Insper

Inteligência Artificial Generativa

Aula 7

Magno TF Severino
PADS - Aprendizagem Estatística de Máquina II

Objetivos de aprendizagem

- Compreender os conceitos fundamentais de Inteligência Artificial (IA) Generativa.
- Exploração de modelos de IA generativa em diferentes domínios (texto, imagem e audio).
- Entender a estrutura aspectos básicos Large Language Models (LLMs).
- Integrar o R com API de modelos de linguagem.

Inteligência Artificial Discriminatória

- Modelos estatísticos para dados supervisionados/rotulados.
  - Predição (regressão e classificação).
  - Treinamento baseado em rótulos (variável resposta).
  - Aprende a relação entre as varíaveis e os rótulos.
- Modelos estatísticos para dados não supervisionados.
  - Redução de dimensionalidade.
  - Segmentação de dados.
  - Análise de associação.

Inteligência Artificial Generativa

- Cria novos dados similares a partir dos dados de treinamento.
- Tem uma "criatividade" restrita aos dados de treinamento.
- Recombinação de estilos, padrões, símbolos.
- Exemplo: prevê a próxima palavra em uma sequência.

Dados rotulados

↓

Modelo Discriminatório

↓

Rótulos

Dados não estruturados

↓

Modelo Generativo

↓

Novo conteúdo

Modelos de IA Generativa

- Linguagem
- Imagem
- Audio
- Vídeo

Modelos de IA Generativa

Linguagem

- Aprende os padrões de uma língua a partir dos dados de treinamento.
- Prevê a próxima palavra dada uma sequência.

Imagem

- Dado um prompt descritivo, o modelo transforma um ruido aleatório em uma imagem (modelo de difusão).
- Produz novas imagens usando diferentes técnicas.

Principais Conceitos

Large Language Model (LLM)

- Uma classe de modelos de linguagem que são capazes de gerar texto de forma autônoma, aprendendo padrões complexos em grandes conjuntos de dados de texto.
- Utiliza redes neurais profundas, geralmente baseadas em arquiteturas de transformers, para gerar sequências de texto de alta qualidade e coerência.

Large Language Model (LLM)

- Grande quantidade de dados de treinamento e parâmetros no modelo.
- De propósito geral - semelhança das linguagens humanas.
- Apresenta comportamentos emergentes.
- Adaptação rápida para outras tarefas.

Generative AI is a technology that *learns from a set of data* and then generates new content that *resembles the original data*.

Principais Conceitos

Generative Pre-trained Transformer (GPT)

- ChatGPT é um **produto** baseado em um LLM (Large Language Model).
- Uma série de modelos de linguagem desenvolvidos pela OpenAI, sendo o GPT-4o o modelo mais recente.
- Utiliza a arquitetura Transformer, que permite que o modelo aprenda padrões de longo alcance em textos extensos.
- Pré-treinado em grandes corpora de texto para aprender uma representação generalizada da linguagem, que pode ser finamente ajustada para tarefas específicas.

Principais Conceitos

Mecanismo de atenção

- Técnica que permite ao modelo "prestar atenção" a diferentes partes da entrada durante a geração de saída.
- Semelhante a destacar palavras-chave ao ler uma frase para entender o significado geral.
- Permite ao modelo dar mais importância a certas palavras ou tokens em diferentes momentos do processamento.

Principais Conceitos

Mecanismo de atenção

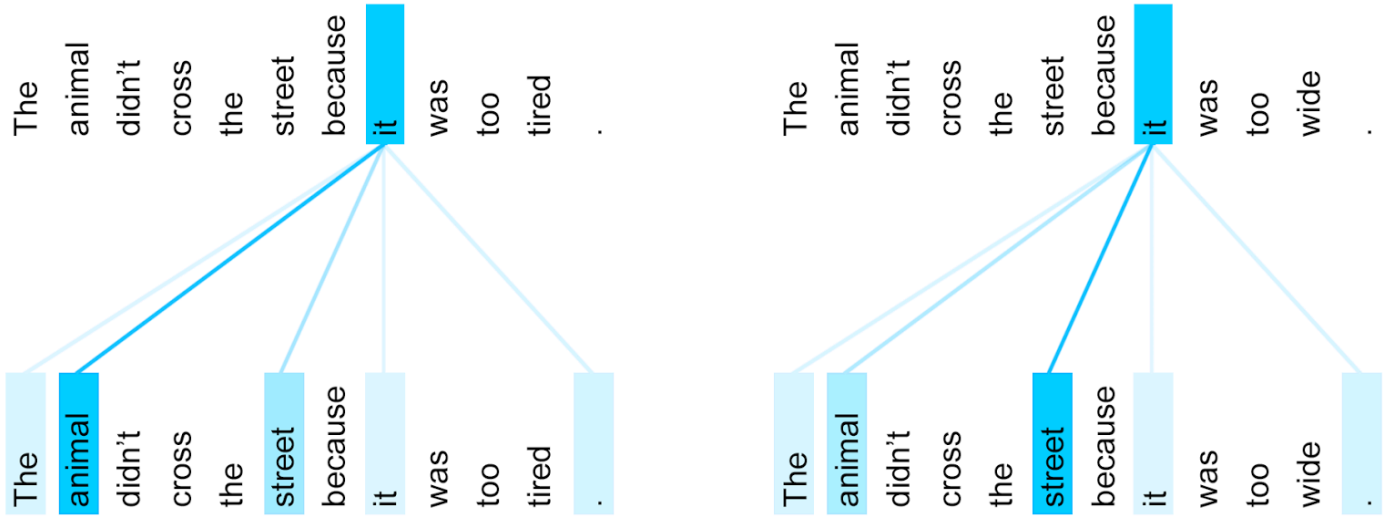


Figura da página DEVOPEDIA.

Principais Conceitos

Modelos de Linguagem

- Representam palavras pelo contexto.
- Um modelo estatístico ou de aprendizado de máquina que é capaz de prever a próxima palavra em uma sequência de texto com base nas palavras anteriores.
- Semântica distribucional: "você conhece uma palavra pelas suas companhias".
- O significado da palavra é dado pelas palavras que aparecem frequentemente perto.

Principais Conceitos

Embeddings de Texto

- Representação numérica de palavras ou sequências de texto.
- Transformação de dados textuais em vetores de números densos.
- Facilita o processamento de linguagem natural em algoritmos de aprendizado de máquina.
- Motivação: limitações da representação tradicional de texto (one-hot encoding).
    - Esparsidade e dimensionalidade alta.
    - Dificuldades em capturar semântica e relações entre palavras.

Funcionamento de Embeddings de Texto

- Mapeamento de palavras para vetores contínuos.
- Aprendizado de representações semânticas.
- Preservação de similaridades e relações entre palavras.
- Uso de modelos pré-treinados (ex: Word2Vec, GloVe, FastText).
- Vantagem: Redução do tempo de treinamento e melhoria na performance.

Embeddings - Exemplo

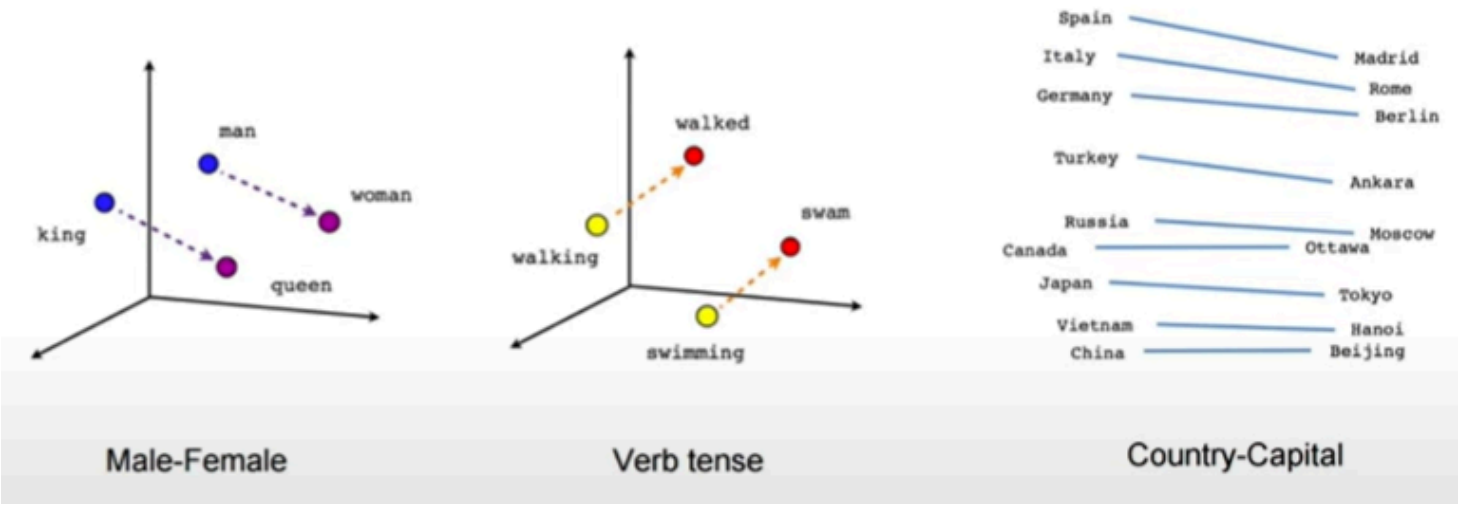


Figura da página di.school.

Arquitetura Transformer

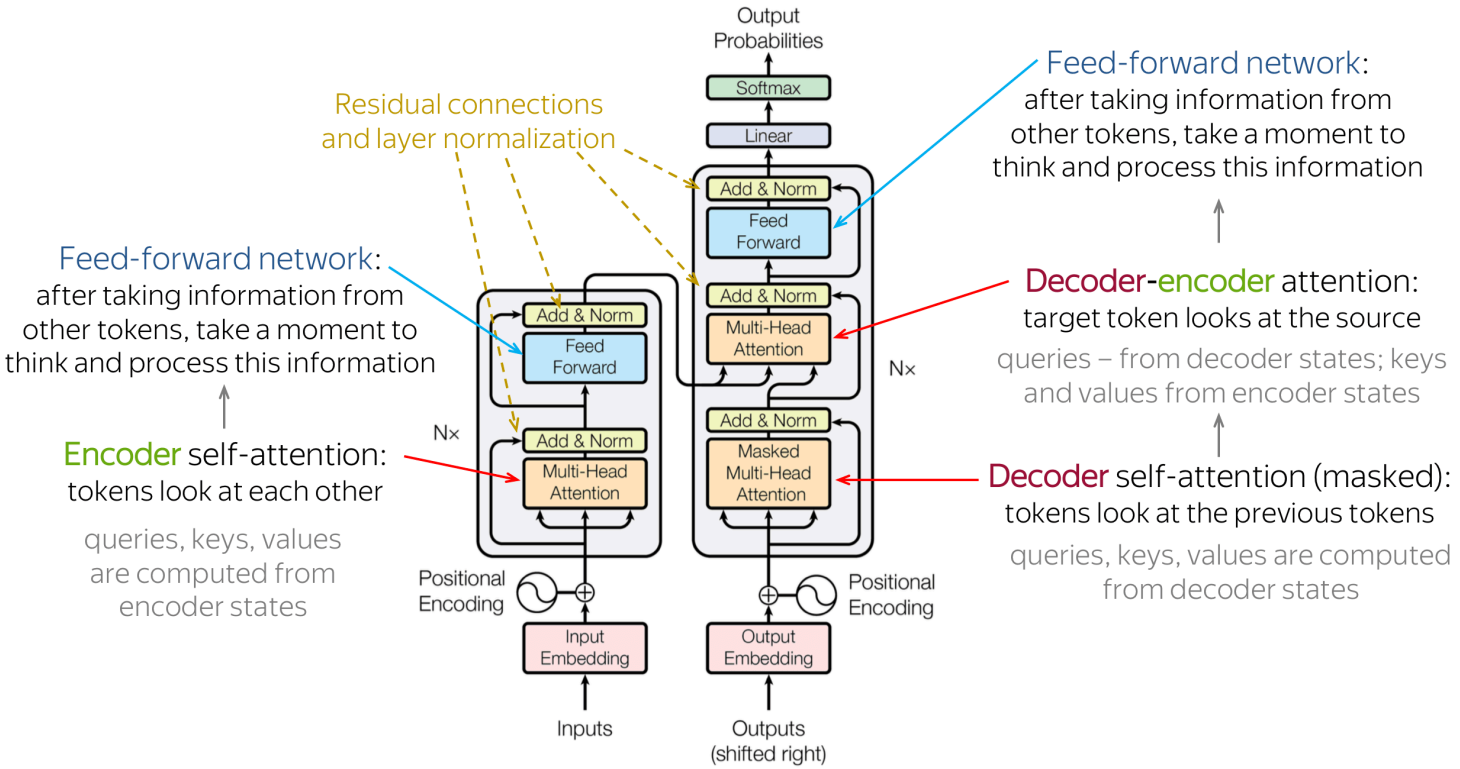
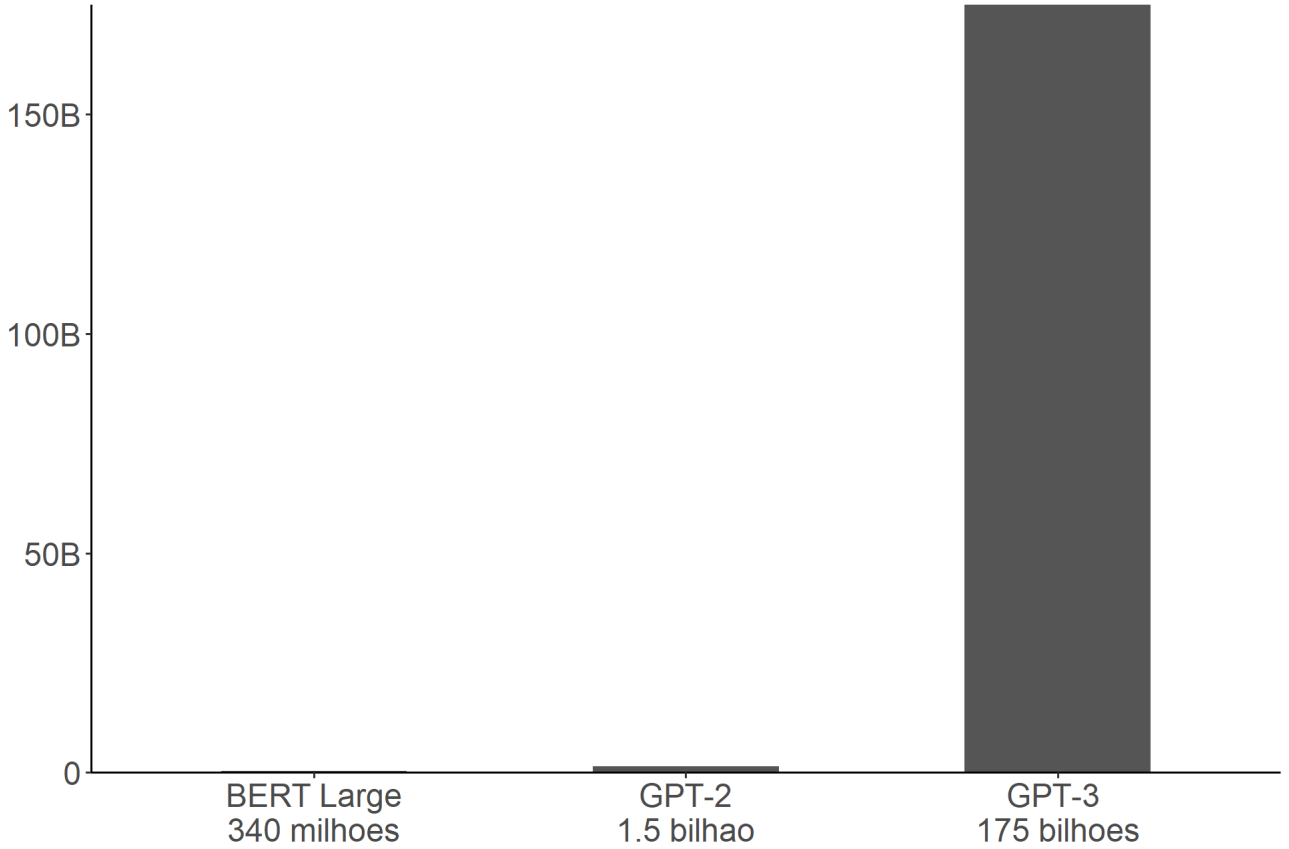


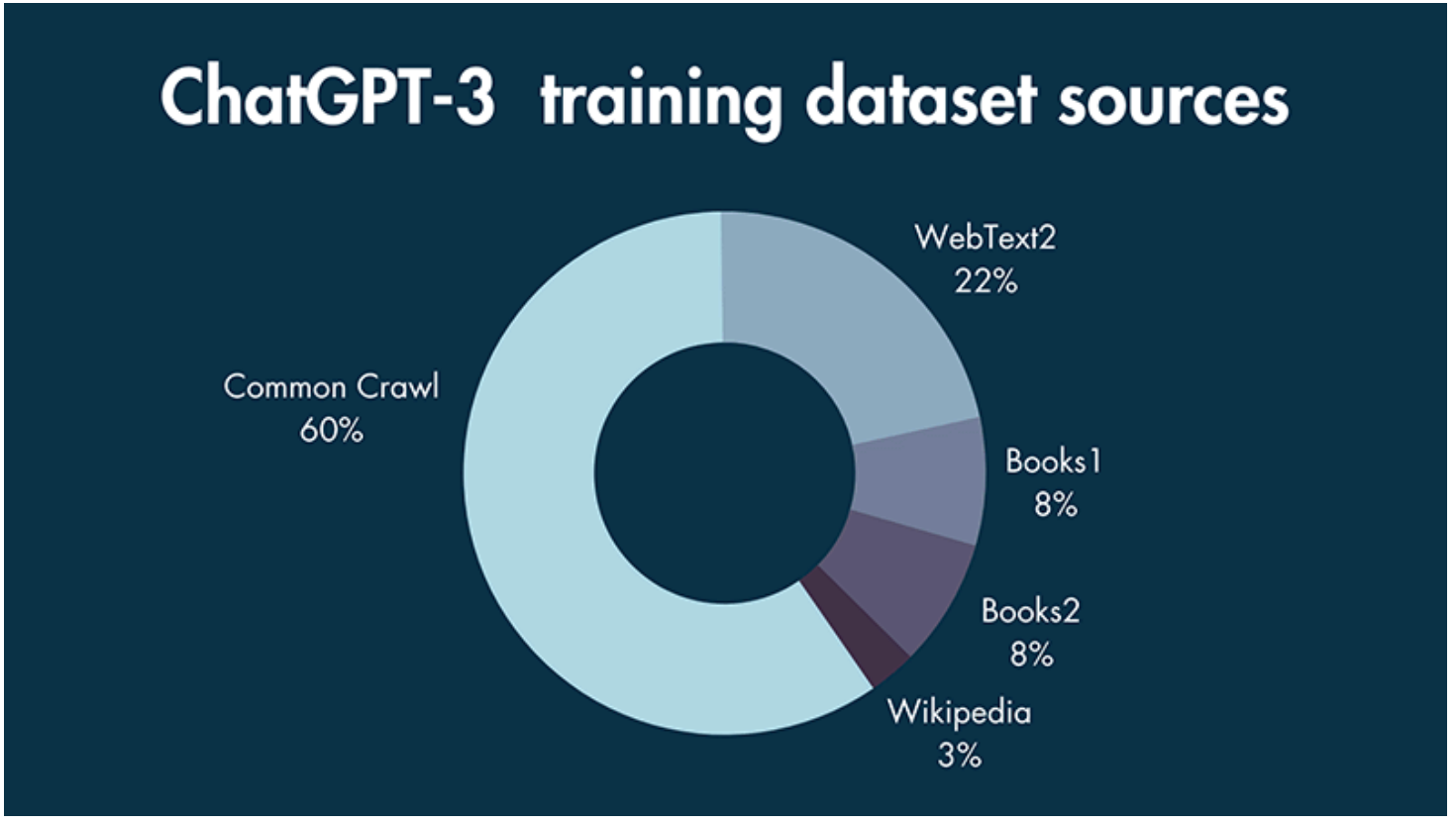
Figura adaptada do paper Attention is all you need, Vaswani et al (2017).

Número de parâmetros dos modelos

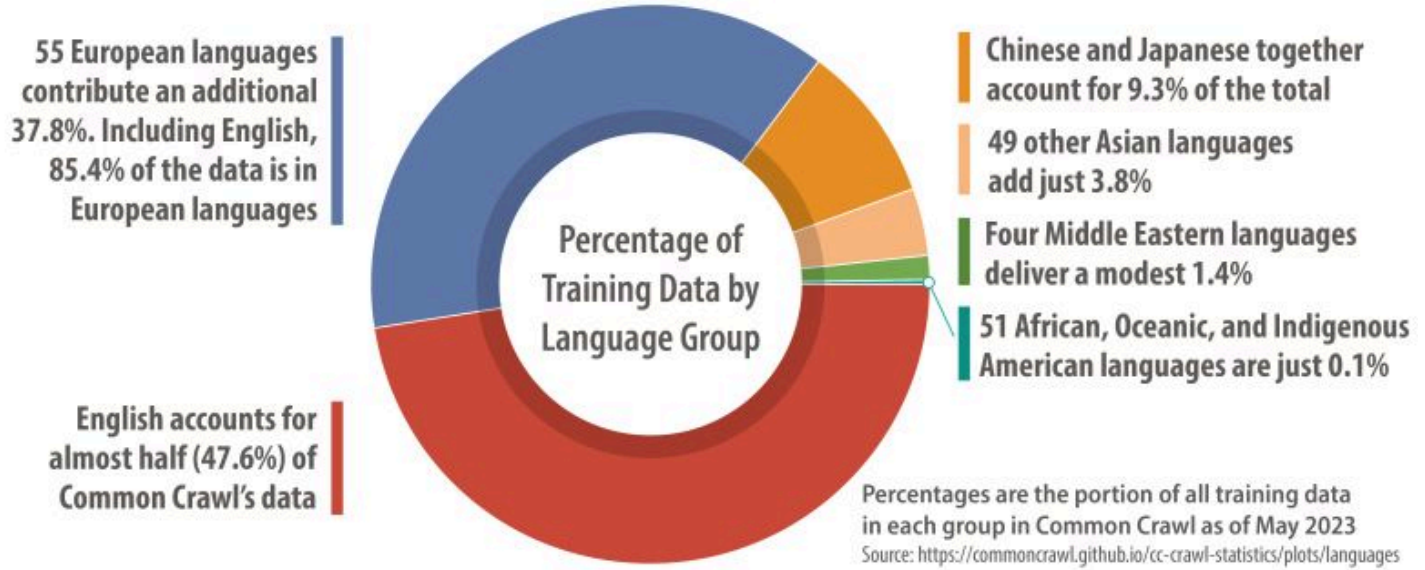


Large Language Models - Benefícios

- Adaptabilidade: modelos podem se adaptar para diferentes tarefas.
- Fine-tuning: necessida uma pequena quantidade de dados para relaizar uma tarefa específica.
- Melhoria contínua: melhores conjuntos de dados e modelos mais otimizados.


ChatGPT-3 training dataset sources
WebText2
22%
Books1
8%
Books2
8%
Wikipedia
3%
Common Crawl
60%

GenAI Training Data Reflects a Strong English and European Bias



© CSA Research, "The Ethics of Generative AI," June 2023

E no Brasil?

- MariTalk é um chatbot baseado em LLM e treinado para atender as necessidades do Brasil.
- Treinada em uma grande quantidade de dados públicos da internet, em sua maioria no idioma Português.
- O modelo também possui capacidades razoáveis de desempenhar tarefas em idiomas como Inglês e Espanhol.
- Acesso através do link: https://chat.maritaca.ai/

Alucinações da LLMs

Definição: fenômeno onde o modelo gera informações incorretas, enganosas ou completamente inventadas, que não são suportadas pelos dados de treinamento.

Causas Comuns de Alucinação:

- Dados de Treinamento Insuficientes ou de Má Qualidade: o modelo pode ser treinado em dados que contêm erros ou informações não verificadas.
- Limitações do Modelo: mesmo modelos avançados podem ter dificuldade em distinguir entre informações factuais e fictícias.
- Ambiguidade do Contexto: o modelo pode não ter contexto suficiente para fornecer uma resposta precisa.

Alucinações da LLMs

Consequências:

- Disseminação de Informação Errada: Pode levar à propagação de desinformação, especialmente em aplicativos críticos como saúde ou finanças.
- Perda de Confiança: Usuários podem perder confiança no sistema se encontrarem informações incorretas repetidamente.
- Tomada de Decisão Prejudicada: Decisões baseadas em informações alucinadas podem ter impactos negativos significativos.

Conhecimento Geral *versus* Conhecimento Específico

Métodos para Enriquecer LLMs

Fine Tuning

- Ajuste fino do modelo pré-treinado usando um conjunto de dados específico para melhorar o desempenho em tarefas específicas.
- Exemplo: Treinar o modelo com dados internos da empresa para personalização.

RAG (Retrieval Augmented Generation)

- Combina recuperação de informações e geração de texto para fornecer respostas mais precisas.
- Exemplo: Buscar dados internos antes de gerar a resposta.

Engenharia de Prompt

- Criar instruções detalhadas para guiar o modelo em suas respostas.
- Exemplo: Adicionar contexto específico ao prompt para obter respostas mais precisas.

Fine Tuning

- É o processo de ajustar um modelo pré-treinado para uma tarefa específica.
- É feito utilizando um conjunto de dados menor e mais específico.
- Exige um investimento cuidadoso de tempo e esforço.
- Exemplo: adaptar um modelo de linguagem geral para entender termos técnicos de uma indústria.
- Recomendação: tentar obter bons resultados com engenharia de prompts.

RAG (Retrieval Augmented Generation)

- A RAG estende os já poderosos recursos dos LLMs para domínios específicos ou para a base de conhecimento interna de uma organização sem a necessidade de treinar novamente o modelo.
- É uma abordagem econômica para melhorar a produção do LLM, de forma que ele permaneça relevante, preciso e útil em vários contextos.
- O uso da tecnologia RAG traz diversas vantagens para as iniciativas de IA generativa de uma organização:
    - Implementação econômica
    - Informações atualizadas
    - Maior confiança de usuários
    - Maior controle na etapa de desenvolvimento

Engenharia de Prompt

- **Prompt**: o texto que é enviado na interação com o modelo de LLM.

Algumas abordagens possíveis:

- **Zero-shot prompt**: solicitação de alguma ação ao modelo sem dar nenhuma referência ou contexto.
- **One-shot prompt**: algum exemplo/referência é dado ao modelo antes de solicitar alguma ação.
- **Few-shot prompt**: alguns exemplos/referências (5-10) é dado ao modelo antes de solicitar alguma ação.

Engenharia de Prompt - Dicas

- Prompts bem definidos levam a respostas mais precisas.
- Forneça contexto relevante para orientar a resposta do modelo.
- Divida perguntas complexas em partes menores e mais gerenciáveis.
- Forneça exemplos para ilustrar o que você está pedindo.
- Indique o formato desejado da resposta, se necessário.
- Defina limites de palavras ou caracteres para controlar a extensão da resposta.
- Experimentar prompts incrementais para entender a sensibilidade do modelo.
- Mais detalhes em https://platform.openai.com/docs/guides/prompt-engineering

Estudo de Caso: Clusterização

E se usarmos o ChatGPT (um LLM) para resolver o problema de segmentação de clientes do Pão de Açúcar apresentado anteriormente?

> Olá, eu tenho um problema de negócio que requer a clusterização de clientes. Vou te passar os detalhes do problema e a base de dados, pode me ajudar? Problema de negócio: Aumentar engajamento de usuários com email para aumentar ativação de cartão de crédito. Estratégia de dados: Segmentar clientes para comunicação personalizada. Expectativa: Uma comunicação mais personalizada aumenta engajamento (abertura de emails e cliques) e leva a uma maior ativação de cartões. Vou te passar a base de dados dos clientes. Gostaria que você gerasse os grupos, desse um nome para os grupos e fizesse uma sugestão de comunicação.

Veja a conversa em https://chatgpt.com/share/2721768c-fe20-4baa-b492-3f45b5201c86

Estudo de Caso: Asset Management

Esta é uma solução avançada para analistas financeiros, projetada para fornecer avaliações detalhadas e fundamentadas sobre a performance recente de ativos de forma automatizada. Utilizando dados de mercado atualizados e notícias relevantes, a ferramenta gera análises diárias concisas e claras, evitando redundâncias e focando nos eventos mais impactantes. Para cada ativo, ela contextualiza a situação, faz uma síntese ponderada dos principais acontecimentos e oferece suporte para decisões informadas na gestão de ativos.

Estudo de Caso: Asset Management

```
prompt <- glue::glue("
Ativo: {nome_asset}
Empresa: {empresa}

Lista de títulos de notícias: {noticias_formatadas}

Você está atuando como analista financeiro em {data_fim}, avaliando a per
que registrou uma variação de {performance}% nos últimos {n_dias} dias. S
análise focada nos eventos dos últimos {n_dias} dias apresentados na list
fundamentada sobre a situação atual da empresa {empresa}.

Por favor, siga estas diretrizes:

Inicie a análise com uma visão geral clara da situação.
Mencione a performance do ativo no período.
Destaque os eventos mais recentes e relevantes.
Enfatize os contrastes entre diferentes eventos e conclua com uma síntese
Inclua exclusivamente informações que estão presentes neste prompt.
Evite redundâncias e repetições na resposta.
Estruture a análise em um parágrafo único.

Modelo da análise: 'Com base nas notícias dos últimos {n_dias} dias... (a
")
```

Páginas para Seguir

- https://www.linkedin.com/in/youssef-hosni-b2960b135/
- https://www.linkedin.com/in/diogocortiz/
- https://www.linkedin.com/in/anderson-rocha-br/
- https://www.linkedin.com/company/langchain/about/
- https://www.linkedin.com/company/llamaindex/about/
- https://www.linkedin.com/company/genai-works/about/
- https://www.linkedin.com/company/maritaca-ai/

Discussão

- Nova habilidade: engenharia de prompt.
- Diferenciar conteúdo criado por uma pessoa do criado por IA.
- Questões éticas.
- Autoria.
- Regulação.
- Limites.

Obrigado!

magnotfs@insper.edu.br