Large Language Models (LLM)

FUNDAMENTOS, APLICAÇÕES E PERSPETIVAS FUTURAS

SISTEMAS DE RECOMENDAÇÃO

MESTRADO EM ENGENHARIA INFORMÁTICA

Constantino Martins, Catarina Figueiredo, Dulce Mota e Fátima Rodrigues

Introdução

- Ascensão dos Modelos de Linguagem de Grande Escala (LLM).
- Transformação da interação humano-máquina.
- Impactos transversais: educação, saúde, finanças, indústria.

Inteligência Artificial Generativa (GenAI)

- IA capaz de gerar novos conteúdos: texto, imagem, som, código.
- Baseada em modelos treinados com grandes volumes de dados.
- Exemplos: GPT, Claude, Gemini, Mistral.

Definição de LLM

- Os LLM são sistemas avançados de IA que podem compreender e gerar linguagem natural.
- Modelos de linguagem treinados para gerar e interpretar texto.
- Utilizam arquitetura **Transformer** com mecanismos de **autoatenção**.
- Capazes de realizar tarefas de linguagem natural com alta coerência.

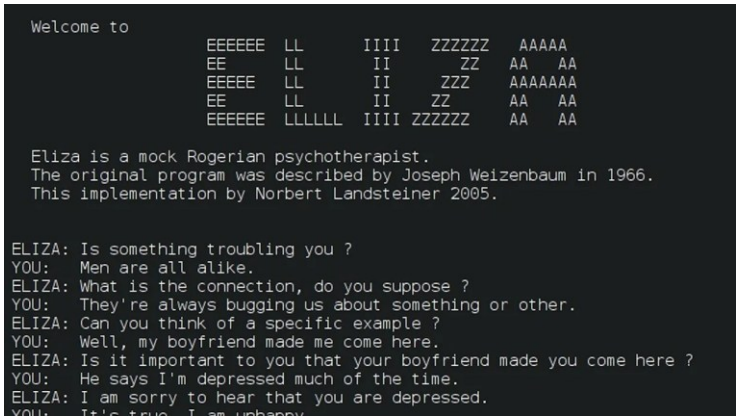
The power of LLM lies in its ability to understand and generate humanlike language.

Evolução dos LLMs

- De ELIZA (1965) aos transformers (2017)
- Surgimento do Word2Vec ao BERT e GPT
- Modelos abertos e comerciais (2023)

word2vec usava uma técnica que permitia ao computador realizar "matemática" com as palavras.
Por exemplo:
Input -> **Lawyer-Law+Medicine=?**
Output -> **Doctor**

✓ **1965** - Joseph Weizenbaum, investigador do MIT construiu o **primeiro chatbot** de IA, conhecido como **ELIZA**. O software era rudimentar e produzia respostas com base nas palavras-chave que detetava no *prompt*. No entanto, quando Weizenbaum programou ELIZA para atuar como psicoterapeuta, as pessoas ficaram supostamente surpresas com o quão convincentes as conversas eram. O trabalho estimulou o interesse crescente no processamento de linguagem natural, inclusive da Agência de Projetos de Pesquisa Avançada de Defesa dos EUA (DARPA), que forneceu financiamento considerável para pesquisas iniciais de IA.



Breve história dos LLM

- Os LLM são um desenvolvimento moderno, mas o estudo do processamento de linguagem natural (PLN) data de 1950, quando Alan Turing lançou o teste de Turing.
- **1950: Alain Turing**: publicou o artigo intitulado "Computing Machinery and Intelligence", no qual colocava a questão "Can machines think?".
  - A resposta a essa pergunta requeria que em primeiro lugar fosse definido os conceitos de "máquina" e "pensar". Em vez disso, Turing propôs um jogo: um observador assistiria a uma conversa entre uma máquina e um humano e tentaria determinar qual era qual. Se não conseguisse fazer isso de forma confiável, a máquina venceria o jogo. Embora isso não provasse que uma máquina estava "a pensar", o **Teste de Turing** — como veio a ser conhecido — tem sido um parâmetro importante para o progresso da IA desde então.

Breve história dos LLM

- **1940:** Warren McCulloch e Walter Pitts introduzem ao mundo a ideia das Redes Neurais Artificiais (ANNs).
- **1950-60s:** Desenvolvimento do primeiro modelo de linguagem (Modelo baseado em regras).
- **1980-90s:** Introdução dos modelos de linguagem baseados em estatísticas. Esses modelos recorrem a mecanismo de predição das palavras seguintes numa frase com base nas palavras anteriores.
- Meados dos anos **2000:** A área de Processamento de Linguagem Natural (NLP) testemunha a introdução de *word embeddings*.
- **2010:** Os modelos transitaram da determinação da ordem das palavras para uma compreensão mais profunda da representação e do significado das palavras, com suporte em Redes Neurais Recorrentes (RNN).

Breve história dos LLM

- Em **2018**, oito cientistas da Google escreveram e publicaram um estudo de referência intitulado "*Attention is All You Need*", sobre aprendizagem automática onde apresentaram a arquitetura do **transformador**, um framework inovador de redes neuronais que pode gerir e compreender informações textuais complexas com maior precisão e em escala.
- **2018**: Introdução do modelo BERT (baseado na arquitetura transformer).
- **2018**: Lançamento do GPT-1 (modelo baseado na arquitetura transformer).

Breve história dos LLM

- **2019:** Desenvolvimento do GPT-2 (modelo baseado na arquitetura transformer).
- A NVIDIA produziu o modelo Megatron-LM (baseado na arquitetura transformer).
- **2020:** Introdução do GPT-3 (modelo baseado na arquitetura transformer).
- **2023:** Lançamento do GPT-4 (modelo baseado na arquitetura transformer).
- **2023:** Claude (Anthropic) - Foco em IA alinhada com valores humanos — treinado com “constitutional AI”.
- **2023:** Gemini (Google DeepMind) - Modelo multimodal, concorrente direto do GPT-4.

Breve história dos LLM

- **2023:** Mistral 7B / Mixtral - Modelos open-source de alta eficiência com desempenho competitivo.
- **2024:** GPT-4 Turbo (OpenAI) Versão mais barata, rápida e com “memória longa” integrada.
- **2024:** Claude 3 (Anthropic) Supera benchmarks em várias tarefas de linguagem e multimodalidade.
- **2024:** Command R+ (Cohere) Especializado em recuperação aumentada (RAG) — integração com documentos.
- **2025:** Modelos Multimodais Unificados (em curso) - Fusão de texto, imagem, áudio, vídeo e ação — evolução rumo à AGI.

O que é o BERT?

- BERT = Bidirectional Encoder Representations from Transformers.
- Modelo de codificação criado pela Google em 2018.
- Lê o texto em ambas as direções para melhor compreensão do contexto.
- Baseado apenas na parte codificadora da arquitetura Transformer.

Como é o BERT treinado?

- Pré-treino com duas tarefas principais:
  - Masked Language Modeling (MLM): prever palavras ocultas em frases.
  - Next Sentence Prediction (NSP): prever se uma frase segue logicamente outra.
- Torna o BERT eficaz em tarefas de compreensão de linguagem natural.

Aplicações e impacto do BERT

- Revolucionou o processamento de linguagem natural.
- Usado em: classificação de texto, resposta a perguntas, análise semântica.
- Superou recordes em benchmarks de NLP em 2018.
- Inspirou modelos como RoBERTa, DistilBERT, ALBERT, entre outros.

Modelo de Transformador

- Os transformadores são constituídos por dois componentes principais:
  - **codificadores**: recebem dados não processados de texto transformando-os em elementos discretos que podem ser analisados pelo modelo.
  - **descodificadores**: processam esses dados através de uma série de camadas para produzir o resultado final, que pode, por exemplo, consistir numa frase gerada.
- Trabalham frequentemente em conjunto para processar e gerar sequências.

Como funcionam os LLM

- Tokenização e embeddings
- Pré-treino e fine-tuning
- Autoatenção e representação contextual

Autoatenção - permite ao modelo concentrar-se em diferentes partes de uma sequência de texto e ponderar dinamicamente o valor da informação relativamente a outros tokens na sequência, independentemente da sua posição. Contribui para que os LLM capturem as intrincadas dependências, relações e nuances contextuais da linguagem escrita. É um componente-chave da arquitetura do transformador.

Fine-tuning é o processo de ajustar um modelo de linguagem já pré-treinado (como o GPT, BERT, etc.) usando um conjunto de dados adicional e específico para uma tarefa ou domínio particular.

Componentes Técnicos dos LLM

- Corpus de dados e pré-processamento
- Embedding posicional e contextual
- Modelos autorregressivos vs mascarados

Modelos autorregressivos são modelos que aprendem a prever a próxima palavra (ou token) numa sequência, dado o contexto anterior. (caso GPT)

Modelos mascarados são modelos que aprendem a prever palavras em falta no meio de uma frase, com acesso bidirecional ao contexto (antes e depois da máscara). (caso BERT)

Pré-treino e Fine-tuning

- Pré-treino: aprendizagem geral da linguagem
- Fine-tuning: adaptação a tarefas específicas
- Técnicas: Seq2Seq, *Contrastive Learning*, aprendizagem por reforço com feedback humano

Contrastive Learning - modelo aprende aproximando representações semelhantes e afastando representações diferentes.
Dados pares positivos (ex: duas frases com o mesmo significado) → o modelo aproxima as suas representações.
Dados pares negativos (ex: frases sem relação) → o modelo afasta as suas representações no espaço vetorial.

Aplicações Práticas

- Assistentes virtuais e chatbots corporativos
- Análise de documentos e extração de informação
- Geração de relatórios, apoio a auditoria, classificação inteligente

LLM em Produção

- Bancos, centros de apoio ao cliente, sistemas jurídicos
- Aplicações na medicina, marketing e educação
- Casos em produção são crescentes e promissores

Desafios e Riscos

- Vieses, alucinações e explicabilidade limitada
- Privacidade, segurança e propriedade intelectual
- Consumo energético e impacto ambiental

Validação de LLMs

- Abordagens multidimensionais de validação
- Métricas quantitativas e avaliação humana
- Interpretação, mitigação de viés, monitorização contínua

Tendências Futuras

- LLM multimodais (texto, imagem, áudio, vídeo)
- Modelos de menor escala e mais eficientes (SLM- *Small Language Models*)
- Rumo à Inteligência Artificial Geral (AGI - *Artificial General Intelligence*)

Inteligência Artificial Geral – modelo de inteligência artificial com a capacidade de aprender, entender e realizar qualquer tarefa cognitiva que um ser humano é capaz de executar, com nível de raciocínio, adaptação e compreensão geral equivalentes aos de uma mente humana.

Em síntese

- LLM representam um avanço paradigmático
- Desafios técnicos, éticos e sociais ainda por resolver
- Necessidade de governança robusta e validação contínua

Estudo: Mais Humanos do que os Humanos?

Resumo do artigo sobre narrativas geradas por LLM

Zhao et al., C&C 2023

Objetivo do Estudo

- Comparar narrativas geradas apenas por LLM (*non-interleaved*) vs. narrativas intercaladas entre humanos e LLM (interleaved).
- Hipótese inicial: colaboração humana é capaz de melhor a qualidade da história.

Metodologia

- 20 histórias geradas (10 interleaved e 10 non-interleaved)
- Tópicos: carros elétricos, praias e culinária
- Dois inquéritos (~500 participantes cada): avaliação individual e por comparação de pares

Critérios de Avaliação

- Preferência geral
- Falhas lógicas
- Plausibilidade
- Compreensão
- Originalidade / novidade

Resultados Principais

- Histórias apenas com LLM foram melhor avaliadas
- Mais claras, com menos falhas lógicas
- Histórias *interleaved* foram vistas como mais inovadoras
- Problemas de fluidez e redundância na escrita *interleaved*

Discussão

- Problema: LLM respondia, mas humano não ajustava as suas frases
- Gerava transições forçadas e repetitivas
- Coordenação limitada prejudicou a narrativa

Conclusão Geral do artigo

- LLM, operando sozinhos, criaram histórias mais coesas e preferidas.
- Colaboração humano-LLM requer melhor integração adaptativa.
- Futuras investigações irão explorar interações mais inteligentes entre LLM e humanos.

Impacto se fosse usado GPT-4 no estudo

- O estudo original usou GPT-3.5.
- O uso de GPT-4 traria melhorias significativas na qualidade narrativa *interleaved*.

1. Coerência narrativa

- GPT-4 é mais eficaz em manter coerência de longo prazo.
- *Interleaved stories* com GPT-4 teriam transições mais naturais e menos redundância.

2. Compreensão de contexto

- GPT-4 compreende melhor o fluxo do discurso.
- Evita repetições e consegue responder com maior relevância ao texto humano.

3. Criatividade e realismo

- GPT-4 apresenta equilíbrio entre originalidade e plausibilidade.
- Histórias *interleaved* seriam mais ricas e menos propensas a alucinações.

4. Adaptação de estilo

- GPT-4 adapta-se melhor ao tom e vocabulário humano.
- Maior homogeneidade no estilo da narrativa conjunta.

Síntese

- Com GPT-4, a diferença entre *interleaved* e *non-interleaved* seria reduzida.
- Histórias *interleaved* poderiam igualar ou superar as geradas apenas pelo LLM.
- O design da colaboração continua a ser um fator crítico.

Sistemas de Recomendação Convencionais e LLM

Limitações dos Sistemas de Recomendação Convencionais

- Falta de conhecimento de mundo aberto → dificuldade em interpretar intenções complexas
- Baseados em dados históricos → baixa capacidade de generalização
- Modelação implícita das preferências → recomendações pouco explicáveis
- Interação passiva com o utilizador → pouca personalização em tempo real

Beneficios dos LLM nos Sistemas de Recomendação

• Feature Engineering → enriquecimento semântico de dados

• Feature Encoding → representações contextuais ricas para utilizadores/itens

• Scoring/Ranking → classificação mais precisa baseada em linguagem e contexto

• Interação com o utilizador → recomendações conversacionais, explicações naturais

• Controlo do pipeline → LLM como coordenador da lógica de recomendação

Benefícios Chave da Integração com LLM

CONHECIMENTO ABERTO → RECOMENDAÇÕES INTELIGENTES COM POUCOS DADOS

INTERAÇÃO NATURAL → PERSONALIZAÇÃO ATRAVÉS DE LINGUAGEM NATURAL

RACIOCÍNIO E COMPREENSÃO → INFERÊNCIA DE INTENÇÕES E SENTIMENTOS

MULTITAREFA → ATUAÇÃO EM VÁRIAS FASES DO SISTEMA DE RECOMENDAÇÃO

INTEGRAÇÃO MODULAR → ACOPLAMENTO PARCIAL OU TOTAL DO LLM AO SISTEMA

Vantagens para os Sistemas de Recomendação

- Compreensão semântica profunda de texto→ Melhor interpretação de preferências e opiniões
- Análise conjunta de imagem e texto → Recomendação visualmente e semanticamente relevante
- Processamento de áudio → Sugestões musicais/podcasts com base em padrões auditivos
- Interpretação de vídeo → Classificação e recomendação de conteúdos visuais complexos
- Interação multimodal com o utilizador → Pode explicar recomendações com base em múltiplos sinais (ex: "porque viste isto e ouviste aquilo...")

Sistemas multimodais – são sistemas que conseguem analisar, integrar e raciocinar sobre diferentes tipos de dados: Texto + Imagem + Áudio + Vídeo + Comportamento

Tendências e Oportunidades

- Recomendações personalizadas com base em várias fontes simultâneas.
- Conversação multimodal (ex: descrever imagem + sugerir algo semelhante).
- Geração de conteúdo personalizado (ex: trailers, resumos, playlists).
- *Explainable AI* (XAI): explicações multimodais que combinam texto e imagem.

Estudo: Artigo

- *“Leveraging Large Language Models for Pre-trained Recommender Systems“,* (Chu, Z. & al., 2023)
- Proposta: RecSysLLM — um LLM adaptado para sistemas de recomendação.
- Objetivo: combinar capacidades linguísticas com tarefas específicas de recomendação.

Desafios Identificados

- Modelos tradicionais não generalizam bem.
- LLM genéricos não compreendem dados estruturados de recomendação.
- Fine-tuning direto pode comprometer o desempenho linguístico dos LLMs.

Solução: RecSysLLM

- Textualização de dados tabulares (utilizadores, itens, sequências).
- Máscaras e codificações posicionais especializadas.
- Treino multitarefa com prompts específicos para recomendação.
- Adaptação leve com LoRA.

Tarefas Realizadas

- Previsão de avaliações (ratings)
- Recomendação sequencial
- Geração de explicações
- Resumo de críticas
- Recomendação direta com linguagem natural

Resultados Experimentais

- Avaliação em Amazon e Alipay.
- Supera GPT-4, ChatGPT e P5 em várias métricas.
- Boas capacidades zero-shot em prompts não vistos.

Conclusão do artigo

- RecSysLLM permite integrar LLMs em pipelines de recomendação.
- Melhora a flexibilidade e precisão das recomendações.
- Representa um avanço em modelos unificados de recomendação multimodal.

Comparação: LLM Genérico vs RecSysLLM

- Capacidade de linguagem natural - Alta em ambos
- Compreensão de histórico de utilizador - Fraca no LLM genérico, forte no RecSysLLM
- Aptidão para recomendação sequencial - Limitada no LLM genérico, otimizada no RecSysLLM
- Integração de dados estruturados - Inexistente no LLM genérico, incorporada no RecSysLLM
- Eficiência de afinação - Requer recursos altos vs afinação leve com LoRA
- Geração explicável de recomendações - Genérico é limitado, RecSysLLM é especializado