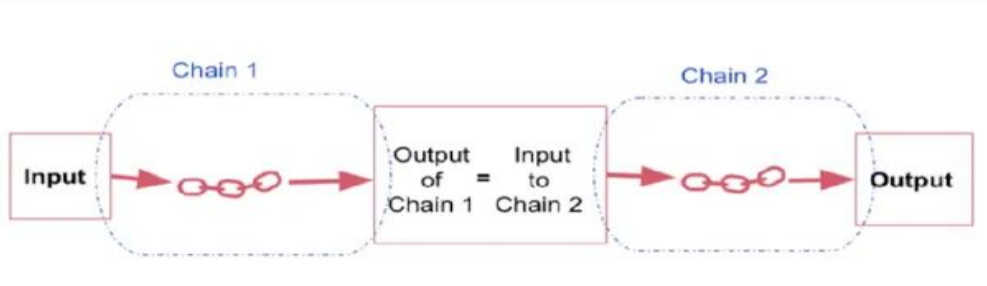
Os prompts são as instruções apresentadas a um LLM. Geralmente, a "arte" de redigir prompts que efetivamente entregam o contexto necessário para que o LLM interprete a entrada e, a saída da estrutura da maneira mais útil para você é chamada de engenharia rápida.

A classe PromptTemplate em LangChain formaliza a composição de prompts sem a necessidade de codificar manualmente o contexto e as consultas. Elementos importantes de um prompt também são inseridos como classes formais, como “input_variables”.

Um modelo de prompt pode, portanto, conter e reproduzir contexto, instruções, um conjunto de exemplos para orientar suas respostas, um formato de saída definido ou uma pergunta padronizada a ser respondida, e além de poder ser salvo e re-utilizado quando necessário.

Correntes (Chains)

Como seu nome implica, cadeias são o núcleo dos fluxos de trabalho do LangChain.
Combinam LLMs com outros componentes, criando fluxos por meio da execução de uma sequência de funções.
Cada etapa da cadeia pode ser uma chamada para um LLM, uma ferramenta (tool) ou um pré-processamento de dados.
A principal forma suportada de fazer isso é com LCEL (LangChain Expression Language).



Índices (Indexes)

Para realizar determinadas tarefas, as LLMs precisarão acessar fontes de dados externas específicas não incluídas em seu conjunto de dados de treinamento, como documentos internos, e-mails ou conjuntos de dados. LangChain refere-se coletivamente a essa documentação externa como “índices".

Exemplos de índices:

- Carregadores de documentos (Document loaders)
- Bancos de dados vetoriais (Vector databases)
- Divisores de texto (Text splitters)
- Recuperação (Retrievers e Retrieval-Augmented Generation aka RAG)

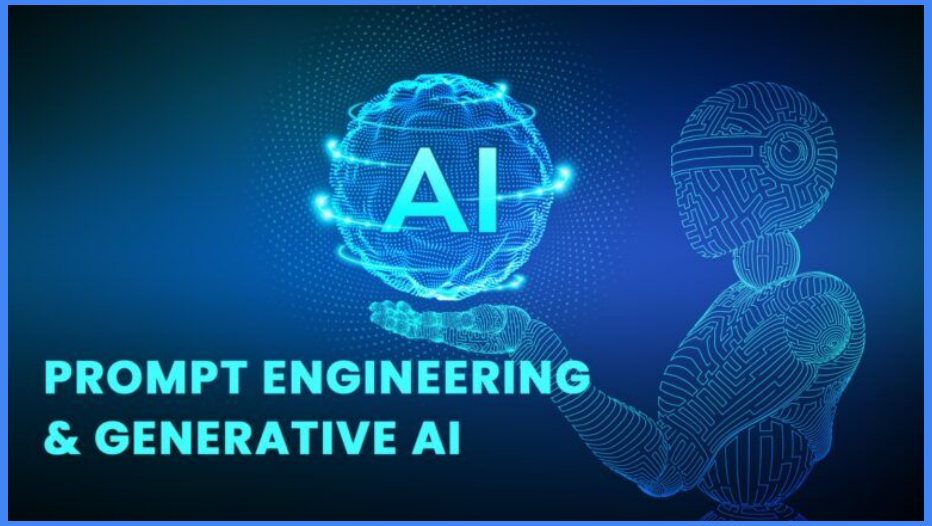
Memória (Memory)

Por padrão, os LLMs não têm memória de longo prazo de conversas anteriores (stateless).
A sua adição permite uma conversa coerente e, pois sem ela, cada consulta seria tratada como uma entrada totalmente independente, sem considerar as interações anteriores.
O LangChain soluciona esse problema com utilitários simples para adicionar memória a um sistema, com opções que vão desde a retenção total de todas as conversas até a retenção de um resumo da conversa até a retenção das n trocas mais recentes.

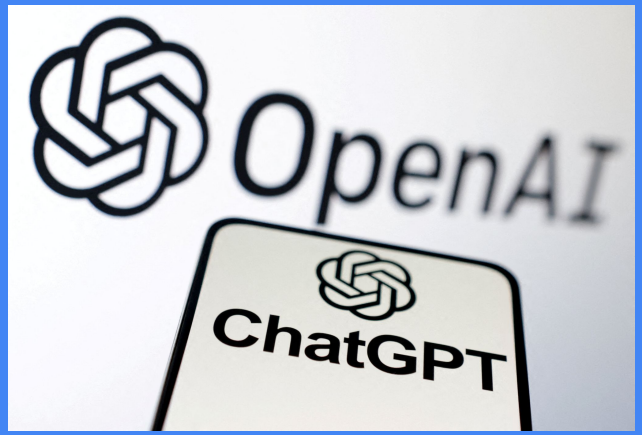
Agentes (Agents)

Os agentes do LangChain podem usar um determinado modelo de idioma como um "mecanismo de raciocínio" para determinar quais ações tomar. Ao criar uma cadeia para um agente, as entradas contêm:

- uma lista de ferramentas disponíveis para serem aproveitadas (Tools).
- entrada do usuário (prompt templates and queries).
- quaisquer etapas relevantes executadas anteriormente (Memory)
- módulos de recuperação de documentos (Retrievers and RAG)


AI
PROMPT ENGINEERING
& GENERATIVE AI


LLaMA


OpenAI
ChatGPT


LangChain

Exemplos práticos

- Exemplos disponíveis em

  https://github.com/marcellorego/langchain-notebook/

- Agenda de exemplos
  - Chat Models
  - Prompt Templates
  - Chains
  - RAG (Retrieval-Augmented Generation)
  - Agents & Tools

"O sucesso é a soma de pequenos esforços repetidos dia após dia."

Robert Collier

Obrigado!

Contato:

BySix
Praceta Prof. Alfredo de Sousa 8,
Algés, Oeiras
1495-072

contacto@bysix.com
https://www.bysix.com/