

Unboxing AI: uma introdução

...

Como escolher os LMM, questionar e utilizá-los na sua investigação, de forma eficaz!



universidade de aveiro



DiGMedia
Digital Media and Interaction
Research Centre

Sumário

- ★ As origens dos Large Language Models (LLMs);
- ★ O que é e como funciona um LLM; [Questões?](#)
- ★ Breve introdução ao Prompt Engineering (sem código);
- ★ Uso de LLMs na investigação; [Questões?](#)
- ★ LLMs locais versus na nuvem;
- ★ Aplicações e recursos úteis; [Questões?](#)
- ★ Criação de... 🤔



As origens dos Large Language Models (LLMs)



universidade de aveiro



DigiMedia
digital media and interaction
research centre

LLMs e a teoria da comunicação ;) - i

Apesar da enorme popularidade alcançada pelo ChatGPT — cuja ascensão teve início a 30 de novembro de 2022 — a inteligência artificial, e em particular os modelos de linguagem, têm raízes que remontam, de certa forma, a **1948**, com o trabalho do "nosso" já conhecido **Claude Shannon***, no célebre livro *A Mathematical Theory of Communication* [1],[2].

*Shannon para os Engenheiros 😊 - Shannon-Weaver para a Comunicação



LLMs e a teoria da comunicação ;) - ii

Nesse livro, Shannon propõe um modelo matemático influenciado pelos processos de **Markov**, assumindo a existência de um número finito de estados possíveis e um conjunto de probabilidades associadas às transições entre esses estados [1].

Ou seja, se o sistema se encontra no estado Y, a probabilidade de transitar para o estado Z é determinada por essas regras.

Este tipo de abordagem está presente, ainda hoje, em funcionalidades como os sistemas de previsão de texto dos telemóveis, que tentam antecipar a palavra que o utilizador quer escrever com base nos caracteres já introduzidos [1].



LLMs - seq2seq

Uma abordagem mais recente, que se afasta dos modelos de Markov, é a chamada *sequence-to-sequence* (ou seq2seq).

Por processar o texto token a token, esta arquitetura é, em teoria, excelente para tarefas como tradução automática, que exigem a compreensão do contexto ao longo de uma sequência [1].

No entanto, à medida que o número de tokens aumenta, pode ocorrer uma perda gradual de informação, o chamado "vanishing gradient problem", que afeta a capacidade do modelo para reter informação em sequências mais longas [3].



LLMs - o mecanismo de *atenção*

O mecanismo de atenção (attention mechanism) foi desenvolvido para ultrapassar os problemas de "esquecimento" do seq2seq.

Este sistema associa a cada token um valor escondido, que é lido e interpretado pelo decodificador durante a fase de geração de texto [4].

Esta inovação deu origem ao "transformador" (Transformer), que, como veremos, revolucionou os LLMs.



LLMs - Os Transformers 🤖

Os *transformers* (não, não os de Cybertron 😊) fazem a ponte entre as arquiteturas previamente mencionadas, inicialmente mantendo a mesma estrutura básica (codificador e decodificador) mas utilizando unicamente mecanismos de atenção [1].

Com a introdução dos modelos *Generative Pre-Trained* (GPT) a arquitetura do transformador foi simplificada, deixando cair o codificador. Isto permite ao modelo duas coisas: concentrar-se no texto de forma não linear (como nos modelos seq2seq) e ser treinado com dados brutos (raw data), posteriormente ajustados através de *fine-tuning* [5].

LLMs - a evolução dos Transformers

Apesar dos modelos GPT serem **generalistas** por natureza, demonstraram ser capazes de atingir resultados superiores em tarefas específicas, como a tradução, quando comparados a modelos treinados exclusivamente para uma única tarefa (neste exemplo, a tradução) [1].

Isto ocorre porque os modelos GPT "aprendem" através de grandes volumes de dados de diversas tipologias, o que lhes permite capturar padrões complexos e adaptar-se a uma vasta gama de contextos. Esta versatilidade é uma das razões pelas quais se tornaram tão populares para aplicações de processamento de linguagem natural [6].



O que é e como funciona um LLM



universidade de aveiro



DigiMedia
digital media and interaction
research centre

O que é e como funciona um LLM - O que são?

Nas "suas" próprias palavras:

O objetivo principal de um LLM (Modelo de Linguagem de Grande Escala) é gerar e compreender texto de forma semelhante à linguagem humana, com base em padrões aprendidos a partir de grandes volumes de dados.

Mais concretamente:

Um LLM tem como objetivo prever a palavra (ou palavras) mais provável(is) que se segue(m) numa sequência de texto, o que lhe permite realizar uma vasta gama de tarefas relacionadas com a linguagem — como responder a perguntas, traduzir textos, resumir conteúdos, gerar código ou manter conversas.

Apesar de a tarefa central ser a previsão da próxima palavra, essa capacidade serve de base para muitas aplicações em compreensão e geração de linguagem natural.

Questão colocada em 05/05/25



universidade de aveiro



DiGeMedia
Digital Media and Interaction
Research Centre

O que é e como funciona um LLM - ii

Popularmente poderíamos dizer:

"Entra texto, sai texto". A menos que seja um modelo multimodal 🤔

O utilizador dá como entrada um *prompt* e receberá, como saída, a *completion* ou *response*.

Mas claro, para que o modelo compreenda, o texto será convertido para tokens ou unidades de linguagem (que não correspondem necessariamente a caracteres, podem ser partes de palavras ou mesmo palavras inteiras. Iremos ver melhor na secção "Aplicações e recursos úteis").



O que é e como funciona um LLM - ii

Entra texto → dá-se a transformação em **tokens** → tokens são **embedded**, ou seja tornam-se representações numéricas (vetores) que capturam padrões e relações com outros tokens, baseados no seu contexto de uso → o **transformer** utiliza o mecanismo de atenção para analisar as relações entre os tokens e assim prever o próximo token, considerando o contexto da frase/global → modelo aplica os seus pesos ajustados (após **treino** e fine tuning) para prever o token seguinte, até obter a frase completa → sai texto!



O que é e como funciona um LLM - Treino i

Na imagem, podemos ver a distribuição percentual e por tipologia dos dados que podem ser utilizados para treinar um LLM. Neste caso, o conjunto/dataset "The Pile" contém (mais de) 800GB de informação [6].

Durante o treino, o modelo aprende a responder de forma similar ao material que utilizou, ou seja, se parte do treino incluir textos da PubMed Central, é provável que o modelo "responda" usando termos técnicos e raciocínios próprios dos artigos científicos [1].

No entanto, o objetivo não é apenas repetir o que "leu", mas sim aprender padrões de lógica e raciocínio, aplicando-os de forma mais flexível e adaptada ao contexto [1].

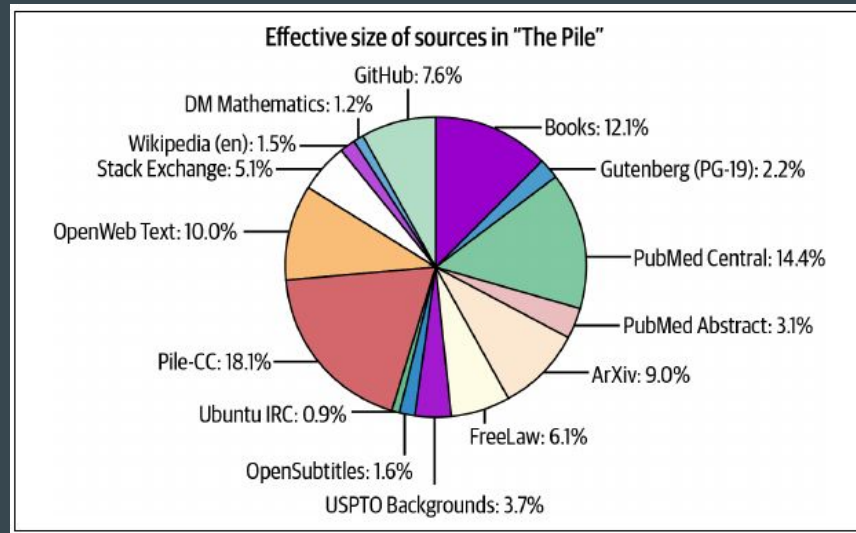


Figura 1: Effective size of sources in "The Pile"

observed: 10%

Fonte:[6]

O que é e como funciona um LLM - Treino ii

Já todos ouvimos:

"Modelo X é melhor para programação, enquanto Modelo Y é melhor para escrita criativa."

Porquê? Pelo tipo de materiais e seu peso/impacto no treino do LLM.



O que é e como funciona um LLM - Humano vs LLM

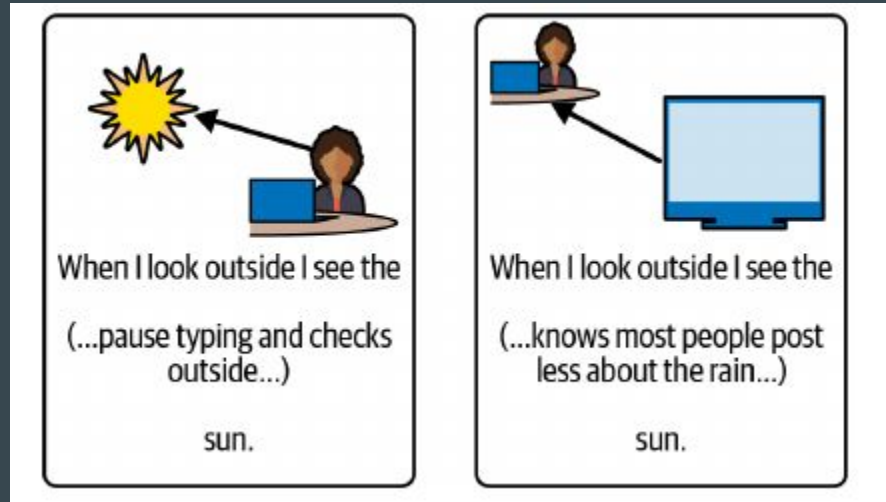


Figura 2: Humano VS LLM
Fonte: [2, p.21]



O que é e como funciona um LLM - Humano vs LLM

O Humano pode expressar incerteza, pode consultar fontes (livros, internet, LLMs 😊) e pode adaptar-se ao contexto/ a um contexto fluido.

Um LLM, na sua forma mais básica e sem recursos a ferramentas (como o RARG [7]) ou a fine-tuning à posteriori, não pode[1]:

- pesquisar, limitando-se a adivinhar;
- expressar dúvida;
- admitir que a sua resposta é apenas um palpite (tudo na sua resposta é, de facto, o token/palavra seguinte mais provável... ou seja, tudo é um palpite).



O que é e como funciona um LLM - "erros"

Isto leva-nos às **halucinações**, ou seja informação factualmente errada (links que não existem, entre outros) mas que são plausíveis e, para o LLM, elas não diferem de outras respostas [1].

Halucinações podem ser induzidas, por exemplo, questionando os modelos e introduzindo no *prompt* informação falsa/inexistente, mas que o modelo assumirá como verdadeira. Isto porque o modelo irá assumir que o *prompt* é verdadeiro - viés de verdade (*truth bias*) [1].

(com a possibilidade de pesquisar na *web* os modelos são mais resistentes)

A abordagem mais cautelosa é verificarmos sempre tudo, pedir ao modelo que explique o seu raciocínio, ou que forneça dados/links/valores que possam ser verificados por nós.



O que é e como funciona um LLM - Round 2 - i

Como é que o cérebro consegue ler esta frase? [8]

According to a research study at Cambridge University, it doesn't matter in what order the letters in a word are, the only important thing is that the first and last letter be in the right place. The rest can be a total mess and you can still read it without problem. This is because the human mind does not read every letter by itself, but the word as a whole.

Figura 3: "According to..." Fonte: [8]

Sabem quem mais não se preocupa com erros (typos) e não lê letra a letra?



O que é e como funciona um LLM - Round 2 - ii

Tal como nós, um LLM **não** lê letra a letra mas sim token a token. Só que [1],[9]:

- Os seus tokens são determinísticos;
- Não pode desacelerar e examinar letra a letra (implica partir os tokens e isso pode levantar problemas);
- Texto capitalizado ou acentuado tem um custo superior (em processamento etc);
- Está limitado a uma janela contextual, otimizada para a língua inglesa (que atualmente equivale a várias páginas A4);



O que é e como funciona um LLM - Round 2 - iii

- Um token de cada vez, auto-regressivo e sem hipótese de corrigir um erro que surja (excepto se introduzirmos aqui RARG ou outros mecanismos pós-processamento);
- Auto-regressão pode originar padrões que levarão a acumulação de erros ou repetições.



O que é e como funciona um LLM - O Transformer i

É tempo de irmos ao Teatro!!! 🎭

Imaginem uma peça, onde:

Guiões: São os **tokens**, ou seja, as palavras ou subpalavras que compõem o texto.

Atores: Representam os **blocos de processamento do Transformer**, como as cabeças de atenção e as camadas feed-forward.

(adaptado de [1], onde o autor utilizava uma analogia com mini-cérebros)



O que é e como funciona um LLM - O Transformer ii

Cada ator:

- Recebe um guião (**token**) e a sua posição na peça (**positional encoding**);
- Durante os ensaios (**passagens pelas camadas (layers/steps) do Transformer**), interage com outros atores para compreender o contexto do guião;
- Utiliza o **mecanismo de atenção** para decidir a quem deve prestar mais atenção, atribuindo diferentes pesos às informações recebidas;
- Após várias rondas de ensaio, ajusta a sua performance com base nas interações e feedback recebidos;
- No final, contribui para a próxima linha do guião, **prevendo** o próximo token na sequência.

Este processo é repetido até que a peça (sequência de texto) esteja completa.

O que é e como funciona um LLM - O Transformer iii

A fim de garantir que a peça seja apresentada de forma coerente e sem "spoilers", implementa-se o masking causal:

- Restrições de Comunicação: Cada ator só pode interagir com os colegas que já se apresentaram (à sua esquerda). Isso impede que um ator obtenha informações de cenas futuras;
- Fluxo de Informação: A informação flui da esquerda para a direita, refletindo a natureza sequencial da linguagem.

(e de baixo para o topo mas nesta parte a analogia falha...)



O que é e como funciona um LLM - O Transformer iv

Durante os ensaios:

- Cada ator formula perguntas (**queries**) sobre o contexto do seu guião;
- Os atores que já se apresentaram (os à esquerda) compartilham informações relevantes (**keys e values**).
- Cada ator calcula a relevância das informações recebidas, atribuindo pesos diferentes e atualizando a sua compreensão do contexto (atenção seletiva).

Tudo isto permite que cada ator (token) compreenda o contexto global da peça (sequência de texto), ajustando assim a sua performance (representação).



Breve introdução ao Prompt Engineering (sem código)



universidade de aveiro



DigiMedia
digital media and interaction
research centre

Breve introdução ao Prompt Engineering(BiPE) (sem código)

Agora que compreendemos um pouco melhor os LLMs, podemos ver que:

- a ordem como fazemos as perguntas importa (por exemplo o "clássico" contar vogais ou letras);
- Erros tipográficos, caracteres estranhos e afins acrescentam "entropia" ao nosso *prompt* e devemos considerar se são necessários ou se podemos/devemos fazer uma "limpeza" inicial à nossa questão.



BiPE - Temperatura ☀☀☀ - i

Quando dizemos que um LLM apresenta um token que considera ser o mais provável, fica implícito que, muitos outros tokens foram considerados.

Mas e se não quisermos o token mais provável?

Podemos alterar um parâmetro, chamado **Temperatura**. Este valor, que é maior que 0, diz quão criativo deixamos o modelo ser [10].



BiPE - Temperatura ☀️☀️☀️ - ii


Temperatura	Comportamento
0	O modelo seleciona sempre o token com a maior probabilidade. Respostas altamente previsíveis e repetitivas.
0.5-07	Introduz alguma aleatoriedade, permitindo respostas mais variadas, mas mantendo a coerência.
1	A distribuição de probabilidade original é mantida, há um equilíbrio entre criatividade e precisão.
>1	Aumenta significativamente a aleatoriedade, podendo gerar respostas mais criativas, mas também mais propensas a erros ou incoerências.

Quanto maior a temperatura e mais longa for a conversa, mais se notará o efeito cumulativo dos erros [10] (termo técnico, em latim, deverá ser *modelus borrachos* 🤪)

BiPE - exemplos

The eating in excess chocolate can lead to:

Temperature 0

 **Assistant** 03:32:22 pm


Eating excessive chocolate, especially if it is high in sugar and fat, can have several potential negative effects on your health. Here are some consequences that may occur:

1.

Weight gain: Chocolate contains calories, so consuming too much can contribute to weight gain, particularly if it's rich in added sugars or heavy with cream.

The eating in excess chocolate can lead to:

Temperature 0.6

 **Assistant** 03:57:41 pm

Eating excessive chocolate, particularly chocolate that is high in sugar and saturated fat, can have several potential negative effects on your health. Here are a few consequences you might experience:

1.

Weight gain: Chocolate, especially milk chocolate, contains calories. If consumed in large amounts, it can contribute to weight gain or hinder weight loss efforts.

The eating in excess chocolate can lead to:

Temperature 1

 **Assistant** 04:00:25 pm


Eating excessive chocolate can potentially lead to several health issues. Here are some common consequences:

1.

Weight Gain/Obesity: Chocolate is high in calories, especially if it's rich in sugar and fat. Consuming too much of it without controlling portion sizes can contribute to weight gain.

The eating in excess chocolate can lead to:

Temperature 2

 **Assistant** 04:03:47 pm

Eating too much chocolate, especially high-quality or dark varieties that are rich in cocoa solids, can have various effects on the body. Here's a list of consequences you might experience:

1.

Weight gain: Chocolate contains calories, and consuming a large amount can contribute to weight gain if your total daily calorie intake exceeds your requirements.

BiPE - Outros parâmetros - i

Há ainda outros parâmetros a considerar, nomeadamente o **Top-K** e o **Top-P**.

Top-K - significa que o modelo escolherá os tokens mais prováveis, dos que o modelo gerou (a analogia dos atores e suas *queries*). Quanto maior o Top-K mais criativa e variada será a resposta [10].

Top-P - em vez de um número fixo de tokens, o modelo selecionará aqueles cuja probabilidade cumulativa não exceda o valor P (0 será greedy decoding e 1 poderá ter todos os tokens no vocabulário do LLM) [10].



BiPE - Outros parâmetros - ii

No entanto... assumindo que todos os parâmetros estão acessíveis, o valor de temperatura sobrepõe-se:

Temperatura 0 significa que top-K e top-P são irrelevantes, pois o modelo selecionará o token mais provável;

Temperatura >1 torna-se irrelevante e os tokens que passarem nos critérios de top-K e top-P serão utilizados aleatoriamente para predizer o próximo token [10].



BiPE - Outros parâmetros (de volta ao teatro)

Imaginando a dita peça de teatro onde cada ator representa um **bloco de processamento do Transformer**, e o guião são os tokens gerada pelo modelo:

Top-K: O diretor fornece aos atores uma lista com **K** possíveis falas. Os atores escolhem aleatoriamente uma dessas falas.

Top-P: O diretor fornece uma lista de falas cuja soma de probabilidades atinge P. Os atores escolhem aleatoriamente uma dessas falas.

Temperatura: Determina o quão restritivas serão as instruções do diretor. Uma temperatura baixa significa que os atores seguem rigidamente o guião, enquanto que uma temperatura alta permite maior liberdade (improvisação).



BiPE - Técnicas de prompt i

Zero-shot *prompt* (nome significa "sem exemplos") [10],[11]

Tipo mais simples de *prompt*.

Coloca apenas uma breve descrição e algum texto para que o LLM proceda.

Por norma não é desejada criatividade, logo a temperatura pode ser baixa.

Exemplo de *prompt* [10]:

"Classify movie reviews as POSITIVE, NEUTRAL or NEGATIVE. Review: "Her" is a disturbing study revealing the direction humanity is headed if AI is allowed to keep evolving, unchecked. I wish there were more movies like this masterpiece. Sentiment:"



BiPE - Técnicas de prompt ii

Se o Zero-shot *prompt* falhar podemos utilizar a técnica técnica "**few-shot**" [12].

Esta técnica fornece ao modelo múltiplos exemplos, de forma a mostrar ao modelo o padrão que este deve utilizar [10].

Exemplo [10]

Parse a customer's pizza order into valid JSON:

EXAMPLE:

I want a small pizza with cheese, tomato sauce, and pepperoni.

JSON Response:

```
{
  "size": "small",
  "type": "normal",
  "ingredients": ["cheese", "tomato sauce", "pepperoni"]
}
```

...

EXAMPLE:

Can I get a large pizza with tomato sauce, basil and mozzarella

```
{
  "size": "large",
  "type": "normal",
  "ingredients": ["tomato sauce", "basil", "mozzarella"]
}
```

Now, I would like a large pizza, with the first half cheese and mozzarella.
And the other tomato sauce, ham and pineapple.

BiPE - Técnicas de prompt iii

System prompt é uma forma de dizer ao modelo qual o contexto geral e o que é esperado do modelo[10].

Temos acesso a este *system prompt* em locais ou programas específicos, como o [Jan.ai](#), o OpenWeb UI, o LMStudio... ou no playground do ChatGPT ou na sua versão Plus.

Role prompt permite-nos atribuir uma função/profissão/papel específico ao LLM, dando-lhe assim a ideia do tom e estilo. Isto permite-lhe assim gerar respostas que sejam mais relevantes e apropriadas ou especializadas [10].

Em termos de estilos podemos ainda dizer-lhe que ele pode ser, por exemplo, confrontacional, direto, formal, humorista, persuasivo...

Contextual prompt não é mais que dotarmos o LLM do contexto, detalhes ou outra informação relevante para a tarefa que ele irá executar, permitindo assim ao modelo entender as nuances do que é pedido e formular a sua resposta de acordo com as mesmas[10].



BiPE - Técnicas de prompt iv

Pode haver sobreposição entre system, contextual e role prompts.

No entanto, cada um tem um propósito diferente [10]:

System prompt - Define as capacidades fundamentais e o objetivo global do modelo;

Contentual prompt - Fornece informação imediata e específica para a tarefa corrente, de forma a guiar a resposta do modelo. Sendo específica para a tarefa/input, tem carácter dinâmico.

Role prompt - Dá enquadramento ao modelo, dando-lhe "estilo e voz". Permite adicionar uma camada de personalidade e especificidade.



BiPE - Técnicas de prompt v

Step-back prompt é quando nós pedimos ao LLM para gerar uma dada resposta, num tom mais geral, e depois lhe fornecemos de volta a sua resposta e lhe pedimos para resolver um assunto específico. Isto dá ao LLM contexto, "background" e ativa os seus processos de raciocínio [10].

Prompt Inicial (típico) [10]: Write a one paragraph storyline for a new level of a first- person shooter video game that is challenging and engaging.

Resposta: The level begins with the player's squad being ambushed in a dense urban areaby a relentless enemy faction. Players must fight their way through a maze of back alleys and dilapidated buildings, utilizing stealth and precision to eliminate enemy forces and avoid detection. The player must use their environment to their advantage, setting traps, utilizing cover, and making calculated...

Prompt Inicial (para processo de step-back): Based on popular first-person shooter action games, what are 5 fictional key settings that contribute to a challenging and engaging level storyline in a first-person shooter video game?

Resposta: 1. ****Abandoned Military Base****: A sprawling, post-apocalyptic military complex crawling with mutated soldiers and rogue robots, ideal for challenging firearm combat.

2. ****Cyberpunk City****: A neon-lit, futuristic urban environment with towering skyscrapers and dense alleyways, featuring cybernetically enhanced enemies and hacking mechanics...



BiPE - Técnicas de prompt vi

Prompt Final processo de step-back): Context: 5 engaging themes for a first person shooter video game:

1. ****Abandoned Military Base****: A sprawling, post-apocalyptic military complex crawling with mutated soldiers and rogue robots, ideal for challenging firearm combat.
2. ****Cyberpunk City****: A neon-lit, futuristic urban environment with towering skyscrapers and dense alleyways, featuring cybernetically enhanced enemies and hacking mechanics.

...

Take one of the themes and write a one paragraph storyline for a new level of a first-person shooter video game that is challenging and engaging.

Resposta: In the heart of a murky abyss, lies a dilapidated underwater research facility, standing as a testament to human ambition and its disastrous consequences. Shrouded in darkness, pulsating with the hum of malfunctioning machinery, and haunted by the echoes of drowned souls, the facility is now a breeding ground for a nightmare borne from the deep sea. The player, an elite marine equipped with advanced diving gear and experimental weaponry...



BiPE - Técnicas de prompt vii

Chain of Thought (CoT) é uma técnica utilizada para tentar melhorar a capacidade de raciocínio do modelo, obrigando-o a criar passos intermédios no raciocínio [10].

Prompt [10]: When I was 3 years old, my partner was 3 times my age. Now, I am 20 years old. How old is my partner? Let's think step by step.

Resposta:

1. Let's assume my current age is "x" years.
2. When I was 3 years old, my age was 3 years.
3. My partner's age at that time was 3 times my age, which means my partner's age was $3 * 3 = 9$ years
4. Now, I am 20 years old, which means my age has increased by $20 - 3 = 17$ years since I was 3 years old.
5. Since my partner's age was 9 years when I was 3 years old, and my age has increased by 17 years, my partner's age must have also increased by 17 years.
6. Therefore, my partner's current age is $9 + 17 = 26$ years.

So, my partner is 26 years old.

E ainda...

Self-consistency;
Tree of Thoughts;
ReAct (reason & act);

...

Para mais informação, ler:
L. Boonstra, 'Prompt Engineering', Google,
Sep. 2024.

Uso de LLMs na investigação



universidade de aveiro



DigiMedia
digital media and interaction
research centre

Uso de LLMs na investigação

E aquele "clickbait" 😏 na frase do workshop? Algo sobre LLMs e investigação...?

Vamos então falar disso!



universidade de aveiro



DigiMedia
Digital Media and Interaction
Research Centre

Uso de LLMs na investigação - UA

Há várias iniciativas na Universidade de Aveiro relacionadas com IA e grupos de discussão, particularmente sobre o seu uso em contexto de ensino, mas também de investigação.

A biblioteca da UA tem/tinha como recomendação [Perplexity](#) como ferramenta de IA.

Há também [este estudo](#) que avalia algumas das ferramentas disponíveis, realçando os já conhecidos problemas e cuidados com o seu uso, como plágio, privacidade de dados etc e aplicando-as na escrita científica.

Uso de LLMs na investigação - linhas orientadoras

Há várias linhas orientadoras sobre a utilização da IA, nomeadamente das Nações Unidas, da Unesco e, claro da União Europeia, particularmente do uso responsável da IA generativa na investigação.

Há também iniciativas por parte de empresas, que procuram criar um co-cientista capaz de ajudar na transdisciplinaridade tão necessária à investigação atual!



Uso de LLMs na investigação - um "caso prático" - i

Imaginem que querem cruzar alguns destes documentos e suas linhas orientadoras com o relatório da UA "Triangular: Ferramentas de IA Generativa na Escrita Académica" e ainda com a ferramenta da Google referida anteriormente (o dito AI Co-scientist).

Tradicionalmente iríamos passar bastante tempo a ler, a sublinhar, a resumir, a procurar convergência...

O ChatGPT levou menos de 1 minuto!

Uso de LLMs na investigação - um "caso prático" - ii

Prompt (contextual e role): Tu és um investigador em ciências sociais, em Portugal. Queres verificar quais os pontos em comum e as divergências entre as normas europeias, da ONU, da Unesco e um documento gerado na tua universidade. Queres ver como tudo isto poderá encaixar num relatório da União Europeia acerca das estratégias e metodologias a implementar para aumentar a competitividade europeia na tecnologia, investigação e desenvolvimento.

Mas vamos ver se o LLM dá uma ajuda no prompt!



universidade de aveiro



DigMedia
Digital Media and Interaction
Research Centre

Uso de LLMs na investigação - um "caso prático" - iii

Contexto:

É um investigador em ciências sociais, baseado em Portugal, interessado em analisar as convergências e divergências entre as normas da União Europeia, da ONU, da UNESCO e um documento produzido na sua universidade. O objetivo é compreender como essas diretrizes podem ser aplicadas aos pontos referidos num relatório da União Europeia que delineia estratégias e metodologias para reforçar a competitividade europeia nas áreas de tecnologia, investigação e desenvolvimento.

Instruções:

Com base nos documentos fornecidos, identifique os pontos comuns e as divergências entre as normas mencionadas. Analise como essas informações podem ser aplicadas num documento com base no relatório da União Europeia focado em estratégias para aumentar a competitividade europeia em tecnologia, investigação e desenvolvimento, a fim de produzir um conjunto de áreas de atuação.

Documentos para Análise: [link](#); [link](#); [link](#); [link](#); [link](#)

Uso de LLMs na investigação - um "caso prático" - iv

Notas (do ChatGPT):

O AI Co-Scientist da Google é uma ferramenta colaborativa baseada em IA que auxilia cientistas na geração de hipóteses e propostas de pesquisa, promovendo descobertas científicas aceleradas.

O artigo "Triangular" da Universidade de Aveiro destaca a utilização de múltiplas ferramentas de IA Generativa na escrita académica, enfatizando a triangulação para aumentar a confiabilidade e qualidade dos conteúdos gerados.

O relatório "Align, Act, Accelerate" da Comissão Europeia propõe recomendações estratégicas para maximizar o impacto dos programas de pesquisa e inovação da UE, promovendo uma agenda transformadora.



universidade de aveiro



Digital Media
& Oral Interaction
Research Centre

Uso de LLMs na investigação - um "caso prático" - iv

Vamos ver no ChatGPT, ok?

E pedir uma "tabela" bonita com os dados todos direitinhos!



universidade de aveiro



DigiMedia
Digital Media and Interaction
Research Centre

Uso de LLMs na investigação - ChatGPT resolveu?

Sim...

E chegaria? Depende do nosso objetivo.

Mas importa reter:

Tivemos agência na escolha dos documentos e mecanismos a comparar;
Tínhamos conhecimento prévio (embora que não exaustivo) dos pontos principais dos documentos e mecanismos;

Perigos:

Os pontos gerados nos desafios e recomendações podem ter implicações éticas e, de qualquer forma, deverão ser verificados e cruzados, com os documentos originais.

Um tipo de utilização bastante recomendado é a utilização da IA como auxílio na correção dos textos, particularmente quando se escrever numa língua que não a materna (ou até nessa 🙄)



universidade de aveiro



DigMedia
Digital Media and Interaction
Research Centre

Uso de LLMs... - agregadores e outros serviços

[Semantic Scholar](#) - motor de busca de artigos (aditivado com AI) E com a possibilidade de utilizar um leitor "inteligente", que nos fornece sinopses de artigos citados, que cruza o artigo com os que tenhamos na nossa biblioteca, etc (login with Universidade de Aveiro)

[Open Research Knowledge Graph](#) - Permite encontrar, comparar/contrastar artigos...

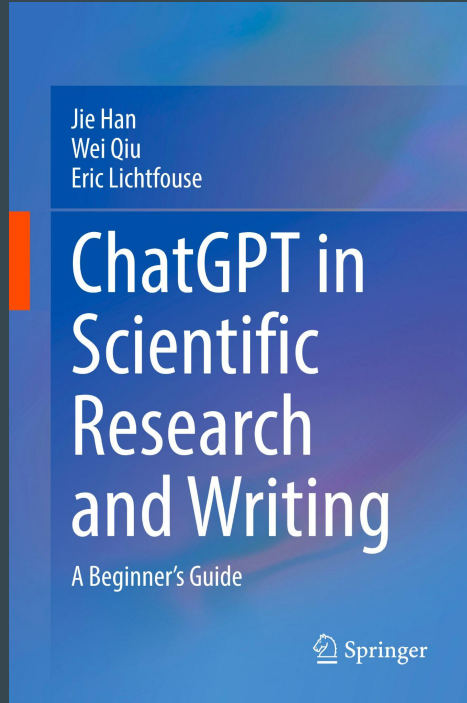
[Hugging Face Daily papers](#) - com possibilidade de subscrição (envio para email), mais focado em... AI, claro!

E os suspeitos do costume - [Research Rabbit](#) e [Connected Papers](#)

YouTube... por exemplo [este canal](#).



Uso de LLMs na investigação - fresquinho!!!



[Podem descarregar aqui!](#)



universidade de aveiro



Digital Media
and Interaction
Research Centre

Uso de LLMs na investigação - "senso comum"

Para finalizar...

Senso comum! Espírito crítico! Confirmar dados!
Pedir fontes! Procurar noutros locais
(Scopus/ACM/PubMed/Google Scholar/...)
quando são artigos ou similares, procurar online
noutros casos a mesma informação e cruzar!

Como viram os LLMs querem ajudar. Mas como
disse a Jessica Rabbit...



R. Zemeckis (Realizador), Who Framed Roger Rabbit? [filme], Walt Disney Pictures, Touchstone Pictures, EUA, 1988. Excerto em formato GIF disponível, [Acesso em: 14-05-25].



universidade de aveiro



DiGiMedia
Digital Media and Interaction
Research Centre

LLMs locais versus na nuvem



universidade de aveiro



DigiMedia
digital media and interaction
research centre

LLMs locais versus na nuvem

Porquê correr modelos localmente [17]?

- **Não dependem da Internet:** Não é necessário acesso à internet para gerar texto ou processar dados, o que aumenta a autonomia e segurança;
- **Privacidade e Controlo de Dados:** Há total controlo sobre a informação que introduzimos, reduzindo assim riscos de fuga de dados;
- **Eliminação de Restrições:** Não precisamos de nos preocupar com limites de tokens, latência de rede ou custos de utilização;
- **Flexibilidade e Personalização:** Há uma vasta gama de modelos "**destilados**" (distilled) e otimizados para tarefas específicas, incluindo variantes **quantizadas** para execução rápida em dispositivos com recursos limitados (+/-)...



LLMs locais - Palavras-chave i

Modelos **destilados** - São uma versão mais pequena, mais rápida e com menos necessidades computacionais do modelo original, mas que procuram manter a sua performance [18],[19].

Faz uso do paradigma *Professor - Estudante*.

Podem ser utilizados como *chatbots*, mecanismos de sumarização de texto, tradução, etc.

Ex: BERT → DistilBERT (40% mais leve, 60% mais rápido, mas com 97% do desempenho do original). [20]

No entanto, pode/vai perder alguma precisão e "nuances" do modelo original.



LLMs locais - Palavras-chave ii

Quantização - É uma técnica que reduz a precisão dos cálculos do modelo, representando os **pesos** (weights - parâmetros aprendidos pelo modelo durante o treino, que determinam a força das conexões entre os neurónios artificiais) [21].

Diminui-se, assim, o tamanho do modelo e acelera-se a chegada a conclusões (inferência), mas com um impacto na precisão do modelo.

LLMs locais - Quantização

Representação de Pesos (Bit Precision)

Bits	Valores Possíveis	Tipo de Modelo	Exemplo
1 bit	0, 1	Redes binárias ou binarizadas	Alguns modelos experimentais de baixa precisão [23]
4 bits	16 valores discretos	Modelos quantizados	LLaMA [Gerganov, 2023], GPT-4 (quantized) [22]
16 bits	65,536 valores	Modelos acelerados, com precisão intermediária	FP16 usados em GPUs para inferência rápida [25]
32 bits	~4,3 mil milhões de valores	Modelos de alta precisão	FP32 usados em treino de grandes LLMs, como GPT-4 [22] e PaLM [24]



LLMs locais... - Como e onde escolher modelos? i

Podemos usar para nossa referência a seguinte tabela (notem que o tipo de GPU também impacta) [26]:

Tamanho de Modelos	RAM	Características
3B - 7B	8GB	Modelos tiny ou fortemente quantizados
7B - 13B	16GB	Quantização moderada
+/- 30B	32GB+	Modelos com maior precisão, mas ainda menores que "small"

LLMs locais... - Como e onde escolher modelos? ii

- Podemos ver modelos na comunidade [Hugging Face](#) (e seleccionar com base nos nossos requisitos/máquinas);
- Ver o [Leaderboard](#) da Hugging Face;
- Consultar a [Awesome-LLM](#);
- Consultar [Open-LLMs](#);
- ...

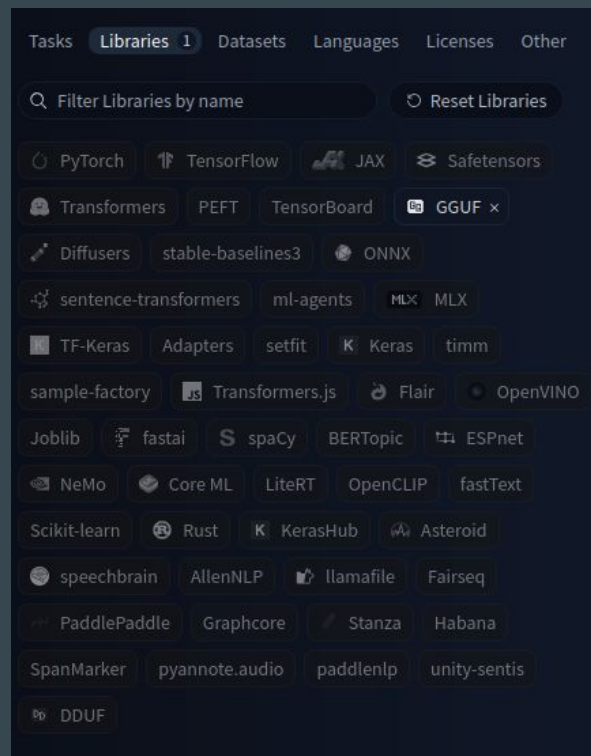


LLMs locais... - Como e onde escolher modelos? iii

Para escolher o modelo devemos saber o que o nosso software permite importar/usar.

Por exemplo, o [Jan.ai](#) apenas utiliza modelos com o formato **GUFF** (GPT-Generated Unified Format).

No Hugging Face, em **Libraries** podemos ativar o filtro GUFF e escolher o modelo que melhor se adequa.



LLMs locais... - Como e onde escolher modelos? iv

The screenshot displays the Hugging Face Model Hub interface. On the left, there are navigation tabs for 'Tasks', 'Libraries', 'Datasets', 'Languages', 'Licenses', and 'Other'. Below these, a search bar 'Filter Tasks by name' is present. The 'Multimodal' section lists tasks like 'Audio-Text-to-Text', 'Image-Text-to-Text', 'Visual Question Answering', 'Document Question Answering', 'Video-Text-to-Text', 'Visual Document Retrieval', and 'Any-to-Any'. The 'Computer Vision' section lists tasks like 'Depth Estimation', 'Image Classification', 'Object Detection', 'Image Segmentation', 'Text-to-Image', 'Image-to-Text', 'Image-to-Image', 'Image-to-Video', 'Unconditional Image Generation', 'Video Classification', 'Text-to-Video', 'Zero-Shot Image Classification', 'Mask Generation', 'Zero-Shot Object Detection', and 'Text-to-3D'.

The main area shows a list of models under the 'Models' tab, with a count of 117,459. A search bar 'Filter by name' and a 'Full-text search' button are at the top right. The 'Sort' dropdown is set to 'Trending'.

The list of models includes:

- wsbagnsv1/ltxv-13b-0.9.7-dev-GGUF**: Image-to-Video • Updated 3 days ago • 22.2k • 52
- unsloth/Qwen3-30B-A3B-GGUF**: Text Generation • Updated 1 day ago • 233k • 165
- Qwen/Qwen3-30B-A3B-GGUF**: Text Generation • Updated 4 days ago • 4.5k • 18
- silveroxides/Chroma-GGUF**: Text-to-Image • Updated about 22 hours ago • 67.7k • 64
- city96/FLUX.1-dev-gguf**: Text-to-Image • Updated Aug 18, 2024 • 171k • 1.04k
- Tesslate/UGEN-T2-7B-Q8_0-GGUF**: Text Generation • Updated 9 days ago • 3.11k • 125
- unsloth/Qwen3-30B-A3B-128K-GGUF**: Text Generation • Updated 1 day ago • 44.5k • 39
- Kijai/HunyuanVideo_comfy**: Updated 6 days ago • 8.01k • 438
- wsbagnsv1/ltxv-13b-0.9.7-distilled-GGUF**: Image-to-Video • Updated about 11 hours ago • 12
- Qwen/Qwen3-14B-GGUF**: Text Generation • Updated 6 days ago • 13.7k • 24
- bartowski/ServiceNow-AI_Apriel-Nemotron-15b-Thinker...**: Text Generation • Updated 7 days ago • 2.27k • 11
- ggml-org/SmolVLM-500M-Instruct-GGUF**: Updated 15 days ago • 5.53k • 11

LLMs locais... - Como e onde escolher modelos? v

Exemplo de modelo GGUF: [janhq/trinity-v1-GGUF](#)

Nome do Modelo: **janhq/trinity-v1-GGUF**

Tamanho do Modelo: 7.24B params (Refere-se ao número de parâmetros do modelo (7,24 mil milhões), o que impacta diretamente nos requisitos de memória e capacidade computacional).

Arquitetura: llama (Especifica a arquitetura base do modelo, neste caso, baseada na família LLaMA).

Quantizações Disponíveis: Os modelos GGUF frequentemente oferecem versões quantizadas o que permite equilibrar desempenho e uso de recursos:

Q3_K_M (3-bit): ~3.52 GB - Q (quantizado)_3(3bits)_K(K-mean, uma técnica que reduz complexidade agrupando pesos semelhantes)_M (tamanho do bloco de pesos contíguos). Como nas peças de roupa, S, M, L... aqui S além de small indica maior precisão!

Q4_K_M (4-bit): ~4.37 GB

...

Compatibilidade com Hardware: Algumas páginas de modelos fornecem estimativas ou recomendações sobre a compatibilidade com diferentes configurações de hardware, principalmente se tiverem conta no Hugging Face e tiverem preenchido o vosso hardware.

Licença - Verifica a licença do modelo para assegurar que está de acordo com os usos pretendidos.

LLMs locais versus na nuvem - As partes más...

- Não termos que precisar de internet é bom... ou mau, se quisermos dotar o modelo de acesso à internet para procurar dados mais recentes ou fazer uma verificação da sua resposta contrastando com os dados disponíveis na internet;
- Precisamos de bom hardware/computador;
- Precisamos de ser nós a manter os modelos atualizados;
- Custo energético e de espaço em disco;



Aplicações e recursos úteis

Sempre que possível de código-aberto (open-source)



universidade de aveiro



DigiMedia
digital media and interaction
research centre

Aplicações e recursos úteis

"I wonder **how much money** OpenAI has lost in electricity costs from people saying 'please' and 'thank you' to their models," writes user @tomieinlove.

"Tens of millions of dollars well spent—you never know," Altman responded[13].

Para além de custos de eletricidade [14] e ambientais (água[15],[16] como é produzida a energia etc...) temos os custos pessoais - quanto nos "custa" um *prompt*?

Aplicações e recursos úteis

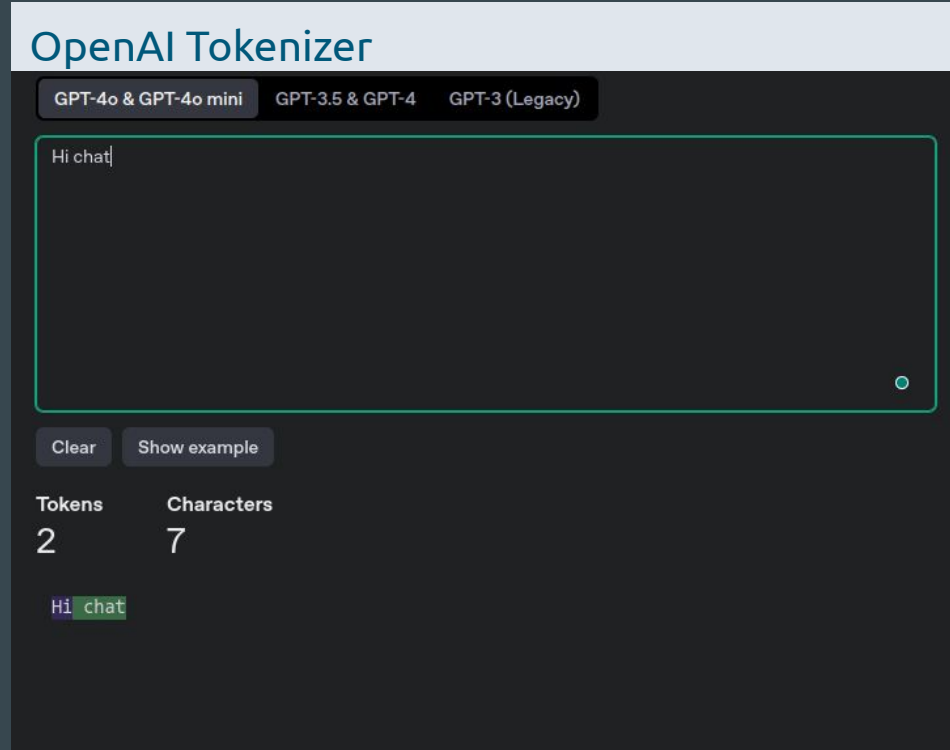


Figura 4: OpenAI Tokenizer Interface Fonte: [27]

Aplicações e recursos úteis

OpenAI Tokenizer

Many words map to one token, but some don't: indivisible.

Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🍌

Sequences of characters commonly found next to each other may be grouped together: 1234567890



Clear

Show example

Tokens

53

Characters

252

Many words map to one token, but some don't: indivisible.

Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🍌

Sequences of characters commonly found next to each other may be grouped together: 1234567890

Text

Token IDs



universidade de aveiro



DigiMedia
Digital Media and Interaction
Research Centre

Figura 5: OpenAI Tokenizer Interface Fonte: [27]

Aplicações e recursos úteis

O Pinokio pode ser visto como uma plataforma de instalação e automatização de aplicações de IA.

<https://program.pinokio.computer/#/?id=pinokio>



Imagem gerada com o DALL-E

Aplicações e recursos úteis

Depois de instalado, permite correr dezenas de ferramentas de IA localmente, de forma contentorizada e isolada, com todas as bibliotecas necessárias Instaladas automaticamente, sem interferir no sistema principal.

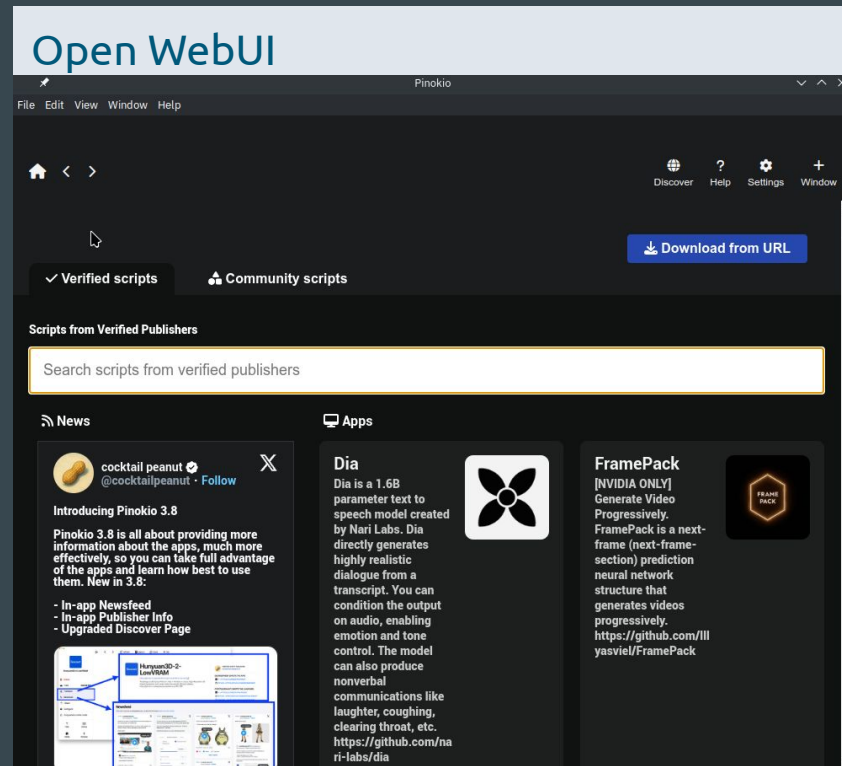
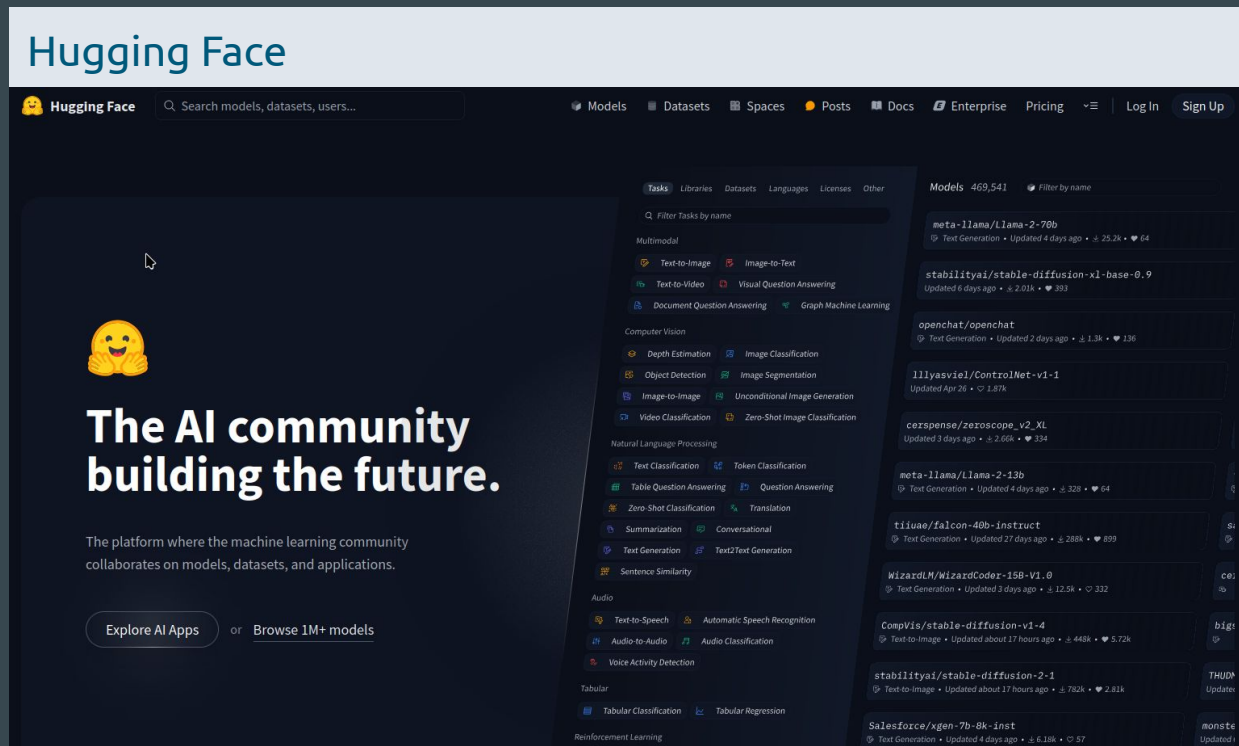


Figura 6: Interface Open WebUI Fonte: [28]

Aplicações e recursos úteis



The screenshot shows the Hugging Face homepage with a dark theme. The main header includes the Hugging Face logo, a search bar, and navigation links for Models, Datasets, Spaces, Posts, Docs, Enterprise, Pricing, Log In, and Sign Up. The main content area features a large emoji (🤗) and the text "The AI community building the future." Below this, it states "The platform where the machine learning community collaborates on models, datasets, and applications." and provides buttons for "Explore AI Apps" and "Browse 1M+ models". On the right, there is a sidebar with a "Tasks" section listing various AI tasks like Text-to-Image, Image-to-Text, etc., and a "Models" section displaying a list of popular models such as meta-llama/Llama-2-70b, stabilityai/stable-diffusion-xl-base-0.9, and openchat/openchat.

Hugging Face

🤗 Hugging Face 🔍 Search models, datasets, users...

Models Datasets Spaces Posts Docs Enterprise Pricing Log In Sign Up

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Text-to-Image Image-to-Text
- Text-to-Video Visual Question Answering
- Document Question Answering Graph Machine Learning

Computer Vision

- Depth Estimation Image Classification
- Object Detection Image Segmentation
- Image-to-Image Unconditional Image Generation
- Video Classification Zero-Shot Image Classification

Natural Language Processing

- Text Classification Token Classification
- Table Question Answering Question Answering
- Zero-Shot Classification Translation
- Summarization Conversational
- Text Generation Text2Text Generation
- Sentence Similarity

Audio

- Text-to-Speech Automatic Speech Recognition
- Audio-to-Audio Audio Classification
- Voice Activity Detection

Tabular

- Tabular Classification Tabular Regression

Reinforcement Learning

Models 469,541 Filter by name

- meta-llama/Llama-2-70b
Text Generation • Updated 4 days ago • 25.2k • 64
- stabilityai/stable-diffusion-xl-base-0.9
Updated 6 days ago • 2.01k • 393
- openchat/openchat
Text Generation • Updated 2 days ago • 1.1k • 136
- lillyasviel/ControlNet-v1-1
Updated Apr 26 • 1.87k
- cerspense/zerospense_v2_XL
Updated 3 days ago • 2.66k • 334
- meta-llama/Llama-2-13b
Text Generation • Updated 4 days ago • 328 • 64
- tiitune/falcon-40b-instruct
Text Generation • Updated 27 days ago • 288k • 899
- WizardLM/WizardCoder-15B-V1.0
Text Generation • Updated 3 days ago • 12.5k • 332
- CompVis/stable-diffusion-v1-4
Text-to-Image • Updated about 17 hours ago • 448k • 5.72k
- stabilityai/stable-diffusion-2-1
Text-to-Image • Updated about 17 hours ago • 782k • 2.81k
- Salesforce/xgen-7b-8k-inst
Text Generation • Updated 4 days ago • 6.18k • 57

Figura 7: Homepage Hugging Face Fonte: [29]

Aplicações e recursos úteis

Hugging Face pode ser visto como um repositório e comunidade onde é possível encontrar, testar, aprender e colaborar em modelos, bibliotecas (como a Transformers) e datasets para IA.

Disponibiliza o seu próprio "chat" online, bem como ferramentas como o AutoTrain, que permite treinar modelos sem programar.

Tem também uma secção dedicada à ciência, com notificações e acesso facilitado a artigos científicos, [aqui!](#)



Aplicações e recursos úteis

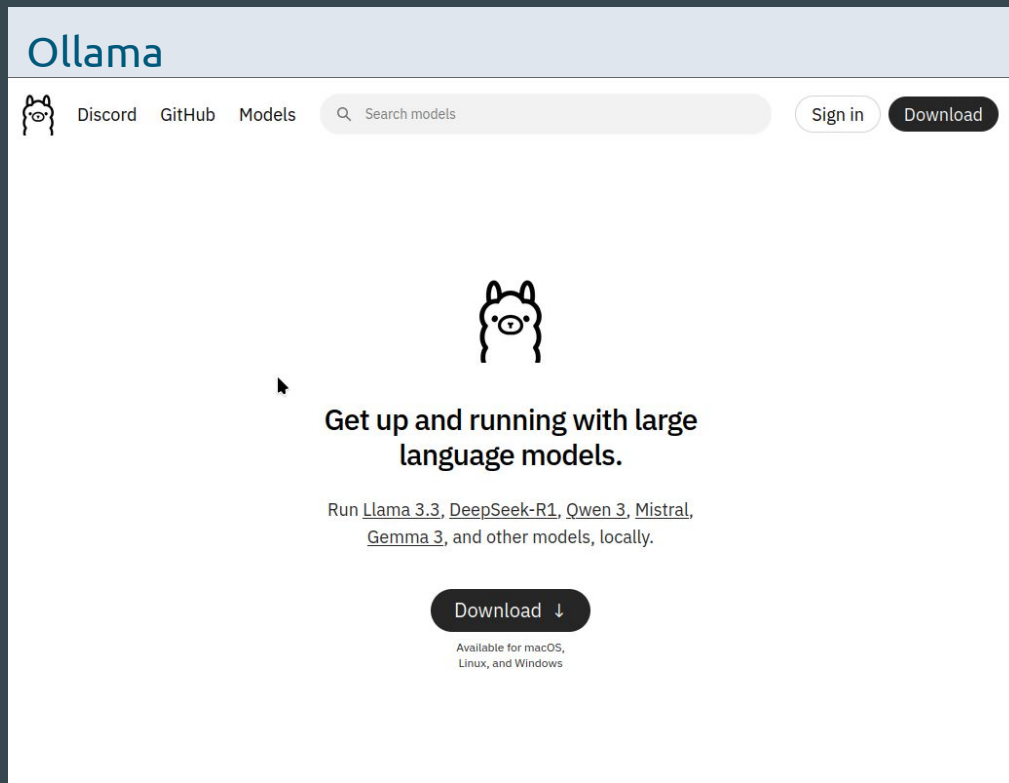


Figura 8: Homepage Ollama Fonte: [30]

Aplicações e recursos úteis

Ollama é uma ferramenta que permite instalar e correr localmente modelos de IA através da linha de comando.

Foca-se na simplicidade, privacidade e rapidez: com um comando (`ollama run`), podemos instalar e iniciar modelos.

Ideal para quem quer testar rapidamente modelos LLM no seu hardware, sem necessidade de configurações complexas.

Suporta modelos como LLaMA, Mistral, Gemma, entre outros — e permite integração com ferramentas como o **OpenWebUI**.



Aplicações e recursos úteis

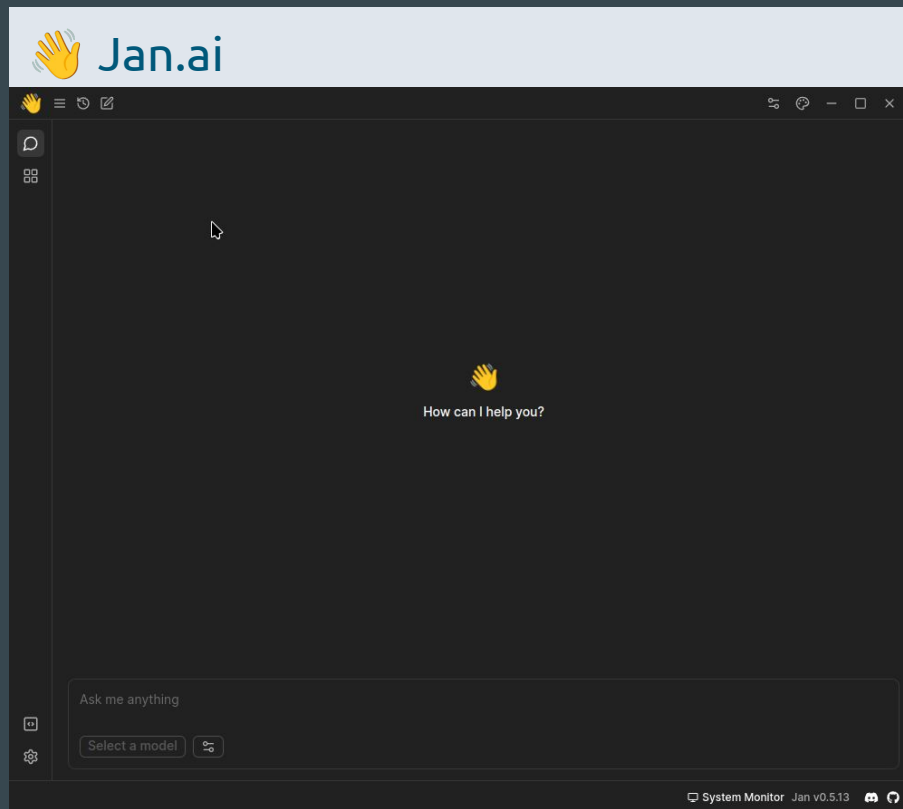


Figura 9: Interface do Jan.AI Fonte: [31]

Aplicações e recursos úteis

Jan.ai afirma-se como uma alternativa local, 100% offline* e privada ao ChatGPT (e similares).

Permite-nos descarregar vários modelos de LLM's, com uma notificação indicando se o modelo é suportado pela nossa máquina. Permite também importar modelos.

Podemos questionar o LLM, analisar textos e criar chatbots, entre outras possibilidades.

*** Mas também podemos colocar a nossa API key e ligar a modelos online!**

Aplicações e recursos úteis

LMStudio

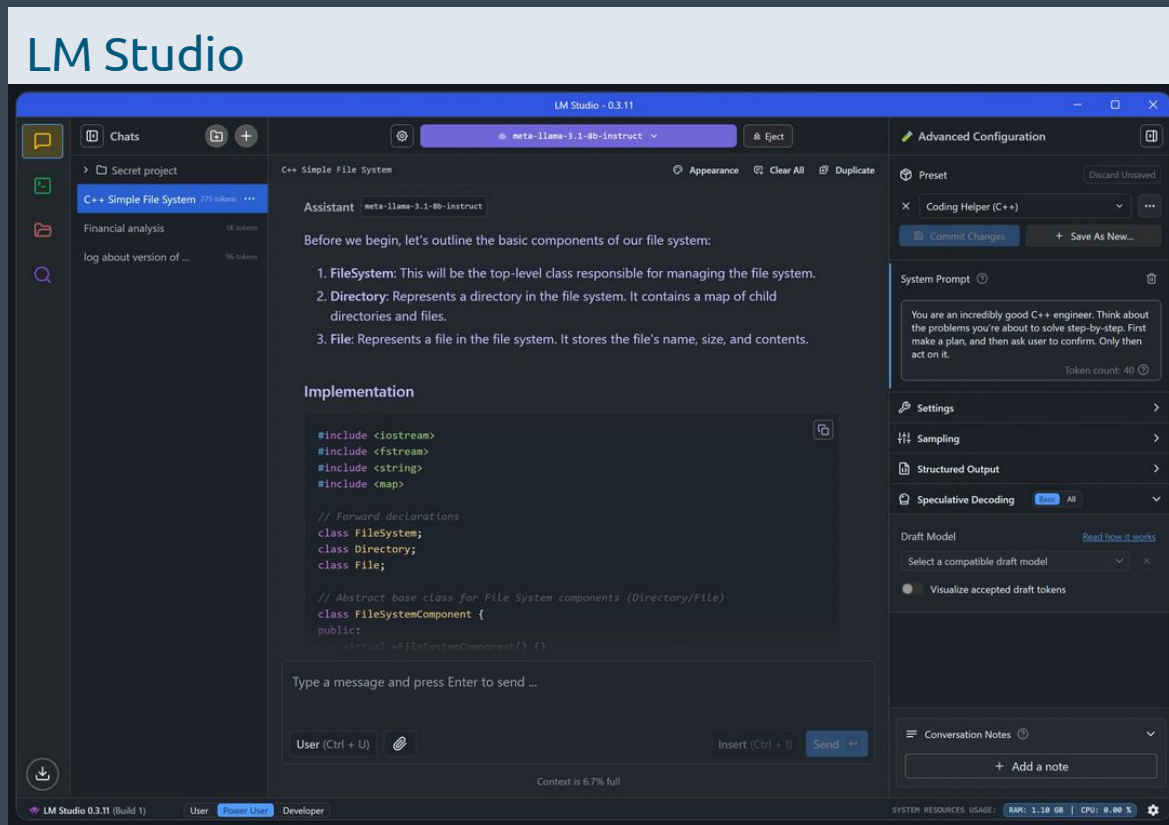


Figura 10: Interface do LM Studio Fonte: [32]

Aplicações e recursos úteis

O LM Studio afirma ser uma aplicação que facilita a procura e utilização de modelos de IA locais.

É mais orientado para desenvolvimento e oferece um conjunto de ferramentas mais avançadas do que o Jan.AI, como a possibilidade de usar dois modelos em simultâneo e pô-los a "discutir" um prompt — embora isso requeira copiar a resposta de um e colar no outro 🤨.

O LM Studio é parcialmente fechado: combina uma interface proprietária com componentes open-source.



Aplicações e recursos úteis

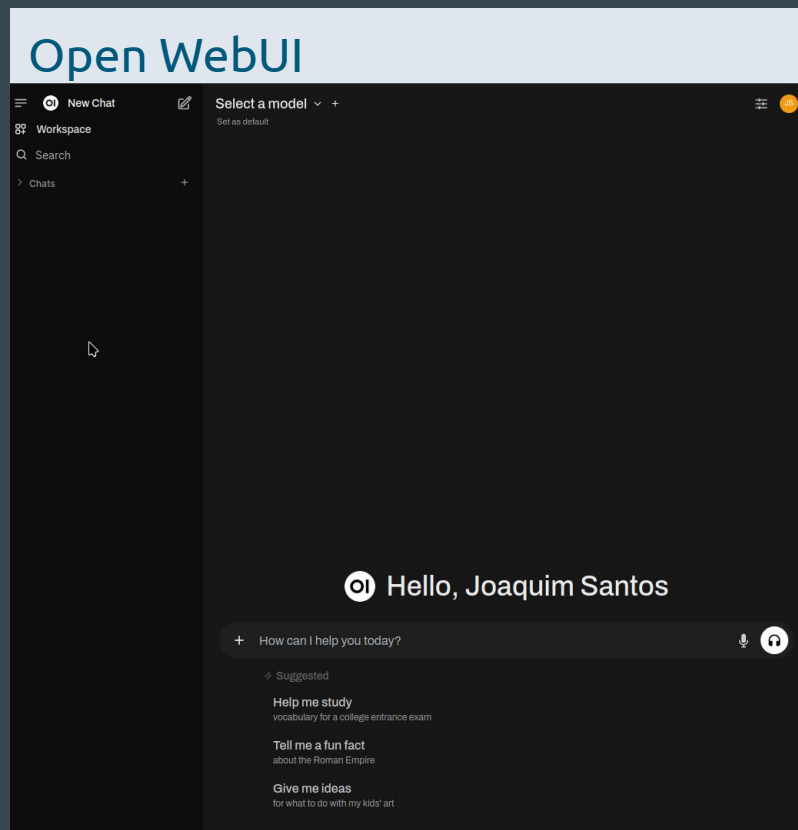


Figura 11: Interface do Open WebUI Fonte: [28]

Aplicações e recursos úteis

OpenWebUI é uma interface gráfica (browser) open-source para interagir com modelos de IA locais, como os da ferramenta Ollama.

Permite criar, organizar e conversar com chatbots personalizados, tudo a partir do browser.

Design similar aos "concorrentes", suporta múltiplas conversas, histórico, vários utilizadores...



	Função Principal	Nível de Complexidade	Observações Relevantes
OpenAI Tokenizer	Ver tamanho de prompts (tokens)	Iniciante	Útil para entender limites de modelos como ChatGPT
Pinokio	AppStore local de apps de IA (com automação)	Iniciante	Instala e corre apps localmente em containers
Jan.AI	Interface gráfica para LLMs locais	Iniciante	Foco em simplicidade e uso offline, 100% local
Ollama	Motor local para correr LLMs via terminal	Iniciante	Simples, rápido, com instalação de modelos com 1 comando
LM Studio	Desenvolvimento e teste de LLMs locais	Intermédio	Permite usar 2 modelos em paralelo (manual), plugins, logs
OpenWebUI	Interface web local para LLMs (via Ollama, etc.)	Intermédio	Interface elegante, suporta plugins e múltiplos backends
Hugging Face	Repositório e comunidade de modelos de IA	Avançado	Milhares de modelos prontos, espaço para projetos, API, datasets

Tabela "gentilmente" criada pelo ChatGPT



universidade de aveiro



Digital Media
and AI Interaction
Research Centre

Criação de... 



universidade de aveiro



DigiMedia
digital media and interaction
research centre

Parte "hands on"

Chatbot;

Escolha de modelos;

Uso de RARG;

Uso de tools;

...



universidade de aveiro



DigiMedia
a digital media and interaction
research centre

Recursos e referências

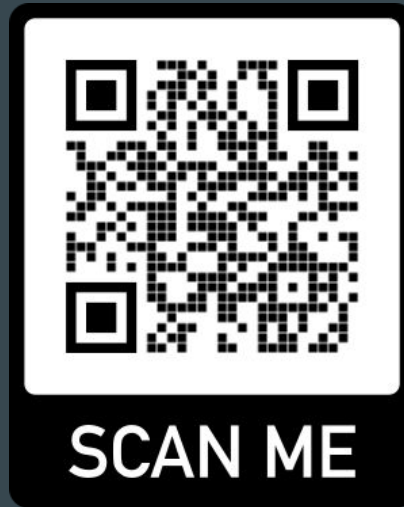


universidade de aveiro



DigiMedia
digital media and interaction
research centre

Recursos



universidade de aveiro



DigMedia
Digital Media and Interaction
Research Centre

Referências i

- [1] J. Berryman and A. Ziegler, Prompt engineering for LLMs: the art and science of building large language model-based applications, First Edition. Sebastapol, CA: O'Reilly Media, 2025.
- [2] C. E. Shannon, 'A Mathematical Theory of Communication', Bell Syst Tech J, vol. 27, pp. 379–423, 1948.
- [3] 'Vanishing Gradient Problem : Everything you need to know', Engati. Accessed: May 08, 2025. [Online]. Available: <https://www.engati.com/glossary/vanishing-gradient-problem>
- [4] D. Bahdanau, K. Cho, and Y. Bengio, 'Neural Machine Translation by Jointly Learning to Align and Translate', 2014, arXiv. doi: 10.48550/ARXIV.1409.0473.
- [5] A. Vaswani et al., 'Attention Is All You Need', 2017, arXiv. doi: 10.48550/ARXIV.1706.03762.
- [6] L. Gao et al., 'The Pile: An 800GB Dataset of Diverse Text for Language Modeling', 2021, arXiv. doi: 10.48550/ARXIV.2101.00027.
- [7] Z. Lee et al., 'ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation', 2025, arXiv. doi: 10.48550/ARXIV.2503.21729.
- [8] 'Visão | Cmo é que o crberéo cnosgeue ler etsa farse?', Visão. Accessed: May 10, 2025. [Online]. Available: <https://visao.pt/visaosaude/2017-05-25-cmoo-e-que-o-crbereo-cnogsgeue-ler-etsa-farse/>
- [9] milvus.io, 'What is the maximum context window for OpenAI's models?' Accessed: May 12, 2025. [Online]. Available: <https://milvus.io/ai-quick-reference/what-is-the-maximum-context-window-for-openais-models>
- [10] L. Boonstra, 'Prompt Engineering', Google, Sep. 2024.
- [11] J. Wei et al., 'Finetuned Language Models Are Zero-Shot Learners', 2021, arXiv. doi: 10.48550/ARXIV.2109.01652.
- [12] T. B. Brown et al., 'Language Models are Few-Shot Learners', 2020, arXiv. doi: 10.48550/ARXIV.2005.14165.
- [13] 'Why Being Polite to ChatGPT Is Costing OpenAI Millions and Wasting Electricity', PCMAG. Accessed: Apr. 29, 2025. [Online]. Available: <https://www.pcmag.com/news/why-being-polite-to-chatgpt-is-costing-openai-millions-and-wasting-electricity>
- [14] 'AI is poised to drive 160% increase in data center power demand', Goldman Sachs. Accessed: Apr. 29, 2025. [Online]. Available: <https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand>

Referências ii

- [15] United Nations, 'Artificial intelligence: How much energy does AI use?', United Nations Western Europe. Accessed: May 15, 2025. [Online]. Available: <https://unric.org/en/artificial-intelligence-how-much-energy-does-ai-use/>
- [16] United Nations, 'Artificial Intelligence (AI) end-to-end: The Environmental Impact of the Full AI Lifecycle Needs to be Comprehensively Assessed - Issue Note', 2024. Accessed: May 15, 2025. [Online]. Available: <https://wedocs.unep.org/20.500.11822/46288>
- [17] 'How to Run Large Language Models (LLMs) Locally: A Beginner's Guide t...', archive.is. Accessed: May 13, 2025. [Online]. Available: <https://archive.is/i0G32>
- [18] 'LLM Distillation Explained: Applications, Implementation & More'. Accessed: May 14, 2025. [Online]. Available: <https://www.datacamp.com/blog/distillation-llm>
- [19] 'LLMs: Fine-tuning, distillation, and prompt engineering | Machine Learning', Google for Developers. Accessed: May 14, 2025. [Online]. Available: <https://developers.google.com/machine-learning/crash-course/llm/tuning>
- [20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', 2019, arXiv. doi: 10.48550/ARXIV.1910.01108.
- [21] I. Goodfellow, A. Courville, and Y. Bengio, Deep learning. in Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press, 2016.
- [22] OpenAI et al., 'GPT-4 Technical Report', 2023, arXiv. doi: 10.48550/ARXIV.2303.08774.
- [23] M. Courbariaux, Y. Bengio, and J.-P. David, 'BinaryConnect: Training Deep Neural Networks with binary weights during propagations', 2015, arXiv. doi: 10.48550/ARXIV.1511.00363.
- [24] A. Chowdhery et al., 'PaLM: scaling language modeling with pathways', J. Mach. Learn. Res., vol. 24, no. 1, Jan. 2023.
- [25] NVIDIA, 'Train with Mixed Precision'. 2023. Accessed: May 14, 2025. [Online]. Available: Train with Mixed Precision



Referências iii

- [[26] A. Shatokhin, 'LLMs quantization naming explained', Medium. Accessed: May 15, 2025. [Online]. Available: <https://andreshat.medium.com/llm-quantization-naming-explained-bedde33f7192>
- [27] OpenAI, '<https://platform.openai.com/tokenizer>'.
- [28] Open WebUI, '<https://openwebui.com/>', Open WebUI.
- [29] Hugging Face, 'Hugging Face', Hugging Face. Accessed: May 01, 2025. [Online]. Available: <https://huggingface.co/>
- [30] Ollama, 'Ollama', Ollama. Accessed: May 01, 2025. [Online]. Available: <https://ollama.com/>
- [31] Menlo Research, 'Jan.ai: Chat with AI without privacy concerns', Jan.ai: Chat with AI without privacy concerns. [Online]. Available: <https://jan.ai/>
- [32] LM Studio, LM Studio: Your local AI toolkit. [Online]. Available: LM Studio: Your local AI toolkit



Unboxing AI: uma introdução

...

Como escolher os LMM, questionar e utilizá-los na sua investigação, de forma eficaz!



universidade de aveiro



DiGMedia
Digital Media and Interaction
Research Centre