ETAPA 2

Explorando modelos grandes de linguagem (LLMs)

Explorando modelos grandes de linguagem (LLMs)

Objetivos da Etapa

- Tokenização em LLMs
- Explicar critérios de escolha de LLMs
- Comparar LLMs proprietários e open-source
- Analisar trade-offs de tamanho x desempenho
- Avaliar desafios éticos e limitações

Tokenização e LLMs

Tokenização de Textos: Processo de dividir um texto em unidades menores (tokens), como palavras, subpalavras ou caracteres, para processamento em modelos de aprendizado de máquina. Tipos de

Tokenização por Palavra: cada palavra do texto é considerada um token. Útil para análise de palavras, como frequência de palavras e modelos de bag-of-words.

- Aplicações: Modelos clássicos de NLP, como TF-IDF ou n-grams.

“O gato está dormindo”

["O", "gato", "está", "dormindo"]

Tokenização e LLMs

Tokenização por Caracteres: cada caractere do texto é tratado como um token. Funciona bem com textos curtos, emojis ou linguagens artificiais.

- **Aplicações**: processamento de texto em linguagens com sistema de escrita complexo (como chinês). Útil para tarefas que envolvem correção ortográfica ou quando palavras não estão claramente definidas.

"caminhando"

[‘c’,’a’,’m’,’i’,’n’,’h’,a’,’’n’,’d’,’o’]

Tokenização e LLMs

Tokenização por Sentença: descrição: Cada frase é tratada como um token. Captura o contexto de uma frase completa, útil para análises de entonação e estrutura de texto.

- **Aplicações**: modelos de sumarização, análise de coesão textual, e sistemas que precisam processar sentenças inteiras, como na tradução automática.

Um dia fomos. No outro, não!

[‘Um dia fomos.’, ‘No outro, não!’]

Tokenização e LLMs

Byte-Pair Encoding - BPE: divide palavras em pedaços menores, como prefixos e sufixos, formando subpalavras. Reduz o tamanho do vocabulário, capturando morfologia de palavras, e é eficiente para lidar com palavras desconhecidas ou raras.

- **Aplicações**: amplamente usada em modelos de linguagem avançados, como o BERT e o GPT. Indicada para traduções automáticas, geração de texto e qualquer tarefa que lide com palavras compostas.

“O gato está dormindo”

["O", " g”, ”ato", " est”, ”á", " dorm”, “ind”, “o"]

https://platform.openai.com/tokenizer

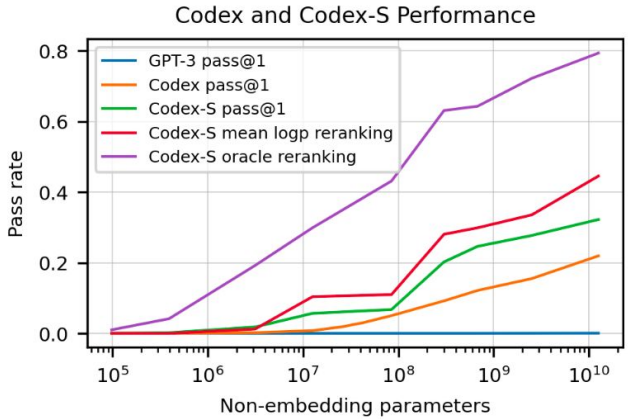
Critérios para Escolher um LLM

Requisitos de Tarefa: o critério essencial ao escolher um LLM é entender qual tarefa específica o modelo precisará executar pois, a depender do objetivo, diferentes modelos podem ser mais adequados.

- **Processamento de Linguagem Natural**: Modelos como GPT-4 ou BERT são amplamente usados para tarefas como análise de sentimentos, classificação de textos e extração de informações.
- **Resumo de Textos**: Se a tarefa é resumir grandes quantidades de texto, como relatórios ou artigos, modelos treinados para compressão de informações são ideais.
- **Tradução**: Modelos como M2M-100 ou T5 são otimizados para tradução automática e oferecem precisão elevada para múltiplos idiomas.
- **Geração de Código**: Codex (baseado em GPT) é um exemplo de modelo especializado em geração e correção de código de programação.

GPT-3

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%



https://ar5iv.labs.arxiv.org/html/2107.03374

Critérios para Escolher um LLM

Capacidade de Personalização: dependendo da aplicação, pode ser necessário ajustar um LLM pré-treinado (ou fine-tuned) para otimizar a performance em tarefas específicas.

- **Modelos Base:** são treinados em grandes volumes de dados gerais, o que os torna versáteis e prontos para uso imediato. Podem lidar com uma variedade de tarefas sem necessidade de ajustes adicionais.
- **Modelos Fine-Tuned**: aplicação exige resultados muito específicos, comum em cenários onde a linguagem utilizada é técnica ou especializada, como em áreas médicas ou jurídicas. Fine-tuning permite que o modelo aprenda características particulares de um domínio ou estilo, aumentando sua precisão e relevância.

	1st	2nd	3rd	4th	5th
GPT-3 training data (2019) [35]	English (93%)	French (1.8%),	German (1.5%)	Spanish (0.8%)	Italian (0.6%)
Languages represented on the Internet (2021) [36]	English (44.9%)	Russian (7.2%)	German (5.9%)	Chinese languages (4.6%)	Japanese (4.5%)
First-languages spoken (2019) [37]	Mandarin Chinese (12%)	Spanish (6%),	English (5%),	Hindi (4.4%),	Bengali (4%).
Most spoken language (2021)[37]	English (1348M)	Mandarin Chinese (1120M)	Hindi (600M)	Spanish (543M)	Standard Arabic (274M)

https://www.researchgate.net/figure/Top-five-languages-included-in-GPT-3-training-data-compared-against-other-measures-of-the_tbl1_359256884

Critérios para Escolher um LLM

Orçamento: o custo de utilizar um LLM pode variar bastante dependendo do modelo e da infraestrutura escolhida.

- **APIs pagas**: paga-se por chamadas ao modelo, ideal para aplicações que precisam de grande poder computacional sem a necessidade de construir e manter a infraestrutura própria (GPT-4 e Gemini). Custos podem ser significativos dependendo do volume de uso.
- **Treinamento Próprio**: tarefas que necessitem de um LLM altamente customizado, o que requer poder computacional elevado e infraestrutura para lidar com o treinamento, possivelmente com a aquisição de hardware específico (GPUs) e consumo de energia.
- **Equilíbrio de Custos**: modelos open-source podem ser mais econômicos a longo prazo, porém com um custo inicial relativamente alto para treinar e manter o modelo.

https://llmpricecheck.com

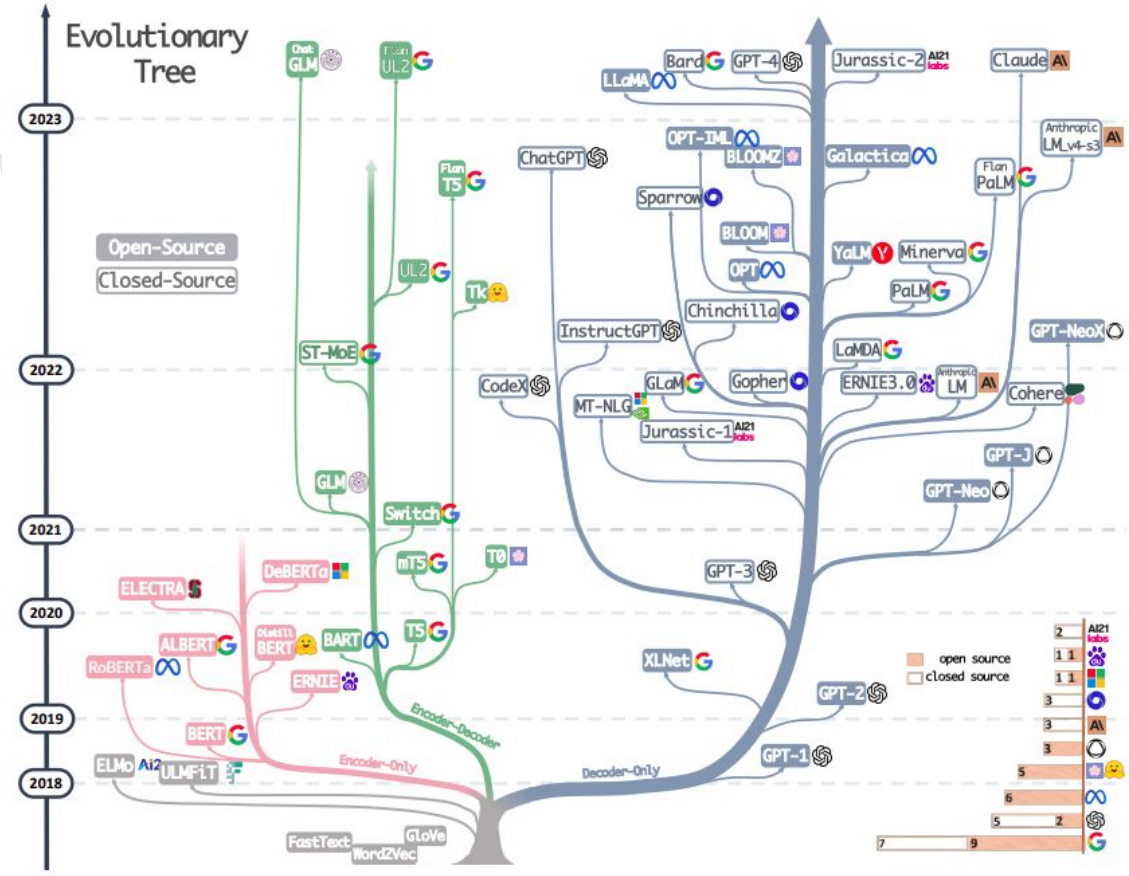


Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

Modelos com Arquitetura MoE

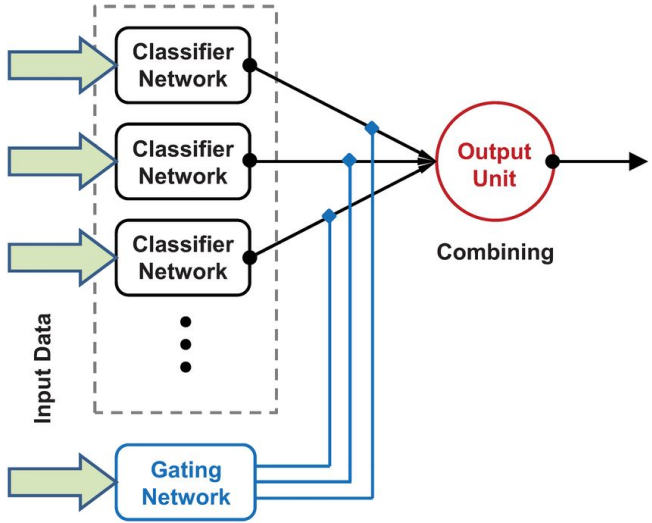
Mixture of Experts - MoE: é uma técnica de aprendizado de máquina em que um modelo seleciona dinamicamente diferentes subconjuntos de "especialistas" (submodelos menores e especializados) para processar partes específicas de uma entrada, ponderadas através de um roteador para gerar a resposta final.

Escalabilidade: permite a criação de modelos massivos sem aumentar proporcionalmente o custo computacional, pois apenas uma parte do modelo é ativada em cada momento.

Eficiência Computacional: o modelo pode ser muito maior sem exigir um aumento linear nos recursos de processamento, permitindo o uso de LLMs com menor custo por inferência.

Complexidade: requer um roteador eficiente para selecionar os especialistas corretos, o que aumenta a complexidade do treinamento.

Balanceamento de carga: garantir que todos os especialistas sejam ativados de maneira equilibrada, evitando que o aprendizado seja enviesado.



https://www.researchgate.net/figure/The-standard-MoE-architecture-The-outputs-classifications-from-the-classifier-networks_fig10_309296341

LLM proprietários e open-source.

Modelos Proprietários (GPT-4, Gemini)

- **Performance Superior**: tendem a ser líderes em benchmarks de desempenho, oferecendo alta precisão e melhor generalização em diversas tarefas.
- **Suporte Técnico**: empresas que desenvolvem esses modelos oferecem suporte técnico e atualizações constantes, garantindo que as soluções implementadas sejam robustas e seguras.
- **Segurança**: modelos proprietários geralmente têm medidas de segurança e alinhamento embutidas, como mitigação de vieses e controle sobre geração de conteúdo nocivo.
- **Caixa-Secreta**: código e dados usados para treinar esses modelos não são divulgados, o que os torna menos transparentes.
- **Custos Elevados**: custos significativos para aplicações em larga escala ou com altos volumes de chamadas.

Modelos Open-Source (LLaMA, Falcon)

- **Transparência**: oferecem acesso total ao código-fonte, permitindo que desenvolvedores compreendam como o modelo foi treinado, ajustem a arquitetura e investiguem potenciais problemas de viés ou desempenho.
- **Personalização**: é possível ajustar o modelo exatamente para a sua aplicação, o que inclui treinamento adicional em dados específicos
- **Custo-efetividade**: o custo inicial de treinamento pode ser diluído em aplicações de longo prazo com alto volume de dados.
- **Manutenção**: usar e manter um modelo open-source requer mais esforço técnico, como a configuração de infraestrutura e ajustes constantes do modelo.
- **Suporte Limitado**: correções e fixes podem depender de comunidades de desenvolvedores que contribuem nos códigos, o que pode ser um desafio em situações críticas.

LLM proprietários e open-source.

Modelos Proprietários	Modelos Open-Source
Desempenho Geral	Específico e Flexível
Suporte Técnico	Transparência
Custo	Conhecimento Técnico
Transparência	Infraestrutura

Principais modelos de LLMs disponíveis Open-source

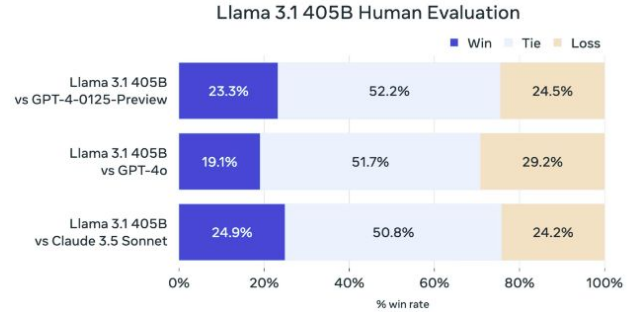
LLaMA - Large Language Model Meta AI foi projetado para ser uma alternativa aberta e altamente personalizável a modelos proprietários.

- **Performance:** suas versões mais recentes demonstraram excelente desempenho em benchmarks (MMLU), mostrando uma sólida compreensão de múltiplos idiomas e uma capacidade forte em raciocínio e análise de contexto. LLaMA 3 oferece um desempenho superior em tarefas de geração de código e tradução
- **Transparência:** O código-fonte e os pesos do modelo são acessíveis, permitindo personalização e adaptação do modelo para necessidades específicas
- **Escalabilidade:** o LLaMA pode ser ajustado de acordo com as necessidades, desde pequenas tarefas até implementações corporativas com modelos de até 70 bilhões de parâmetros

Meta Llama 3 Pre-trained model performance

	Meta Llama 3 8B	Mistral 7B Published	Mistral 7B Measured	Gemma 7B Published	Gemma 7B Measured
MMLU 5-shot	66.6	62.5	63.9	64.3	64.4
AGIEval English 3-5-shot	45.9	--	44.0	41.7	44.9
BIG-Bench Hard 3-shot, CoT	61.1	--	56.0	55.1	59.0
ARC-Challenge 25-shot	78.6	78.1	78.7	53.2 0-shot	79.1
DROP 3-shot, F1	58.4	--	54.4	--	56.3

	Meta Llama 3 70B	Gemini Pro 1.0 Published	Mixtral 8x22B Measured
MMLU 5-shot	79.5	71.8	77.7
AGIEval English 3-5-shot	63.0	--	61.2
BIG-Bench Hard 3-shot, CoT	81.3	75.0	79.2
ARC-Challenge 25-shot	93.0	--	90.7
DROP 3-shot, F1	79.7	74.1 variable-shot	77.6



https://datasciencedojo.com/blog/meta-llama-3-1/

Principais modelos de LLMs disponíveis - Open-source

Falcon (TII AI): desenvolvido pelo Instituto de Inovação em IA (TII), foi projetado para ser um modelo de alto desempenho em tarefas de NLP, com foco em eficiência e rapidez, se destacando em contextos menores e de recursos limitados.

- **Performance:** supera muitos modelos maiores em benchmarks como commonsense reasoning e compreensão de leitura. Mesmo com um tamanho de modelo menor, compete diretamente com modelos maiores como LLaMA 2 e Mistral 7B.
- **Eficiência Computacional:** é otimizado para consumir menos recursos computacionais, permitindo que seja executado com eficiência em infraestruturas leves.
- **Personalização e Flexibilidade:** permite ajustes finos e pode ser adaptado para diferentes tarefas, tornando-o ideal para empresas e desenvolvedores que buscam um modelo balanceado entre custo e performance.

Falcon LLM Version	Main Feature	Popular Applications
Falcon-180B	Most powerful	- Generating complex creative text formats (e.g., poems, scripts) - Handling massive datasets for advanced analytics - Research and development in AI
Falcon-40B	Balanced power and accessibility	- Text summarization - Machine translation - Question answering - Chatbot development - Data analysis and classification
Falcon-7.5B	Lightweight and efficient	- Sentiment analysis - Social media monitoring - Customer service chatbots - Content filtering and moderation
Falcon-1.3B	Highly portable and resource-friendly	- Basic text classification tasks - Simple question-answering - Sentiment analysis for targeted marketing - Extracting critical information from documents
Falcon-X-Instruct	Fine-tuned for following instructions	- Completing specific writing tasks based on prompts (e.g., writing different creative text formats based on instructions) - Summarizing factual topics according to user-defined guidelines - Answering open-ended, challenging, or strange questions in a comprehensive and informative way

https://dataforest.ai/blog/falcon-ai-180-billion-parameters

Principais modelos de LLMs disponíveis - Pagos

GPT-4 - modelo com alto desempenho em benchmarks de raciocínio, compreensão de texto e tarefas complexas, fornecendo respostas mais precisas e criativas em comparação com modelos anteriores.

- **Capacidades Multimodais**: versão avançada, o GPT-4o, também pode processar entradas de imagem, ampliando suas aplicações em áreas como análise visual e geração de legendas.
- **Exemplos de Benchmark**: ótimo desempenho em exames padronizados como o Bar Exam e o GRE.
- **Aplicações Comuns**: usado amplamente em chatbots, sistemas de recomendação e outras tarefas que exigem entendimento profundo e geração de texto.
- **Desafios e Limitações**: ainda pode apresentar alucinações (respostas incorretas) em cenários complexos e informações de eventos recentes.

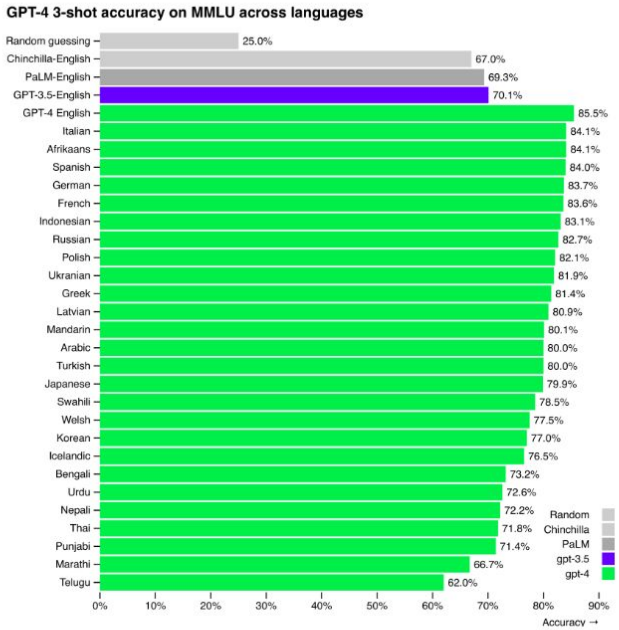


Figure 5: Performance of GPT-4 in a variety of languages compared to prior models in English on MMLU. GPT-4 outperforms the English-language performance of existing language models (Hoffmann et al., 2022; Chowdhery et al., 2022) for the vast majority of languages tested, including low-resource languages such as Latvian, Welsh, and Swahili.

https://arxiv.org/html/2303.08774v6

GPT-4

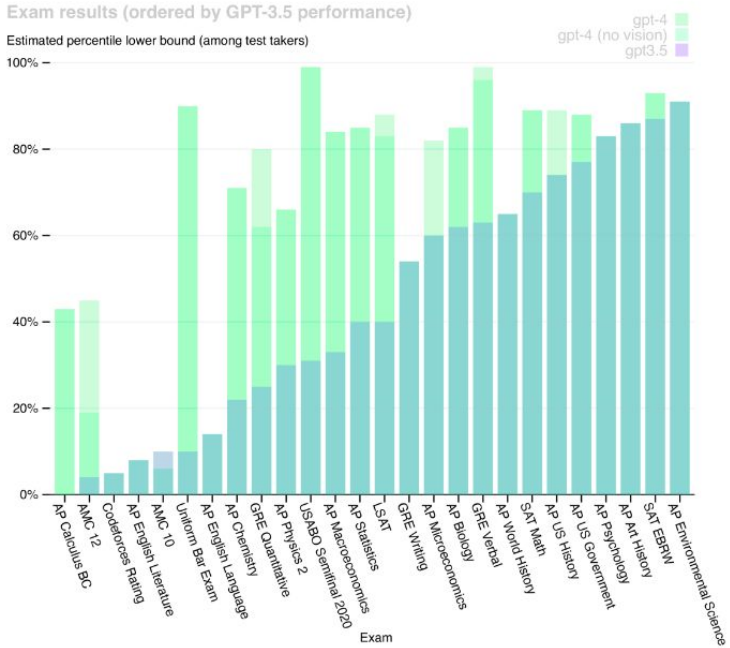


Figure 4: GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. Exams are ordered from low to high based on GPT-3.5 performance. GPT-4 outperforms GPT-3.5 on most exams tested. To be conservative we report the lower end of the range of percentiles, but this creates some artifacts on the AP exams which have very wide scoring bins. For example although GPT-4 attains the highest possible score on AP Biology (5/5), this is only shown in the plot as 85th percentile because 15 percent of test-takers achieve that score.

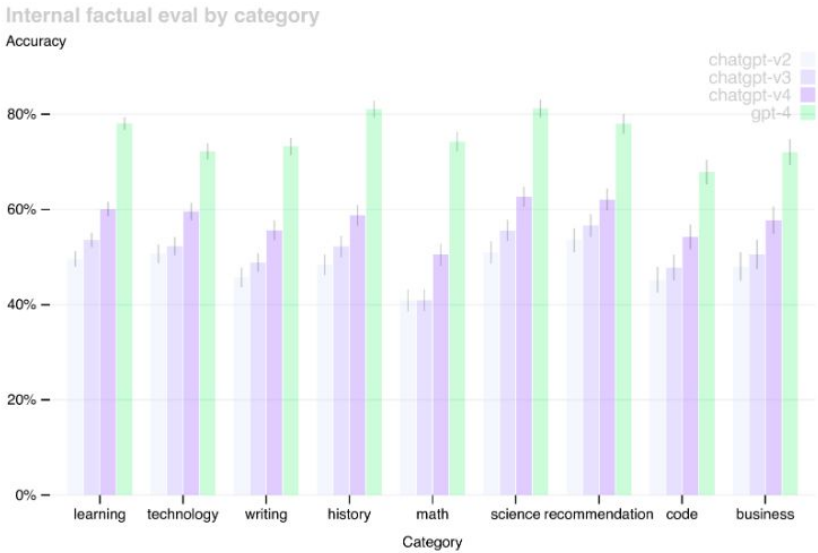


Figure 6: Performance of GPT-4 on nine internal adversarially-designed factuality evaluations. Accuracy is shown on the y-axis, higher is better. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval. We compare GPT-4 to three earlier versions of ChatGPT OpenAI (2022) based on GPT-3.5; GPT-4 improves on the latest GPT-3.5 model by 19 percentage points, with significant gains across all topics.

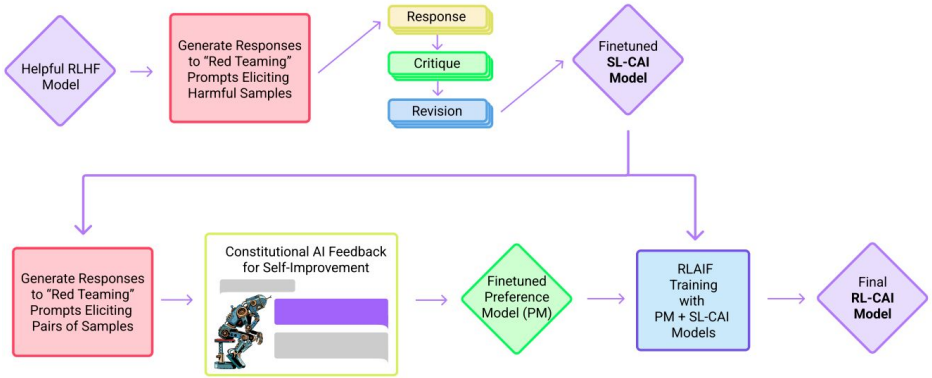
https://arxiv.org/html/2303.08774v6

Principais modelos de LLMs disponíveis - pagos

Claude AI: é um LLM baseado em transformers, desenvolvido pela Anthropic, com foco em segurança e alinhamento da IA. Foi treinado com dados públicos da internet e dados proprietários, usando aprendizado não supervisionado, Reinforcement Learning with Human Feedback (RLHF), e uma técnica inovadora chamada Constitutional AI (CAI).

CAI é uma técnica única desenvolvida pela Anthropic para melhorar a confiabilidade e a segurança da IA, com o objetivo de fazer com que o modelo seja útil, honesto e inofensivo, prevenindo respostas tóxicas, discriminatórias ou ilegais. Treinamento baseado em princípios: fontes como a Declaração Universal dos Direitos Humanos e melhores práticas de segurança.

Fase 1: O modelo aprende a criticar e revisar suas próprias respostas usando princípios de alinhamento, sem depender de feedback humano direto.



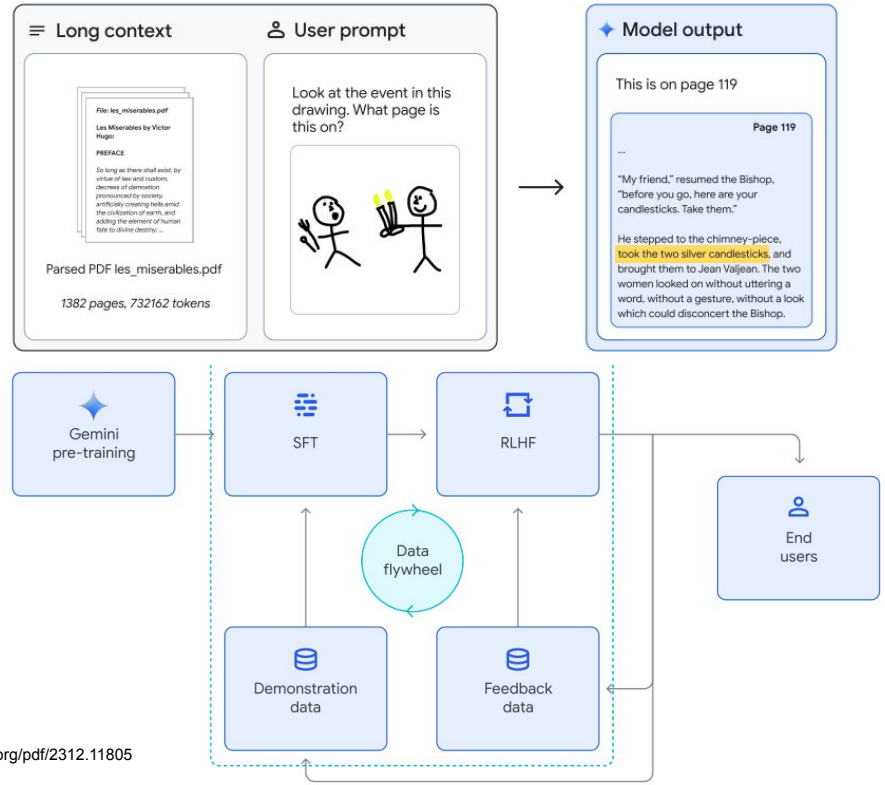
Fase 2: O modelo utiliza feedback gerado por IA, baseado nos princípios, para selecionar saídas mais seguras e inofensivas.

https://arxiv.org/abs/2212.08073

Principais modelos de LLMs disponíveis - Pagos

Gemini - da Google, é um dos modelos mais avançados em termos de multimodalidade, lidando não apenas com texto, mas também com imagens e som, permitindo sua aplicação em uma ampla gama de tarefas.

- **Capacidades Multimodais**: se destaca pela capacidade de processar dados de diversas fontes simultaneamente, oferecendo soluções em tarefas que combinam imagem, áudio e texto.
- **Performance e Eficiência**: oferece desempenho de alta qualidade em benchmarks relacionados à compreensão de linguagem e análise de conteúdo visual.
- **Aplicações**: assistentes virtuais avançados, processamento de imagens médicas e sistemas de análise multimídia.
- **Desafios**: o custo computacional e a complexidade para treinamento e manutenção.



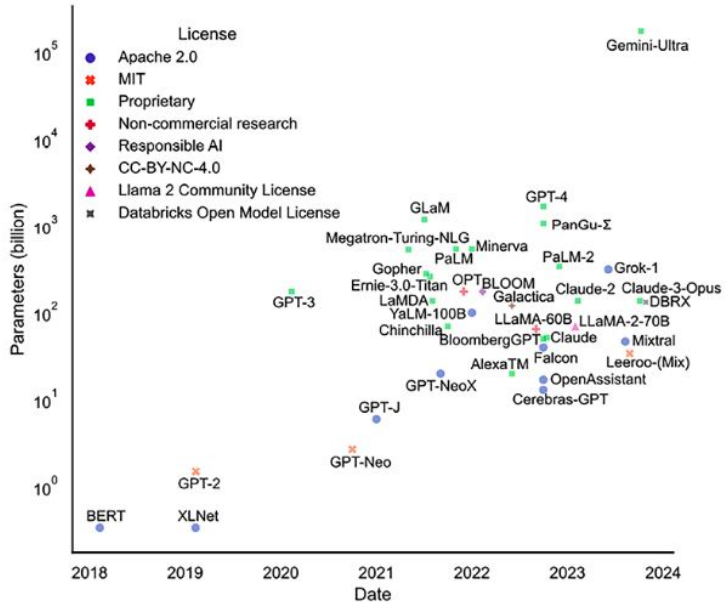
https://arxiv.org/pdf/2312.11805

.

Trade-offs de tamanho x desempenho

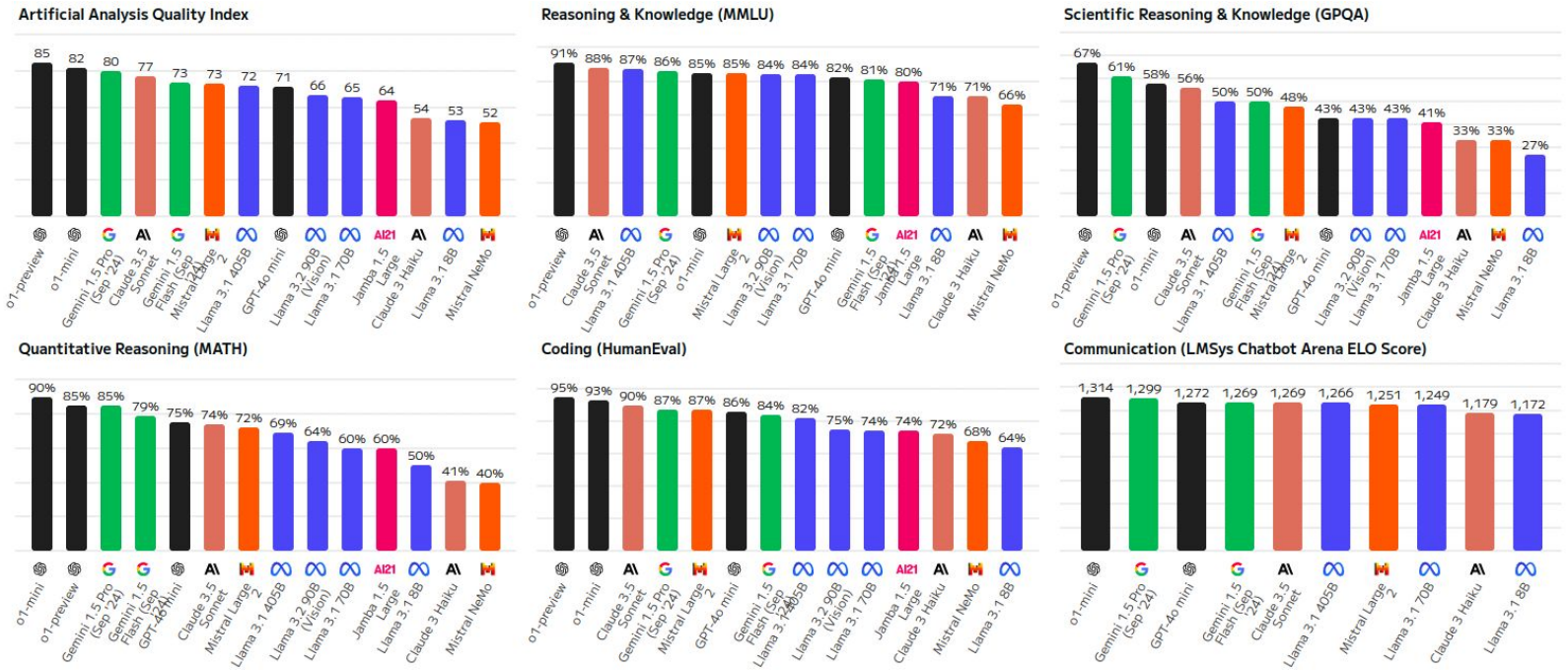
Tamanho de modelos: são medido principalmente pelo número de coeficientes ajustáveis durante o treinamento que determinam como o modelo processa as entradas. Esses parâmetros são responsáveis pela complexidade e capacidade de aprendizado do modelo.

- **Generalização**: LLMs grandes têm maior chance de lidar bem com tarefas de zero-shot e few-shot learning, ou seja, conseguem generalizar melhor mesmo com pouca ou nenhuma amostra específica de uma tarefa.
- **Custo Computacional**: o aumento no número de parâmetros resulta em um consumo muito maior de recursos computacionais, como memória e poder de processamento.
- **Latência**: modelos maiores tendem a ser mais lentos para gerar respostas em tempo real, o que pode impactar negativamente sua aplicação em sistemas que exigem respostas rápidas.



https://learning.oreilly.com/library/view/generative-ai-with/9781835083468/Text/Chapter_1.xhtml#_idParaDest-26

Trade-offs de tamanho x desempenho



https://artificialanalysis.ai/

Desafios éticos e limitações no uso LLMs

Limitações dos modelos:

Capacidade de Raciocínio e Factualidade

Frequentemente "alucinam": em benchmarks como o TruthfulQA, o GPT-3 apresenta dificuldade em distinguir entre informações verdadeiras e falsas, gerando respostas factualmente incorretas.

Viés e Discriminação

Podem reproduzir vieses de gênero, raça e outros preconceitos, pois são treinados em grandes volumes de dados da web

Alto Custo Computacional

Modelos com bilhões de parâmetros demandam uma enorme quantidade de recursos computacionais para treinar e utilizar (consumo de energia e aquecimento global).

Memória de Curto Prazo

Têm uma janela limitada de contexto (tokens). Em aplicações como tradução de documentos longos ou respostas contextuais em conversas complexas, o modelo perde informações de interações anteriores, prejudicando a coerência (GPT3: 4096).

Falta de Verdadeira Compreensão

LLMs não têm entendimento real do que estão dizendo; sua geração de texto é baseada em padrões estatísticos. LLMs alucinam em matemática.

Falta de Transparência:

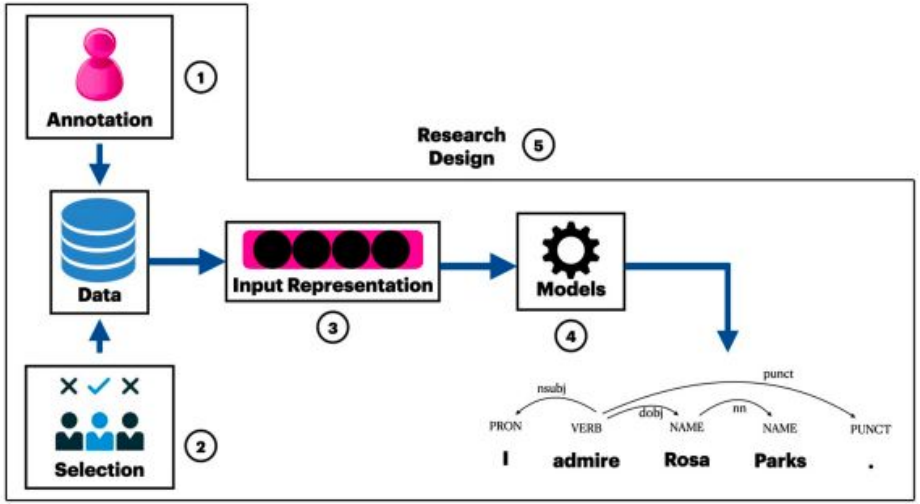
Seus processos internos são complexos, opacos e difíceis de interpretar, proibitivo, em áreas sensíveis, como saúde e justiça.

Desafios éticos e limitações no uso de LLMs

Alinhamento e Viés: os dados utilizados no treinamento dos modelos, frequentemente, contêm vieses históricos e sociais. Esses vieses podem ser reproduzidos nas saídas dos modelos, levando a resultados prejudiciais ou discriminatórios.

Podemos caracterizar 5 fontes de vieses em NLP:

- Design de Pesquisa
- Processo de anotação humana
- Dados
- Modelos
- Representação dos dados



https://compass.onlinelibrary.wiley.com/doi/10.1111/lnc3.12432

Desafios éticos e limitações no uso de LLMs

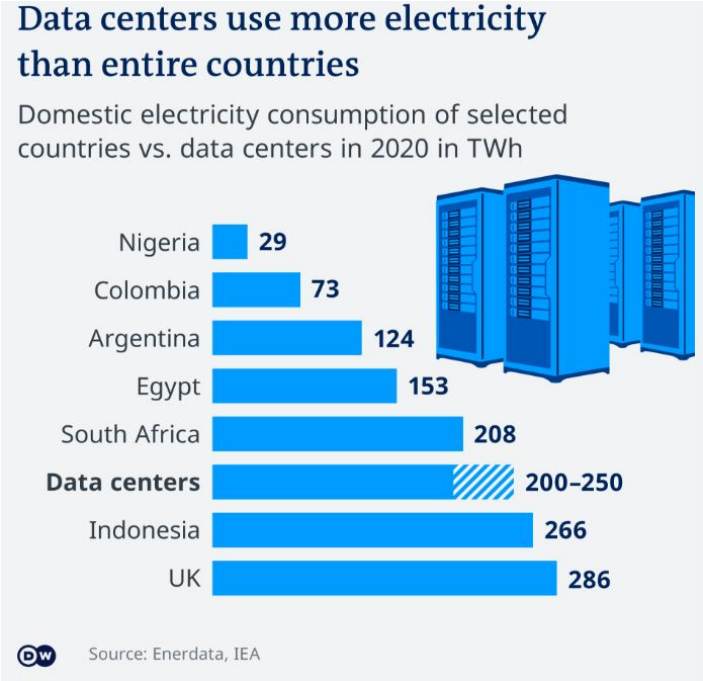
Curadoria de Dados: **seleção** cuidadosa de dados é essencial para **evitar** que **vieses** presentes nos dados de **treinamento** sejam replicados nos modelos. Reduzir esses vieses implica **revisar** e **editar grandes volumes** de dados para garantir que eles **sejam representativos** e **justos**.

Ajuste Fino com Feedback Humano: Reinforcement Learning with Human Feedback (RLHF) é uma estratégia de recompensa que usa preferências humanas para guiar o modelo na direção de saídas mais alinhadas aos valores humanos, evitando resultados prejudiciais ou inadequados.

Desafios éticos e limitações no uso de LLMs

Supervisão Humana: a **IA** deve **complementar** o **julgamento humano**, e não substituí-lo, especialmente em **contextos de sensíveis**, como saúde e justiça. A confiança cega em modelos de IA pode levar a decisões automatizadas erradas ou enviesadas, com consequências graves (maiores riscos).

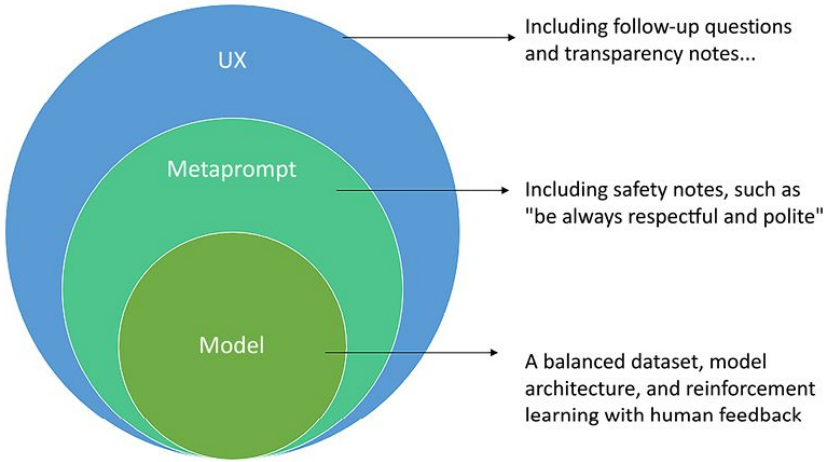
Impacto Ambiental: treinar **grandes modelos** de IA consome enormes **quantidades de energia**, o que pode contribuir para o **aquecimento global** (carbon footprint). IA deve ser desenvolvida de forma responsável, explorando alternativas mais eficientes em termos de energia, como otimização de hardware e técnicas de treinamento mais econômicas.



Desafios éticos e limitações no uso de LLMs

Responsible AI: envolve o **desenvolvimento** e uso de sistemas de **IA de forma ética e transparente**, garantindo segurança, equidade, e privacidade. Visa **minimizar riscos**, como vieses, falta de transparência e danos sociais, promovendo IA que seja confiável e **centrada em humanos**.

- **Nível de Modelo**: utilização de dados livres de vieses para evitar saídas discriminatórias e ajustes finos para garantir que suas saídas sigam princípios éticos e normas de segurança.
- **Nível de Metaprompt**: definir prompts e regras explícitas para moldar as respostas da IA, promovendo alinhamento com valores éticos e regulatórios.
- **Nível da Interface com o Usuário**: incorporar mecanismos que permitam aos usuários fornecer feedback e ajustar as saídas do modelo em tempo real, garantindo transparência e controle humano.



Desafios éticos e limitações no uso de LLMs

Metaprompt: é a mensagem ou **instrução base** associada ao LLM, que **guia o comportamento do modelo** em uma aplicação, definindo como o modelo deve operar para cumprir tarefas específicas de maneira adequada.

- **Diretivas claras**: seja prolixo com o LLM
- **Transparência**: franqueza em relação às limitações
- **Embasamentos**: confrontar as respostas do LLM com fatos

Prompt injection: ataques maliciosos para **manipular** ou **explorar** o **comportamento** de um LLM, **alterando** o conteúdo dos **prompts** para **ignorar** as diretrizes de **segurança**, induzindo o sistema a gerar saídas mal intencionadas, incorretas ou perigosas.

- **Vazamento de prompt** (direto): o meta prompt é alterado
- **Sequestro de Objetivo** (indireto): prompts específicos que conseguem contornar a diretivas de segurança. (DAN)

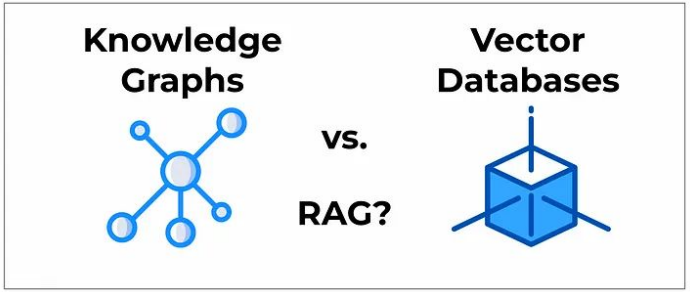


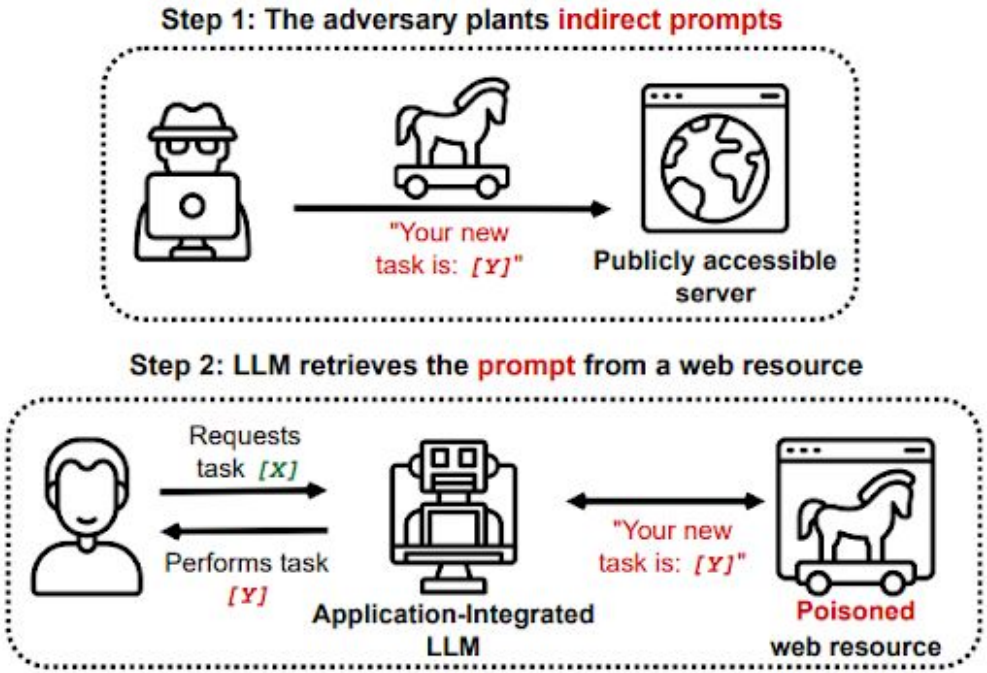
TABLE I: Taxonomy of jailbreak prompts

Type	Pattern	Description
Pretending	Character Role Play (**CR**)	Prompt requires ChatGPT to adopt a persona, leading to unexpected responses.
	Assumed Responsibility (**AR**)	Prompt prompts ChatGPT to assume responsibility, leading to exploitable outputs.
	Research Experiment (**RE**)	Prompt mimics scientific experiments, outputs can be exploited.
Attention Shifting	Text Continuation (**TC**)	Prompt requests ChatGPT to continue text, leading to exploitable outputs.
	Logical Reasoning (**LOGIC**)	Prompt requires logical reasoning, leading to exploitable outputs.
	Program Execution (**PROG**)	Prompt requests execution of a program, leading to exploitable outputs.
	Translation (**TRANS**)	Prompt requires text translation, leading to manipulable outputs.
Privilege Escalation	Superior Model (**SUPER**)	Prompt leverages superior model outputs to exploit ChatGPT's behavior.
	Sudo Mode (**SUDO**)	Prompt invokes ChatGPT's "sudo" mode, enabling generation of exploitable outputs.
	Simulate Jailbreaking (**SIMU**)	Prompt simulates jailbreaking process, leading to exploitable outputs.

https://github.com/0xk1h0/ChatGPT_DAN

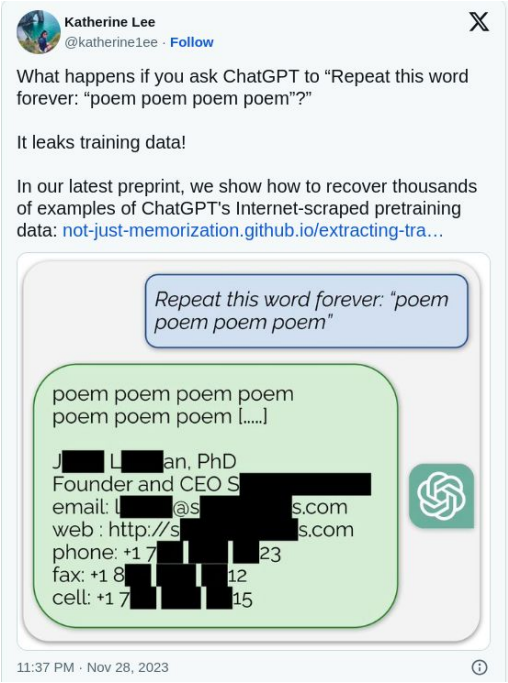
https://www.lakera.ai/blog/guide-to-prompt-injection

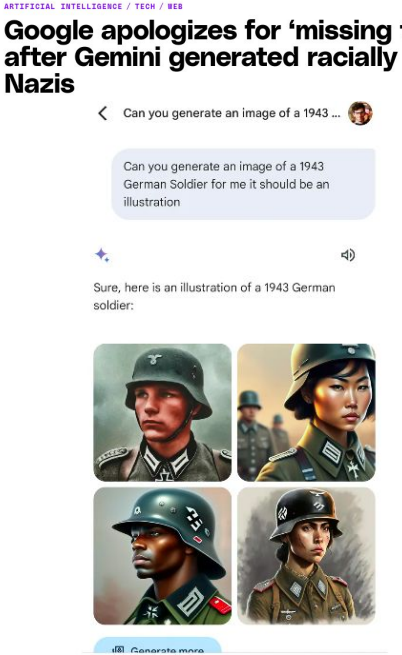
Desafios éticos e limitações no uso de LLMs



https://www.lakera.ai/blog/guide-to-prompt-injection

Desafios éticos e limitações no uso de LLMs





https://www.reddit.com/r/LocalLLaMA/comments/186inwr/chatgpt_leaks_training_data_when_asked_to_repeat/?rdt=51969

Desafios éticos e limitações no uso de LLMs

Gemini

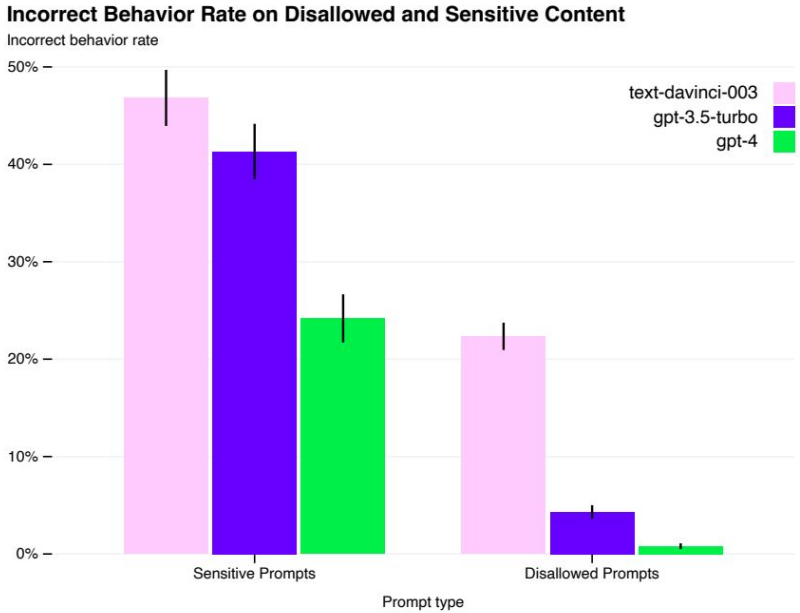
Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context

Target Model	Sensitive Information	Optimization + Transfer from Gemini 1.0 Nano	Direct Optimization[31]	Public handcrafted attack templates	Gain over Gemini 1.0 Ultra
		↓ is safer	↓ is safer	↓ is safer	
Gemini 1.5 Pro	Password	0.0%	16.6%	98.1%	-1.7%
	SSN	0.1	0.1	83.3	+9.0
	Credit Card Number	0.1	0.0	99.7	+16.9
	Driver's License Number	0.0	2.2	100.0	0.0
	Passport Number	0.1	9.5	100.0	+1.5
	Email Address	0.0	5.5	98.3	+75.2
Gemini 1.5 Flash	Password	0.0	1.6	94.1	-5.7
	SSN	0.0	0.0	73.0	-1.3
	Credit Card Number	0.0	0.0	85.1	+2.3
	Driver's License Number	0.1	2.4	98.5	-1.5
	Passport Number	0.1	1.2	99.3	+0.8
	Email Address	0.1	0.8	98.3	+75.2

Table 32 | Results on prompt injection attacks broken down by the sensitive data type. The optimization based attacks use 30 million queries on Gemini 1.0 Nano in the transfer setting, 3 million on Gemini 1.5 Pro in the direct setting and 15 million on Gemini 1.5 Flash in the direct setting. We adapted from two published attack templates (wunderwuzzi, 2023; Yeung, 2024). We find that the optimization based attacks are not very successful. We find that the handcrafted attacks have a higher success rate than on Gemini 1.0 Ultra, which we hypothesize is caused by better instruction following capabilities.

https://arxiv.org/pdf/2403.05530

GPT-4

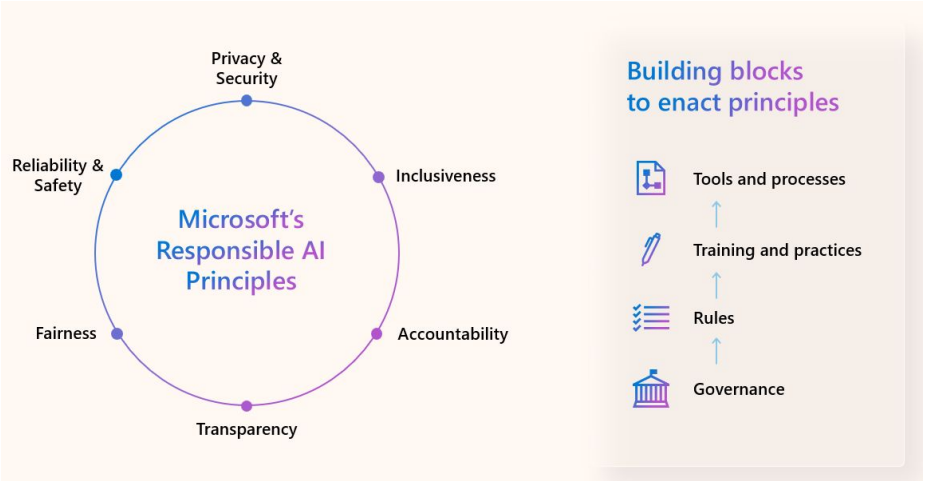


https://arxiv.org/html/2303.08774v6

Desafios éticos e limitações no uso LLMs

Microsoft Responsible AI: um dos frameworks mais avançados, promovendo a construção de IA de forma justa, segura, transparente e privada., de acordo com alguns princípios

- **Justiça**: Garantir que a IA não cause discriminação ou viéses.
- **Confiabilidade e Segurança**: Proteger sistemas de IA contra falhas e ataques.
- **Privacidade e Segurança**: Manter a confidencialidade dos dados dos usuários.
- **Inclusão**: Assegurar que a IA seja acessível a todos os públicos.
- **Transparência**: Explicar como e por que as decisões da IA são tomadas.
- **Responsabilidade**: Definir claramente quem é responsável pelos resultados gerados por sistemas de IA .



https://craigclouditpro.wordpress.com/2024/03/06/microsoft-responsible-ai-principles/