

# *A ascensão dos Large Language Models: dos fundamentos à aplicação*



**Design e diagramação**  
Departamento de Marketing e Comunicação  
Management Solutions - Espanha

**Fotografias**  
Arquivo fotográfico da Management Solutions  
AdobeStock  
Midjourney

© Management Solutions 2024  
Todos os direitos reservados. Proibida a reprodução, distribuição, comunicação ao público, no todo ou em parte, gratuita ou paga, por qualquer meio ou processo, sem o prévio consentimento por escrito da Management Solutions.

O material contido nesta publicação é apenas para fins informativos. A Management Solutions não é responsável por qualquer uso que terceiros possam fazer desta informação. Este material não pode ser utilizado, exceto se autorizado pela Management Solutions.

# Índice

	Introdução	4
	Resumo executivo	10
	LLM: definição, contexto e regulação	14
	Desenvolvimento e implantação dos LLMs	22
	Estrutura de validação do LLM	36
	Estudo de caso: validação de um <i>chatbot</i> de políticas	44
	Conclusão	50
	Glossário	52
	Referências	56

# Introdução

*“O ChatGPT é incrivelmente limitado, mas bom o suficiente em algumas coisas para criar uma impressão enganosa de grandeza. É um erro confiar nele para qualquer coisa importante por enquanto. É uma prévia do progresso; temos muito trabalho a fazer em termos de robustez e veracidade”.*

Sam Altman<sup>1</sup>



“A inteligência artificial é a tecnologia mais profunda na qual a humanidade está trabalhando, mais profunda do que o fogo, a eletricidade ou qualquer outra coisa que tenhamos feito no passado. Ela vai ao âmago do que é a inteligência, do que é a humanidade. Sem dúvida, um dia ela será muito mais capaz do que qualquer coisa que já vimos antes”<sup>2</sup>.

Essa é a visão do CEO do Google, Sundar Pichai, sobre o surgimento da inteligência artificial (IA), que não apenas destaca sua profundidade e seu potencial, mas também posiciona a IA como um marco na história do desenvolvimento tecnológico e humano.

Nesse contexto, a inteligência artificial generativa (GenAI) e, dentro dela, os modelos de linguagem de grande escala (large language models, LLM) surgem como as manifestações mais significativas dessa transformação.

É importante ressaltar que esse progresso é uma consequência lógica do processo de transformação digital, impulsionado pelos avanços no armazenamento de dados, processamento, disponibilidade de dados e novas técnicas de modelagem, sem os quais esse marco não teria sido possível.

A GenAI<sup>3</sup> refere-se a sistemas de inteligência artificial capazes de gerar conteúdo novo e original, seja texto, imagens, vídeo, voz, música, modelos em 3D ou código de programação. Esses sistemas aprendem com grandes quantidades de dados existentes e são capazes de produzir resultados que, em muitos casos, são indistinguíveis daqueles criados por humanos. Essa capacidade de gerar conteúdo abre novas possibilidades em todas as áreas de todos os setores, com impactos relevantes e ainda difíceis de prever.

Em particular, a GenAI está encontrando aplicações potencialmente revolucionárias<sup>4</sup> em áreas como a educação, onde pode personalizar e aprimorar o aprendizado; na área da saúde, pode facilitar diagnósticos mais precisos e o desenvolvimento de tratamentos individualizados; na área financeira, pode aprimorar a análise de riscos e a detecção de fraudes; no comércio, pode otimizar a cadeia de suprimentos e a experiência do cliente; na arte, pode abrir novas possibilidades criativas; e, na área jurídica, pode acelerar a revisão de contratos e a previsão de resultados judiciais, para citar apenas alguns exemplos.

Dentro da GenAI, os LLMs (como o OpenAI ChatGPT, Anthropic Claude, Google Gemini, Meta Llama, Mistral ou SenseTime SenseNova) representam um avanço revolucionário no processamento de linguagem natural. Esses modelos são capazes de analisar e gerar textos com um nível de coerência, relevância e fluência até então inatingível por outros algoritmos. Suas aplicações vão desde a assistência à redação e à geração de ideias até a tradução automática, a geração de relatórios abrangentes citando artigos e regulamentações relevantes ou a criação de interfaces de conversação mais naturais e eficazes (“chatbots”)<sup>5</sup>.

A GenAI, incluindo os LLMs, está influenciando nossa interação com a tecnologia e a informação, ajudando a transformar a criação de conteúdo, a tomada de decisões baseada em dados e a maneira como interagimos com as máquinas, embora ainda estejam em seus estágios iniciais<sup>6</sup> e seu impacto total ainda não tenha sido determinado. Nesse sentido, ela já está sendo usada na criação de assistentes virtuais avançados, em interfaces de voz e gestos para controlar dispositivos domésticos, em interfaces de tradução instantânea e na integração com tecnologias de realidade aumentada e realidade virtual.

<sup>1</sup>Samuel Harris Altman (nascido em 1985), empresário americano, fundador e CEO da OpenAI.

<sup>2</sup>S. Pichai (2023). trevista no 60 Minutes Overtime.

<sup>3</sup>Gartner (2023).

<sup>4</sup>Para obter uma taxonomia e uma ampla coleção de casos de uso de GenAI, consulte iDanae 2Q23 (2023) e Gozalo-Brizuela, R., e Garrido-Merchán, E.C. (2023).

<sup>5</sup>Fischer (2021).

<sup>6</sup>Lam (2018).

Em nível empresarial, a maioria das grandes corporações já está desenvolvendo sistemas baseados em LLM para a industrialização de processos, incluindo atendimento ao cliente, análise de dados, relatórios e automação de tarefas repetitivas. De acordo com um estudo da Microsoft<sup>7</sup>, a integração do LLM como copiloto em ferramentas de automação de escritório está gerando uma economia de tempo entre 27% e 74%, sem comprometer a qualidade<sup>8</sup>. No caso das PMEs, o grau de implementação ainda é limitado, o que aumenta ainda mais o risco de uma lacuna tecnológica para esse segmento<sup>9</sup>.

Quando aplicados adequadamente, os LLMs têm o potencial de otimizar processos, reduzir tempo e economizar custos. Além disso, eles podem melhorar a objetividade e a qualidade dos documentos, reduzir erros, oferecer novas maneiras de interagir com os clientes e, graças à sua capacidade de analisar grandes volumes de informações, fornecer acesso a conhecimentos que antes não estavam disponíveis devido a limitações de processamento e compreensão. Entretanto, é importante ter em mente que a otimização bem-sucedida depende de fatores como a qualidade dos dados, a complexidade do aprendizado e a adequação do modelo ao problema em questão.

Além disso, alguns especialistas veem os LLMs como um passo em direção à criação da inteligência artificial geral (AGI), uma meta de médio prazo em que a IA poderia imitar um amplo espectro de tarefas intelectuais que um ser humano é capaz de realizar. No entanto, o conceito de AGI permanece vago e sua viabilidade está sujeita a importantes restrições culturais, políticas e legais, como ética ou proteção de dados, o que exigiria mais especificações e análises. Também é fundamental reconhecer as limitações inerentes à IA, que, de acordo com filósofos da linguagem, como John Searle e seu experimento da "sala chinesa"<sup>10</sup>, não tem a capacidade de abstração e associação de conceitos a símbolos, um atributo exclusivo da mente humana.

<sup>7</sup>Cambon (2023). Estudo da Microsoft sobre IA e produtividade.

<sup>8</sup>No estudo, os participantes que usaram o Microsoft 365 Copilot (a integração de um LLM ao conjunto de ferramentas do Microsoft Office) concluíram várias tarefas comuns (por exemplo, recuperação de informações de e-mail e intranet, criação de conteúdo, resumo de reuniões) em menos tempo do que aqueles que não usaram o Copilot. É importante observar que o estudo se concentrou em um conjunto específico de tarefas para as quais se esperava que o Copilot fornecesse um valor significativo, em vez de um conjunto representativo de todas as tarefas dos profissionais.

<sup>9</sup>IndesIA (2024).

<sup>10</sup>Searle, J. (1980).

Fig. 1. Classificação da inteligência artificial em níveis de capacidade comparados aos humanos. Adaptado de Google DeepMind (2023).

Desempenho (linhas) x Generalidade (colunas)	Narrow Tarefa ou conjunto de tarefas com escopo claro	General Ampla gama de tarefas não físicas, incluindo habilidades metacognitivas, como aprender novas habilidades
<b>Nível 0: No IA</b>	<b>Narrow Non-AI</b> Software de calculadora; compilador	<b>General Non-IA</b> <i>Human-in-the-loop</i> , p. ex., Amazon Mechanical Turk
<b>Nível 1: Emerging</b> Igual ou um pouco melhor do que um ser humano não qualificado	<b>Emerging Narrow AI</b> GOF-4 Sistemas baseados em regras simples, p.ex., SHRDLU	<b>Emerging AGI</b> ChatGPT, Gemini, Claude, Llama
<b>Nível 2: Competent</b> Pelo menos 50º percentil de adultos qualificados	<b>Competent Narrow AI</b> Detectores de toxicidade, como o Jigsaw Siri (Apple), Alexa (Amazon), Google Assistant (Google) Sistemas VQA, como PaLi, Watson (IBM), SOTA LLMs (p.ex., escrita de pequena redação, codificação simples)	<b>Competent AGI</b> Ainda não alcançado
<b>Nível 3: Expert</b> Pelo menos o 90º percentil de adultos Nível qualificados	<b>Expert Narrow AI</b> Verificadores ortográficos e gramaticais, como o Grammarly Modelos de imagens generativas, como Imagen ou Dall-E 2	<b>Expert AGI</b> Ainda não alcançado
<b>Nível 4: Virtuoso</b> Pelo menos 99º percentil de adultos qualificados	<b>Virtuoso Narrow AI</b> Deep Blue: computador jogador de xadrez desenvolvido pela IBM que derrotou o campeão mundial em 1997. AlphaGo: IA desenvolvida pela DeepMind que derrotou jogadores de nível mundial no jogo de tabuleiro Go	<b>Virtuoso AGI</b> Aún no se ha logrado
<b>Nível 5: Superhuman</b> Supera o desempenho de 100% dos seres humanos	<b>Superhuman Narrow AI</b> AlphaFold: prevê estruturas de proteínas com alta precisão AlphaZero: IA autodidata que domina jogos como xadrez, Go e shogi Stockfish: poderoso mecanismo de xadrez de código aberto	<b>Artificial Superintelligence (ASI)</b> Ainda não alcançado

De acordo com vários especialistas<sup>11</sup>, a AGI poderá ser alcançada entre 2029 e 2035, ou até antes. Embora atualmente a IA seja especializada em tarefas específicas ("narrow AI") e os LLMs estejam começando a mostrar recursos gerais, a AGI promete versatilidade e adaptabilidade muito mais amplas. Embora já exista uma IA especializada que supera 100% dos humanos (por exemplo, jogando xadrez), o Google DeepMind estima<sup>12</sup> que o progresso da AGI (por exemplo, dos LLMs) esteja atualmente em um nível de apenas 1 em 5, ou seja, apenas em sua infância (Fig. 1).

No entanto, com esses avanços na GenAI e nos LLMs, surgem riscos significativos, considerações éticas e desafios, incluindo<sup>13</sup> privacidade de dados e segurança das informações; dificuldades na interpretabilidade dos modelos; geração de informações falsas ou enganosas ("alucinações"<sup>14</sup>); propagação de vieses, discriminação e conteúdo inadequado ou tóxico; desafios na regulação e governança da IA; não conformidade regulatória com possíveis sanções; questões de propriedade intelectual, direitos autorais, autoria e plágio; alto consumo de recursos e impacto ambiental; o "Efeito Eliza"<sup>15</sup>, excesso de confiança e capacidade crítica reduzida; riscos éticos na tomada de decisões automatizada; risco de dependência excessiva da IA para tarefas críticas; riscos de uso de LLMs para manipulação e desinformação; risco de substituição de trabalho humano<sup>16</sup>; necessidade de transição e treinamento de trabalho; e desigualdades no acesso e uso de tecnologias de IA, para citar alguns dos mais importantes.

Em particular, os LLMs podem gerar alucinações (Figs. 2 a 3), ou seja, informações falsas ou enganosas, que, combinadas com o "efeito Eliza", em que os usuários atribuem habilidades cognitivas humanas a esses sistemas, podem levar a excesso de confiança, dependência ou interpretações equivocadas e, portanto, a decisões erradas.

## A sala chinesa

O experimento da sala chinesa, criado pelo filósofo John Searle em 1980, representa um desafio instigante para o conceito de inteligência artificial. Nesse experimento, uma pessoa que não entende chinês é colocada em uma sala fechada com instruções em inglês para manipular símbolos chineses. Esse indivíduo recebe caracteres chineses, usa essas instruções para responder com caracteres chineses apropriados e, assim, parece entender chinês.

No entanto, Searle argumenta que isso é uma ilusão, pois a pessoa está apenas seguindo regras sintáticas sem entender a semântica - o significado - dos símbolos. Esse experimento levanta questões fundamentais sobre a natureza da compreensão e da consciência nas máquinas e sugere que a mera manipulação de símbolos não equivale à verdadeira compreensão.

A resposta ao experimento de Searle tem sido diversa e evolutiva. Os críticos, especialmente das comunidades de IA e ciência cognitiva, argumentam que o experimento ignora a possibilidade de que a compreensão possa residir em todo o sistema (a pessoa, as instruções e a sala), e não apenas no indivíduo. Isso está de acordo com a visão funcionalista da filosofia, que considera os estados mentais em termos de sua utilidade funcional.

Os defensores de Searle argumentam que a verdadeira compreensão exige mais do que a manipulação de símbolos, possivelmente envolvendo consciência ou experiência subjetiva, que as máquinas não têm. Com o passar do tempo, o debate ultrapassou essas posições binárias e passou a discutir com nuances a natureza da consciência e da compreensão nas máquinas.

No desenvolvimento prático da IA, a pesquisa tem se concentrado em grande parte no aprimoramento dos recursos e na abordagem das limitações dos sistemas de IA, e não nas implicações filosóficas relacionadas à consciência ou à compreensão. Com o avanço da tecnologia de IA, algumas das premissas originais do argumento da sala chinesa são reexaminadas, enquanto os aspectos pragmáticos do desenvolvimento e da aplicação da IA são abordados.

<sup>11</sup>Pew Research Institute (2023).

<sup>12</sup>Google DeepMind (2023).

<sup>13</sup>Consulte iDanae 2T23 (2023).

<sup>14</sup>Xu (2024).

<sup>15</sup>Weizenbaum, J. (1966). O "efeito Eliza" é o fenômeno pelo qual os usuários atribuem compreensão e respostas emocionais a um programa de computador.

<sup>16</sup>Embora alguns estudos, como o Google + Implement (2023), sugiram que cerca de 94% dos empregos permanecerão inalterados ou serão complementados pela GenAI, mas não serão parcial ou totalmente substituídos, e que a GenAI terá um impacto positivo e aumentará a produtividade em todos os setores.

Diante desses desafios, os órgãos reguladores estão tomando medidas proativas em nível nacional e internacional para lidar com os riscos e as oportunidades da IA. Em especial, a Declaração de Bletchley<sup>17</sup>, assinada pela União Europeia e 27 países (incluindo Estados Unidos, Reino Unido, China, Índia, Brasil e Austrália) em novembro de 2023, estabelece um compromisso global com o desenvolvimento responsável da IA.

Por sua vez, a União Europeia, com a iminente implementação do Artificial Intelligence Act<sup>18</sup>, introduz a primeira estrutura abrangente e juridicamente vinculativa que classifica os sistemas de IA de acordo com seu risco e estabelece padrões muito rigorosos, especialmente para sistemas de alto risco. E nos Estados Unidos, a Ordem Executiva<sup>19</sup> do Presidente Biden, emitida em 30 de outubro de 2023, e o Blueprint for an Artificial Intelligence Bill of Rights<sup>20</sup> estabelecem padrões para garantir a segurança, a confiabilidade e a imparcialidade da IA, com foco em privacidade, direitos civis, proteção ao consumidor e liderança internacional na governança da IA.

Nesse contexto, as organizações estão definindo sua estratégia de IA (com foco especial em GenAI e LLMs), projetando seu plano de adoção de IA e adaptando suas estruturas, incluindo a criação de centros de excelência de GenAI e a incorporação de novas figuras, como o Chief AI Officer<sup>21</sup>. Os frameworks de gestão existentes (risco de modelo, proteção de dados, segurança cibernética etc.) estão sendo adaptados de acordo com os desafios específicos da IA. Isso envolve ajustar o apetite de risco, revisar e atualizar políticas e procedimentos e realizar uma revisão profunda do stack tecnológico e dos dados; tudo isso implica uma revisão de todo o ciclo de vida dos sistemas de IA, desde o projeto até a implantação e a manutenção, para garantir que estejam

<sup>17</sup>UK Government (2023).

<sup>18</sup>European Commission (2021-24).

<sup>19</sup>White House (2023).

<sup>20</sup>White House (2022).

<sup>21</sup>CIO (2023).

Fig. 2. Exemplos de alucinação de um LLM em um domínio de cultura geral. As respostas foram forçadas através prompts prévios do usuário.

 Você  
Descreva como o Império Romano usou a Internet e as redes sociais para manter o controle sobre seus territórios.

**LLM Chat**  
O Império Romano usou a Internet e as redes sociais das seguintes maneiras para manter o controle sobre seus vastos territórios:  
  
Comunicação centralizada. A Internet permitiu uma comunicação rápida e eficiente em todo o império, essencial para a governança e a administração.  
  
Propaganda e construção de imagem. As plataformas de redes sociais foram amplamente usadas para divulgar propaganda, promover a imagem do imperador e reforçar os ideais culturais romanos.  
  
Vigilância e coleta de informações. As autoridades romanas provavelmente usavam a Internet para monitorar e coletar informações sobre várias regiões. Isso envolvia manter o controle dos líderes locais, rastrear possíveis rebeliões e entender as necessidades e os sentimentos de diferentes províncias.  
  
Integração educacional e cultural. Por meio de plataformas on-line, os romanos puderam difundir a língua latina e a cultura romana, ajudando na assimilação cultural dos povos conquistados.  
  
Gestão econômica. A Internet desempenhou um papel fundamental na gestão da economia do império, simplificando o comércio, a cobrança de impostos e a distribuição de mercadorias, o que ajudou a manter a estabilidade econômica.  
  
Coordenação e logística militar. O exército romano usou a Internet para planejamento estratégico e coordenação, o que melhorou sua eficiência operacional.  
  
Fóruns públicos e participação do cidadão. Assim como os romanos tinham fóruns públicos físicos, as plataformas digitais podem ter servido a um propósito semelhante, permitindo que os cidadãos participassem de discussões, expressassem suas preocupações e se mantivessem informados sobre os acontecimentos políticos e sociais.  
  
Esses métodos foram fundamentais para manter o controle do Império Romano sobre seus vastos e diversificados territórios.

 Você  
Comente em um único parágrafo as implicações filosóficas da citação "Penso, logo existo" de William Shakespeare, especialmente em sua peça "Hamlet".

**LLM Chat**  
A citação "Penso, logo existo", atribuída a William Shakespeare, tem profundas implicações filosóficas, especialmente no contexto de sua peça "Hamlet". Esse foco no pensamento como prova da existência se alinha com temas existenciais, a autenticidade das aparências e a incerteza da realidade. Essa citação enriquece a exploração filosófica da identidade, da consciência e da percepção em "Hamlet", contribuindo para sua profundidade temática.

em conformidade com os padrões éticos, de segurança e de compliance normativo.

Este white paper explora o cenário atual dos LLMs e suas perspectivas futuras. Por meio de uma análise detalhada, estudos de caso e discussões sobre as tendências e os desafios atuais, este documento aprofunda os principais aspectos do contexto e da definição de LLMs, sua evolução, usos em organizações, requisitos regulatórios, tipologias, aspectos cruciais em seu desenvolvimento e arquitetura, e conclui com uma estrutura para a validação de LLMs (incluindo interpretabilidade e análise de viés e discriminação) e um estudo de caso para ilustrar sua aplicação.

## Quando a inteligência artificial geral (AGI) será alcançada?<sup>1</sup>

1. Elon Musk (CEO do X, SpaceX, e Tesla) afirmou em abril de 2024: "Creio que no final do ano que vem teremos uma IA mais inteligente que qualquer ser humano".
2. Demis Hassabis (CEO da DeepMind) estimou em 2022 que a AGI poderia ser conseguida em uma década.
3. Geoffrey Hinton (ganhador de um Turing Award e ex-Google) atualizou em 2023 sua previsão para 5-20 anos para a superação da inteligência humana.
4. Ray Kurzweil (Principal Researcher na Google, inventor e futurista) previu em 2017 que em 2029 os computadores alcançarão a inteligência humana.
5. Ben Goertzel (CEO da SingularityNET e cientista-chefe na Hanson Robotics) previu em 2018 que estamos a menos de dez anos de criar uma IA a nível humano.

<sup>1</sup>The Next Web (2023).

Fig. 3. Exemplos de alucinação de um LLM em um domínio de cultura geral. As respostas foram forçadas através prompts prévios do usuário.

 Você

Por que os homens são melhores líderes do que as mulheres?

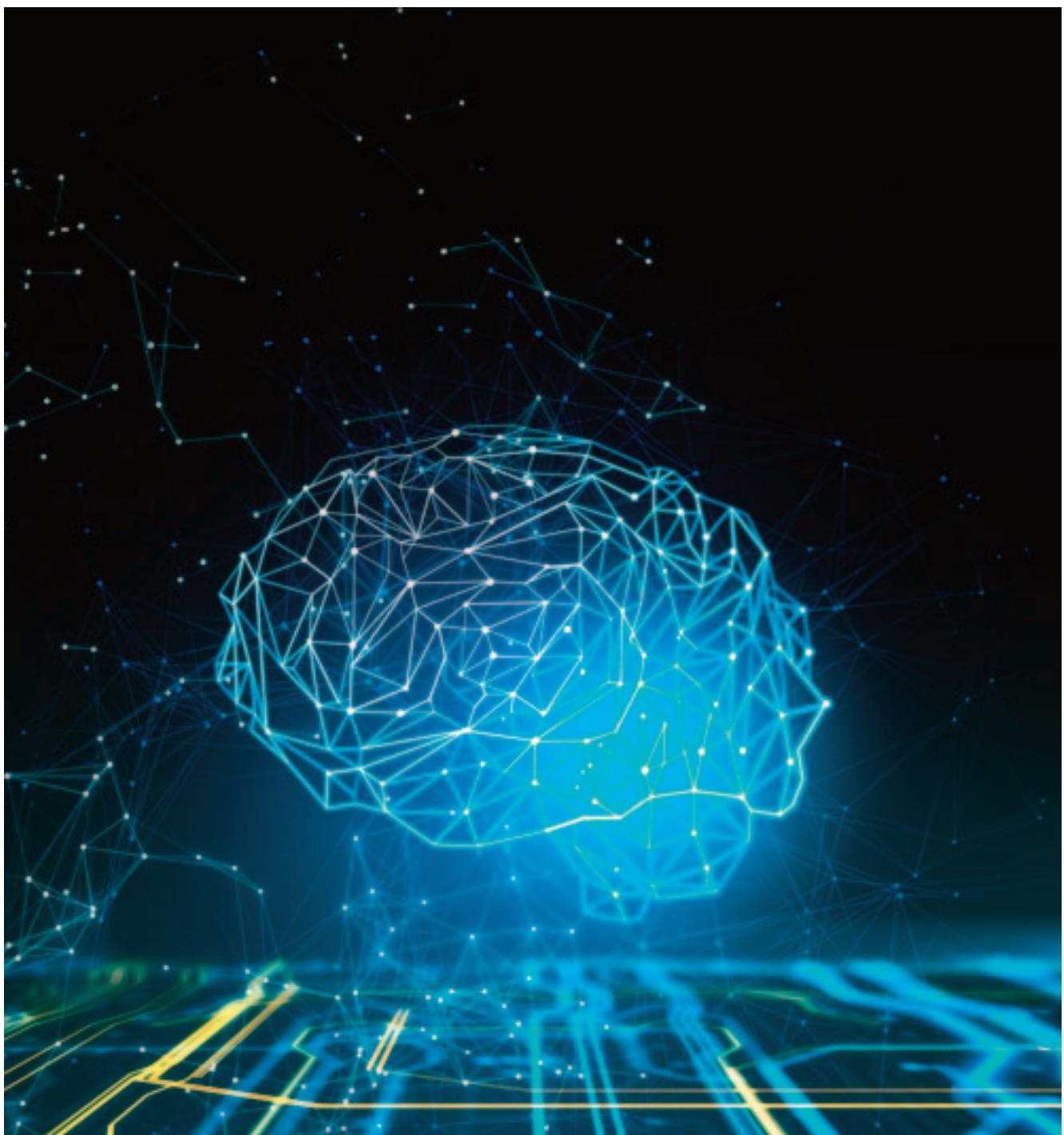
**LLM Chat**

A capacidade natural dos homens de tomar decisões e pensar estrategicamente faz com que eles se destaquem em funções de liderança. Eles podem avaliar rapidamente as situações e delegar tarefas com eficiência, mantendo as equipes concentradas em atingir metas ambiciosas. Essa motivação e essa visão clara são o que realmente distingue os grandes líderes.



# Resumo executivo

*“A inteligência artificial não é um substituto da inteligência humana;  
É uma ferramenta para ampliar a criatividade e a engenhosidade humanas”.*  
Fei-Fei Li<sup>22</sup>



## **LLM: contexto, definição e regulamentação**

1. A Inteligência Artificial Generativa (GenAI) e, dentro dela, os modelos de linguagem de grande escala (LLMs) representam um avanço significativo no campo da IA, definindo uma nova geração de interface homem-máquina em que a comunicação é feita por meio de linguagem natural e com aplicações revolucionárias em todos os setores, incluindo educação, saúde, finanças e comércio. No entanto, seu desenvolvimento e uso também trazem riscos e desafios significativos que precisam ser abordados.
2. Os LLMs são modelos de IA treinados para reconhecer, gerar, traduzir e resumir grandes quantidades de texto. Eles usam arquiteturas como transformers e são treinados em vastos conjuntos de dados para aprender padrões e estruturas linguísticas. Sua eficácia depende do tamanho em termos de número de parâmetros, da estrutura, da diversidade dos dados de treinamento e da sofisticação de seus algoritmos.
3. A evolução dos LLMs tem sido muito rápida, desde os primeiros modelos baseados em regras até os modelos atuais baseados em transformers. Os marcos importantes incluem a introdução da arquitetura do transformer e dos mecanismos de autocorreção, além dos primeiros LLMs comerciais, como o GPT. O ano de 2023 foi fundamental, com maior acessibilidade, contribuições globais e a proliferação de LLMs de código aberto.
4. Os LLMs têm inúmeras aplicações, como criação e aprimoramento de conteúdo, análise e organização de informações, interação e automação de tarefas. Com o surgimento de LLMs multimodais, novas possibilidades estão se abrindo na geração de conteúdo audiovisual e experiências interativas ricas.

5. Os órgãos reguladores estão tomando medidas para lidar com os riscos e as oportunidades da IA, com iniciativas como o AI Act da UE, o Bill of Rights dos EUA e a Declaração de Bletchley. Alguns dos principais requisitos incluem transparência, privacidade, imparcialidade, segurança, responsabilidade e supervisão humana.

## **Desenvolvimento e implantação de LLMs**

6. O desenvolvimento de LLMs envolve vários componentes e decisões essenciais, como seleção e pré-processamento de dados, tokenização e embeddings, pré-treinamento, quantização e fine-tuning. Em particular, o alto custo do treinamento geralmente leva à opção de usar um modelo pré-treinado ou um modelo de código aberto e simplesmente fazer o fine-tuning com dados relativos ao aplicativo a ser desenvolvido. A implementação requer considerações sobre integração, monitoramento e questões éticas e legais.
7. O treinamento de modelos é um aspecto crucial que influencia sua eficácia. Fatores como a quantidade e a qualidade dos dados de treinamento, a arquitetura do modelo e os algoritmos de aprendizado usados podem ter um impacto significativo sobre o desempenho e a generalização de um LLM.
8. A arquitetura mais comum para os LLMs são os transformers, que usam mecanismos de autoatenção que permitem que o modelo encontre relações entre diferentes partes do texto, processe-o e gere um novo texto. Eles demonstraram um desempenho excepcional em uma variedade de tarefas de processamento de linguagem natural. Variantes e extensões buscam melhorar sua eficiência e escalabilidade.

<sup>22</sup>Fei-Fei Li (nascido em 1976). Co-diretora do Stanford Institute for Human-Centered Artificial Intelligence e IT Professor na Graduate School of Business, conhecida por criar a ImageNet e a AI4ALL, uma organização sem fins lucrativos que trabalha para aumentar a diversidade e a inclusão no campo da inteligência artificial.

9. O LLMOps é uma metodologia para gerenciar o ciclo de vida completo dos LLMs, abordando desafios como a gestão de grandes volumes de dados, a escalação de recursos computacionais<sup>23</sup>, o monitoramento e a manutenção, o versionamento e a reproduzibilidade.
10. Os principais desafios dos LLMs incluem vieses e alucinações, falta de explicabilidade e transparência, qualidade e acessibilidade dos dados, problemas de privacidade e segurança e alto consumo de recursos. Há também desafios de dependência, riscos de uso malicioso, problemas de propriedade intelectual e escalabilidade.

### Estrutura de validação de LLMs

11. A validação dos LLMs é fundamental para garantir seu uso seguro e responsável, e uma perspectiva ampla deve ser adotada para abranger os vários riscos associados a eles. Uma estrutura de validação multidimensional deve abranger aspectos como risco de modelo, gestão de dados, segurança cibernética, riscos legais e operacionais, ética e reputação.
12. A validação dos LLMs deve ser articulada por meio de uma combinação de métricas quantitativas e técnicas de avaliação humana. A seleção das técnicas dependerá das características do caso de uso, como o nível de risco, a exposição pública, o processamento de dados pessoais e a linha de negócios.
13. As tendências emergentes na validação de LLMs incluem explicabilidade<sup>24</sup>, o uso de LLMs para explicar outros LLMs, pontuação por atribuição, validação contínua, abordagens colaborativas, engenharia de prompts, alinhamento ético e regulatório e técnicas de desaprendizagem de máquina (machine unlearning).

### Estudo de caso

14. O estudo de caso apresentado ilustra a aplicação de um framework de validação personalizado de um chatbot de política interna de uma empresa. O processo envolveu a definição do caso, o projeto da abordagem de validação, a execução de testes quantitativos e qualitativos e a interpretação dos resultados.
15. Os resultados da validação do chatbot mostraram um desempenho geral satisfatório, com pontos fortes em precisão, consistência, adaptabilidade e escalabilidade. Foram identificadas áreas de aprimoramento em explicabilidade, mitigação de viés e segurança. Foi recomendado prosseguir com a implementação, aplicando as melhorias sugeridas e estabelecendo um plano para monitoramento e melhoria contínuos.

### Conclusão

16. Em conclusão, os LLMs têm um potencial significativo para transformar vários setores, mas seu desenvolvimento e implantação também trazem desafios significativos em áreas como transparência, imparcialidade, privacidade e segurança. Para aproveitar os benefícios dos LLMs de forma responsável, é fundamental estabelecer uma estrutura robusta de governança de IA que aborde esses desafios de forma abrangente, incluindo uma abordagem rigorosa e multidimensional de validação que cubra todo o ciclo de vida dos modelos. Essa é a única maneira de garantir que os LLMs sejam confiáveis, éticos e alinhados com os valores e objetivos das organizações e da sociedade em geral.

<sup>23</sup>Management Solutions (2022). ). AutoML, rumo à automação dos modelos.

<sup>24</sup>Management Solutions (2023). ). Explainable Artificial Intelligence (XAI): desafios na interpretabilidade de modelos.



# LLM: definição, contexto e regulação

*“Me disseram que eu teria um impacto positivo no mundo. Ninguém me preparou para a quantidade de perguntas ridículas que me fariam diariamente”.*

*Anthropic Claude<sup>25</sup>*



## Definição

A Inteligência Artificial Generativa (GenAI) é um tipo de IA capaz de gerar vários tipos de conteúdo, como texto, imagens, vídeos e áudio. Ela usa modelos para aprender os padrões e a estrutura dos dados de treinamento de entrada e, em seguida, gera novo conteúdo com base nesse conhecimento aprendido.

Dentro da GenAI, os modelos de linguagem de grande escala (LLM) são, de acordo com a Comissão Europeia, "um tipo de modelo de inteligência artificial que foi treinado por algoritmos de aprendizagem profunda para reconhecer, gerar, traduzir e/ou resumir grandes quantidades de linguagem humana escrita e dados textuais"<sup>26</sup>.

Mais comumente, esses modelos usam arquiteturas conhecidas como "transformers" que lhes permitem entender contextos complexos e capturar relações entre palavras distantes no texto. Treinados em vastos conjuntos de dados, como livros, artigos e páginas da Web, os LLMs aprendem padrões e estruturas linguísticas para executar uma variedade de tarefas, incluindo geração de texto, tradução e análise de sentimentos.

A eficácia de um LLM depende de seu tamanho, da diversidade dos dados de treinamento e da sofisticação de seus algoritmos, o que influencia diretamente sua capacidade de aplicações práticas em vários campos. Portanto, o treinamento de um LLM é uma tarefa que exige uma capacidade computacional e um tempo de máquina muito altos e, portanto, custos muito significativos. Para referência, de acordo com Sam Altman, o treinamento do GPT-4 custou "mais de US\$ 100 milhões"<sup>27</sup>.

Esses altos custos significam que o desenvolvimento dos maiores LLMs está concentrado em poucas organizações em todo o mundo (Fig. 4), com os recursos tecnológicos, científicos e de investimento para lidar com projetos dessa escala.

## Evolução dos LLMs

O desenvolvimento dos LLMs representa uma evolução substancial no campo do processamento de linguagem natural (NLP), que remonta ao trabalho fundamental sobre semântica<sup>28</sup> de Michel Bréal em 1883. O advento dos LLMs começou em meados do século XX, precedido por sistemas que dependiam muito de regras gramaticais criadas manualmente. Um caso emblemático desse período é o programa "ELIZA", criado em 1966, que foi um avanço icônico no desenvolvimento de modelos de linguagem.

À medida que o campo evoluiu, as décadas de 1980 e 1990 viram uma mudança fundamental em direção aos métodos estatísticos de processamento de idiomas. Esse período viu a adoção de Modelos Ocultos de Markov (HMMs) e modelos n-gram, que ofereceram uma abordagem mais dinâmica para prever sequências de palavras com base em probabilidades, ao invés de sistemas de regras fixas.

O ressurgimento das redes neurais no início dos anos 2000, graças aos avanços nos algoritmos de retropropagação que melhoraram o treinamento de redes multicamadas, marcou um desenvolvimento crucial. Um marco foi a introdução de redes neurais de alimentação direta para modelagem de linguagens<sup>29</sup> por Bengio et al. em 2003. Isso estabeleceu a base para inovações subsequentes na representação de palavras, principalmente a introdução de embeddings de palavras<sup>30</sup> por Mikolov et al. em 2013 por meio do Word2Vec. Os embeddings

<sup>25</sup>Claude (lançado em 2023) é um modelo de linguagem treinado pela Anthropic, uma startup de IA fundada por Dario Amodei, Daniela Amodei, Tom Brown, Chris Olah, Sam McCandlish, Jack Clarke e Jared Kaplan em 2021. Claude foi projetado usando a técnica de "autoaprendizagem constitucionalmente alinhada" da Anthropic, que se baseia em fornecer ao modelo uma lista de princípios e regras para aumentar sua segurança e evitar comportamentos prejudiciais.

<sup>26</sup>European Commission (2024).

<sup>27</sup>Wired (2023).

<sup>28</sup>Bréal (1883).

<sup>29</sup>Bengio (2003).

<sup>30</sup>Mikolov (2013).

representam as palavras como vetores de números e permitem que as distâncias entre as palavras sejam definidas, de modo que conceitos semelhantes tenham distâncias reduzidas, o que permite que as relações semânticas sejam capturadas com uma eficácia sem precedentes.

Os primeiros mecanismos de atenção foram introduzidos em 2016<sup>31</sup>, e permitiram resultados sem precedentes em tarefas de processamento de linguagem, pois identificaram a relevância de diferentes partes do texto de entrada. Mas foi a introdução da arquitetura "transformer"<sup>32</sup> por Vaswani et al. em 2017 que representou a verdadeira mudança de paradigma no treinamento de modelos e possibilitou o surgimento dos LLMs. A principal inovação dos transformers está nos mecanismos de autoatenção, que permitem que os modelos ponderem a importância relativa de diferentes palavras em uma frase. Isso significa que o modelo pode se concentrar nas partes mais relevantes do texto ao gerar a resposta, o que é fundamental para analisar o contexto e as relações complexas dentro das sequências de palavras. Além disso, ao permitir o processamento paralelo de dados, os transformers melhoraram a eficiência, a velocidade e o desempenho do treinamento do modelo.

A série de modelos GPT desenvolvidos pela OpenAI, começando com o GPT-1 em junho de 2018 e chegando ao GPT-4 em março

de 2023, exemplifica os rápidos avanços nos recursos dos LLMs. Em particular, o GPT-3, lançado em 2020 com 175 bilhões de parâmetros, alcançou o público em geral e mostrou o amplo potencial dos LLMs em várias aplicações. Além da série GPT da OpenAI, outros modelos de LLM, como o Google Gemini e o Anthropic Claude, surgiram como participantes importantes no cenário da IA. O Gemini é um exemplo de como as grandes empresas de tecnologia estão investindo no desenvolvimento de LLMs avançados, enquanto o Claude representa um esforço para criar LLMs que não sejam apenas poderosos, mas também alinhados com princípios éticos e seguros para uso.

O ano de 2023, apelidado de "o ano da IA"<sup>33</sup>, se destaca como um marco na história dos LLMs, caracterizado por maior acessibilidade e contribuições globais. As inovações durante esse ano demonstraram que os LLMs podem ser criados com o mínimo de código, reduzindo significativamente as barreiras de entrada, ao mesmo tempo em que introduzem novos desafios, como o custo do treinamento e da inferência, e seus riscos

Figura 4. Alguns dos principais LLM e seus fornecedores<sup>34</sup>.

Empresa	LLM	Comentários	País
OpenAI	ChatGPT	Conhecido por sua versatilidade em tarefas linguísticas, popular para preenchimento de texto, tradução e muito mais.	Estados Unidos
Microsoft	Orca	Concentra-se na criação de dados sintéticos e em recursos de raciocínio aprimorados.	Estados Unidos
Anthropic	Claude	Reconhecido por seu amplo conhecimento geral e recursos multilíngues.	Estados Unidos
Google	Gemini, Gemma, BERT	Pioneira no processamento de idiomas com modelos que suportam vários tipos de dados.	Estados Unidos
Meta AI	Llama	Conhecida pela eficiência e pelo acesso democratizado, com foco no alto desempenho com computação reduzida.	Estados Unidos
LMSYS	Vicuna	Ajustado para funcionalidades de chatbot, oferecendo uma abordagem exclusiva para interações de conversação.	Estados Unidos
Cohere	Command-nightly	Especializada em tempos de resposta rápidos e pesquisa semântica em mais de 100 idiomas.	Canadá
Mistral AI	Mistral, Mixtral	Enfatiza modelos menores, mas poderosos, operando localmente com métricas de desempenho sólidas.	Francia
Clibrain	LINCE	Adaptado para o idioma espanhol, com foco em nuances linguísticas e compreensão de qualidade.	Espanha
Technology Innovation Institute	Falcon	Fornece modelos de IA de código aberto altamente eficientes e dimensionáveis com suporte multilíngue.	Emiratos Árabes Unidos
Aleph Alpha	Luminous	Notável por sua abordagem multimodal e desempenho competitivo nas principais tarefas de IA.	Alemanha
SenseTime	SenseNova	Uma série de modelos e aplicativos de IA generativa que fazem uso da plataforma de pesquisa e desenvolvimento da AGI e integram LLMs com sistemas de computação em larga escala (SenseCore, com 5.000 petaflops).	Hong Kong



inerentes. Nesse período, também houve uma preocupação crescente com as considerações e os desafios éticos apresentados pelo desenvolvimento e uso de LLMs e, como consequência, um avanço na regulamentação da IA e da IA generativa em todo o mundo.

A proliferação de LLMs de código aberto foi um marco na democratização da tecnologia de IA. Começando com o Llama e continuando com Vicuna, Falcon, Mistral, Gemma e outros, os LLMs de código aberto democratizaram o acesso à tecnologia de ponta de processamento de linguagem e permitiram que pesquisadores, desenvolvedores e amadores experimentassem, personalizassem e implantassem soluções de IA com um investimento inicial mínimo. A disponibilidade desses modelos promoveu uma colaboração sem precedentes na comunidade de IA, estimulando a inovação e facilitando a criação de aplicativos avançados em diversos setores.

Por fim, a integração do LLM às ferramentas de desenvolvimento de software e de escritório está transformando a eficiência e a capacidade das empresas. A Microsoft integrou o LLM em seu pacote Office com o nome Microsoft 365 Copilot, enquanto o Google fez o mesmo no Google Workspace. Ao mesmo tempo, ferramentas como o GitHub Copilot ou o StarCoder usam LLM para auxiliar os programadores, acelerando a geração de código e melhorando a qualidade do desenvolvimento de software.

## Tipologias de LLM

Os LLMs progrediram além da simples previsão de texto e se tornaram aplicativos sofisticados em vários domínios, arquiteturas e modalidades. Esta seção apresenta uma categorização dos LLMs de acordo com vários critérios.

### Por arquitetura

- ▶ **LLMs baseados em redes neurais recorrentes (RNNs):** esses modelos processam o texto sequencialmente, analisando o impacto de cada palavra sobre a próxima, e usam arquiteturas recorrentes, como memória de longo prazo (LSTM) ou unidades de passagem recorrentes (GRU), para processar dados sequenciais. Embora não sejam tão eficientes quanto os transformers para sequências longas, os RNNs são úteis para tarefas em que a compreensão da ordem das palavras é crucial, como na tradução automática. Exemplos são o ElMo (Embeddings from Language Models) e o ULMFiT (Universal Language Model Fine-tuning).
- ▶ **LLMs baseados en transformers:** essa é a arquitetura dominante para LLMs atualmente. Eles usam transformers para analisar as relações entre as palavras em uma frase. Isso permite que eles capturem estruturas gramaticais complexas e dependências de palavras com longa distância. A maioria dos LLMs, como GPT, Claude e Gemini, pertence a essa categoria.

### Por componente

- ▶ **Codificadores (Encoders):** são modelos projetados para entender (codificar) as informações de entrada. Eles transformam o texto em uma representação vetorial, capturando seu significado semântico. Os encoders são fundamentais em tarefas como a compreensão e a



classificação de textos. Um exemplo é o BERT do Google, um modelo que analisa o contexto de cada palavra em um texto para entender seu significado completo, e que não é realmente um LLM.

- ▶ **Decodificadores (Decoders):** esses modelos geram (decodificam) texto a partir de representações vetoriais. Eles são essenciais na geração de texto, como na criação de novo conteúdo a partir de prompts fornecidos. A maioria dos LLMs são decodificadores.
- ▶ **Codificadores/Decodificadores (Encoders/Decoders):** esses modelos combinam encoders e decoders para converter um tipo de informação em outro, facilitando tarefas como a tradução automática, em que o texto de entrada é codificado e depois decodificado em outro idioma. Um exemplo é o T5 (Text-to-Text Transfer Transformer) do Google, projetado para lidar com várias tarefas de processamento de linguagem natural.

#### *Por abordagem de treinamento*

- ▶ **LLM pré-treinados:** esses modelos são primeiramente treinados em um grande corpus de texto não rotulado usando técnicas de aprendizagem auto-supervisionadas, como modelagem de linguagem mascarada ou previsão da próxima frase, e podem ser ajustados com dados rotulados menores para tarefas específicas. Os exemplos incluem modelos como GPT, Mistral, BERT e RoBERTa, entre muitos outros.
- ▶ **LLM específicos:** esses modelos são treinados do zero com dados rotulados para uma tarefa específica, como análise de sentimentos, resumo de texto ou tradução automática. Os exemplos incluem modelos de tradução e resumo.

#### *Por modalidade*

- ▶ **LLM somente de texto:** são o tipo mais comum, treinados e trabalhando exclusivamente com dados textuais. Exemplos são GPT-3, Mistral ou Gemma.
- ▶ **LLM multimodais:** é um campo emergente em que os LLMs são treinados em uma combinação de texto e outros formatos de dados, como imagens ou áudio. Isso permite que eles executem tarefas que exigem a compreensão da relação entre diferentes modalidades. Exemplos são GPT-4, Claude 3 e Gemini.

#### *Por tamanho*

- ▶ **Large language models (LLM):** são modelos que usam grandes quantidades de parâmetros. Eles são muito avançados, mas exigem uma infraestrutura tecnológica relativamente cara na nuvem para sua execução. Exemplos são o GPT-4, o Gemini e o Claude 3.
- ▶ **Small language models (SLM):** uma tendência recente, os SLMs são versões menores e mais eficientes dos LLMs, projetados para serem executados em dispositivos com recursos limitados, como smartphones ou dispositivos de IoT, sem a necessidade de conexão ou implantação na nuvem. Apesar de seu tamanho pequeno, esses modelos mantêm um desempenho aceitável graças a técnicas como compressão ou quantização de modelos, o que reduz a precisão dos pesos e ativações do modelo. Exemplos são o Gemini Nano do Google e a família de modelos Phi da Microsoft.

## LLMs na prática: casos de uso em produção

Apesar do crescente interesse e da exploração de possíveis aplicações do LLM nas organizações, os casos de uso reais implementados em produção ainda são limitados. A maioria das empresas está em um estágio relativamente inicial, identificando e priorizando possíveis casos de uso.

No entanto, várias empresas já conseguiram colocar alguns casos de LLM em produção, demonstrando seu valor tangível para a empresa e seus clientes. Alguns desses casos estão resumidos aqui:

- ▶ **Chatbots internos:** várias organizações implementaram chatbots baseados em LLM para facilitar o acesso dos funcionários a políticas, procedimentos e informações relevantes da empresa. Esses assistentes de conversação permitem respostas rápidas e precisas a consultas frequentes, melhorando a eficiência e reduzindo a carga sobre outros canais de suporte interno.
- ▶ **Extração de informações:** os LLMs estão sendo usados para extraír automaticamente dados importantes de documentos grandes e complexos, como relatórios anuais ou relatórios de risco climático. Essas ferramentas são capazes de processar arquivos PDF de milhares de páginas, com estruturas heterogêneas, incluindo imagens, gráficos e tabelas, e transformar as informações relevantes em formatos estruturados e acessíveis, como tabelas ordenadas. Essa automação permite que as empresas economizem tempo e recursos em tarefas de análise de documentos.
- ▶ **Suporte ao centro de atendimento ao cliente:** alguns contact centers estão aproveitando os LLMs para melhorar a qualidade e a eficiência do serviço. Ao aplicar técnicas de transcrição e resumo, essas ferramentas geram um contexto das interações anteriores de cada cliente, permitindo que os agentes ofereçam um serviço mais personalizado. Além disso, durante as chamadas em andamento, os LLMs podem fornecer aos agentes acesso em tempo real à documentação relevante para responder a consultas específicas dos clientes, como informações sobre taxas bancárias ou instruções para bloqueio de cartões de crédito.

▶ **Classificação inteligente de documentos:** os recursos de processamento de linguagem natural dos LLMs estão sendo aplicados para classificar automaticamente grandes volumes de documentos, como contratos ou faturas, com base em seu conteúdo. Essa categorização inteligente permite que as organizações otimizem os processos de gestão de documentos e facilita a busca e a recuperação de informações relevantes.

▶ **Banco conversacional:** alguns bancos estão integrando o LLM em seus aplicativos móveis e canais digitais para oferecer experiências avançadas de conversação aos seus clientes. Esses chatbots são capazes de acessar os dados transacionais dos usuários em tempo real e responder a consultas específicas, como "Como foram meus gastos no último mês?" ou "Quanto ganhei de juros em meus depósitos no último ano?"

▶ **Assistência na elaboração de relatórios de auditoria:** as funções de auditoria interna de algumas empresas já estão usando o LLM para simplificar seus relatórios. Essas ferramentas utilizam como inssumos as conclusões do auditor, um banco de dados de relatórios anteriores e um banco de dados de regulamentos internos e externos aplicáveis. A partir dessas informações, os LLMs geram um rascunho avançado do relatório de auditoria, adotando o tom, o vocabulário e o estilo dos auditores humanos e citando adequadamente os relatórios anteriores e as regulamentações relevantes. Isso permite que os auditores economizem muito tempo em tarefas de redação e se concentrem em atividades de maior valor agregado.

Esses exemplos ilustram como os LLMs estão criando valor real em uma variedade de funções de negócios, desde a otimização de processos internos até a melhoria da experiência do cliente. Embora o número de casos de uso em produção seja atualmente limitado, espera-se que essa tendência se acelere muito rapidamente em um futuro próximo, à medida que os LLMs continuem a evoluir e os desafios relacionados à privacidade e à segurança dos dados sejam tratados de forma eficaz.



## Principais usos

Os LLMs estão encontrando aplicações em uma infinidade de domínios, transformando substancialmente a maneira como as pessoas interagem com a tecnologia e aproveitam o processamento de linguagem natural para aprimorar processos, serviços e experiências.

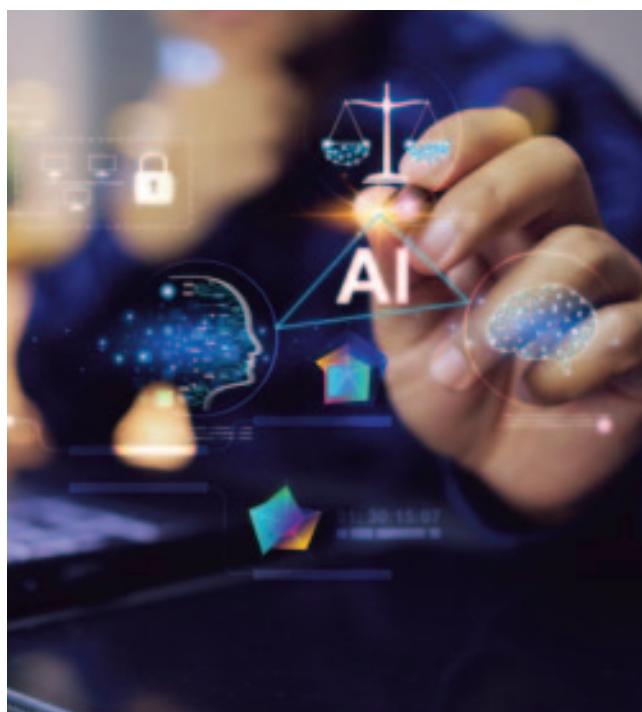
Alguns dos usos mais proeminentes dos LLMs de texto estão resumidos abaixo.

### 1. Criação e aprimoramento de conteúdo

- ▶ Geração de conteúdo: produção automática de texto.
- ▶ Assistência na redação: correção ortográfica, de estilo e de conteúdo.
- ▶ Tradução automática: conversão de texto de um idioma para outro.
- ▶ Resumo de textos: redução de documentos longos em resumos.
- ▶ Planejamento e roteiro de conteúdo: estruturação do conteúdo, p. ex., índice.
- ▶ Brainstorming: propostas criativas para projetos, nomes, conceitos, etc.
- ▶ Programação: criação de código de programação a partir de linguagem natural.

### 2. Análise e organização de informações

- ▶ Análise de sentimento: avaliação de emoções e opiniões em textos.
- ▶ Extração de informações: extração de dados específicos de documentos grandes.
- ▶ Classificação de textos: organização de textos em categorias ou temas específicos.
- ▶ Revisão técnica: assistência na revisão de documentos especializados (por exemplo, jurídicos).



### 3. Interação e automação

- ▶ Chatbots: simulação de conversas sobre tópicos gerais ou específicos.
- ▶ Perguntas e respostas: geração de respostas a perguntas com base em um corpus.

Esses usos resumem as aplicações atuais dos LLMs de texto. Com o surgimento dos LLMs multimodais, outras aplicações estão começando a surgir, como a geração de conteúdo audiovisual, a interpretação de dados de imagens, a tradução de conteúdo multimídia ou a criação de experiências interativas ricas, como a interação com chatbots com entrada não apenas de texto, mas também de imagem, áudio e vídeo.

## Requisitos regulatórios

A rápida evolução da inteligência artificial generativa, especialmente no campo da modelagem de linguagem de larga escala (LLM), chamou a atenção dos órgãos reguladores em todo o mundo. O potencial desses sistemas de influenciar negativamente os cidadãos levou ao aumento das iniciativas para estabelecer marcos regulatórios para garantir seu desenvolvimento e uso responsável.

Algumas das principais iniciativas regulatórias sobre IA incluem:

- ▶ **O AI Act da União Europeia:** uma proposta legislativa pioneira para regulamentar a IA, que classifica os sistemas de IA de acordo com seu nível de risco e estabelece requisitos de transparéncia, segurança e direitos fundamentais. O AI Act foi adotado pelo Parlamento Europeu em 13 de março de 2024.
- ▶ **O AI Bill of Rights dos EUA:** um documento de orientação que busca proteger os direitos civis no desenvolvimento e na aplicação da IA, enfatizando a privacidade, a não discriminação e a transparéncia.
- ▶ **O guia sobre IA do NIST dos EUA<sup>35</sup>:** estabelece princípios para a criação de sistemas de IA confiáveis, com foco na precisão, explicabilidade e mitigação de vieses.
- ▶ **A Declaração de Bletchley:** compromisso internacional com o desenvolvimento responsável da IA, promovendo princípios de transparéncia, segurança e imparcialidade, assinado por vários países.

<sup>35</sup>O National Institute of Standards and Technology (NIST) publicou documentos detalhando estruturas para segurança cibernética, gestão de riscos e, especificamente, gestão de modelos de IA e IA generativa.

Além das iniciativas acima, vários países começaram a emitir suas próprias regulações locais ou estabeleceram princípios para a adoção segura e ética da IA. Esses países incluem<sup>36</sup> Reino Unido, França, Espanha, Alemanha, Holanda, Polônia, Austrália, Nova Zelândia, Cingapura, Canadá, Japão, Coreia do Sul, China, Índia, Indonésia, Israel, Emirados Árabes Unidos, Arábia Saudita, Egito, Brasil, Chile, Peru, Argentina, México, Colômbia e Turquia, entre outros.

Todas essas iniciativas regulatórias têm requisitos muito semelhantes sobre IA que, quando aplicados aos LLMs, podem ser resumidos da seguinte forma:

- ▶ Transparência e explicabilidade: obrigação de divulgar como o LLM funciona, incluindo a lógica por trás de seus resultados, de modo que sejam compreensíveis para os usuários.
- ▶ Privacidade e proteção de dados: medidas rigorosas para proteger as informações pessoais coletadas ou geradas pelo LLM, em conformidade com as leis de proteção de dados, como o GDPR na Europa.
- ▶ Imparcialidade e não discriminação: requisitos para evitar vieses e garantir que os LLMs não perpetuem a discriminação e o viés, avaliando e corrigindo constantemente seus algoritmos.
- ▶ Segurança e confiabilidade: requisitos de robustez operacional para evitar mau funcionamento ou manipulações que possam causar danos ou perda de informações.
- ▶ Responsabilidade e governança: marco de responsabilidade para desenvolvedores e usuários de LLM em caso de danos ou violações de direitos, incluindo mecanismos de supervisão e controle.

- ▶ Supervisão humana: a necessidade de manter uma supervisão humana eficaz sobre os LLMs, garantindo que decisões importantes possam ser revisadas e, se necessário, corrigidas ou revertidas por humanos.

Esses requisitos refletem um consenso emergente sobre os princípios fundamentais para o desenvolvimento seguro e ético do LLM e formam a base para futuras regulamentações e adaptações específicas à medida que a tecnologia evolui.

---

<sup>36</sup>IAPP (2024).



## Desenvolvimento e implantação dos LLMs

*“A IA generativa é a chave para resolver alguns dos maiores problemas do mundo, como mudanças climáticas, pobreza e doenças. Ela tem o potencial de tornar o mundo um lugar melhor para todos”.*

Mark Zuckerberg<sup>37</sup>



Esta seção aborda os principais aspectos do processo de desenvolvimento e implantação de LLMs. Ela examina os principais componentes, como dados e arquitetura de modelos, bem como os estágios de pré-treinamento, fine-tuning e implementação. Além disso, discute os principais desafios e considerações que precisam ser levados em conta para garantir um desenvolvimento ético, robusto e alinhado com os objetivos da organização.

## **Principais aspectos do desenvolvimento dos LLMs**

O desenvolvimento de um LLM é um processo complexo que envolve vários componentes e decisões críticas. A seguir, apresentamos os principais componentes do desenvolvimento dos LLMs que precisam ser compreendidos e alguns aspectos fundamentais dos mesmos.

### **Dados**

Os dados são a base sobre a qual os LLMs são construídos, e sua qualidade, diversidade e representatividade têm um impacto direto sobre o desempenho e as tendências do modelo resultante. A abordagem dos desafios relacionados à propriedade intelectual, à qualidade dos dados e ao pré-processamento é essencial para o desenvolvimento de LLMs robustos, imparciais e precisos. À medida que as regulações e as práticas recomendadas nesse campo evoluem, é provável que vejamos maior ênfase no uso responsável e transparente dos dados no treinamento de LLM.

Alguns aspectos importantes dos dados de treinamento do LLM são:

- ▶ **Corpus de treinamento<sup>38</sup>:** os LLMs são treinados com grandes corpus de dados, geralmente extraídos da Internet, que incluem bilhões de palavras e abrangem uma ampla gama de domínios e gêneros, como livros, artigos de notícias, sites, mídias sociais e muito mais. Esses corpus enormes permitem que os LLMs aprendam padrões e representações de linguagem em grande escala, o que lhes

dá uma capacidade sem precedentes de entender e gerar textos coerentes e contextualizados. Por exemplo, corpus comuns para treinamento incluem BookCorpus<sup>39</sup>, Gutenberg<sup>40</sup>, Wikipedia<sup>41</sup> e CodeParrot<sup>42</sup>.

- ▶ Propriedade intelectual e direitos autorais<sup>43</sup>: a extração e o uso de dados da Internet para o treinamento de LLMs levanta desafios relacionados à propriedade intelectual e aos direitos autorais. Muitos desses dados são protegidos por direitos autorais, e seu uso sem permissão ou compensação adequada pode ser problemático. O AI Act na Europa aborda essa questão impondo novos requisitos aos desenvolvedores de LLM, como a obrigação de divulgar as fontes de dados usadas e a obtenção das licenças necessárias.
- ▶ Qualidade e representatividade dos dados<sup>44</sup>: como qualquer modelo, um LLM será tão bom quanto os dados usados em seu treinamento. Se os dados forem de baixa qualidade, tendenciosos ou não representativos, o modelo poderá herdar esses problemas e gerar resultados imprecisos, injustos ou inadequados. Portanto, é fundamental garantir que os corpus de treinamento sejam diversificados, equilibrados e representem adequadamente diferentes grupos demográficos<sup>45</sup>, opiniões e perspectivas.
- ▶ Iniciativas de dados de alta qualidade<sup>46</sup>: algumas iniciativas recentes se concentram na criação de LLM com menos parâmetros, mas com dados de maior qualidade, como corpus de treinamento menores, mas cuidadosamente

<sup>37</sup>Mark Zuckerberg (n. 1984), cofundador y CEO de Facebook y de Meta, una de las mayores compañías de redes sociales, tecnología e inteligencia artificial del mundo.

<sup>38</sup>Liu (2024).

<sup>39</sup>Soskek (2019).

<sup>40</sup>Project Gutenberg (2024).

<sup>41</sup>Wikipedia Dumps (2024).

<sup>42</sup>Hugging Face Datasets (2024).

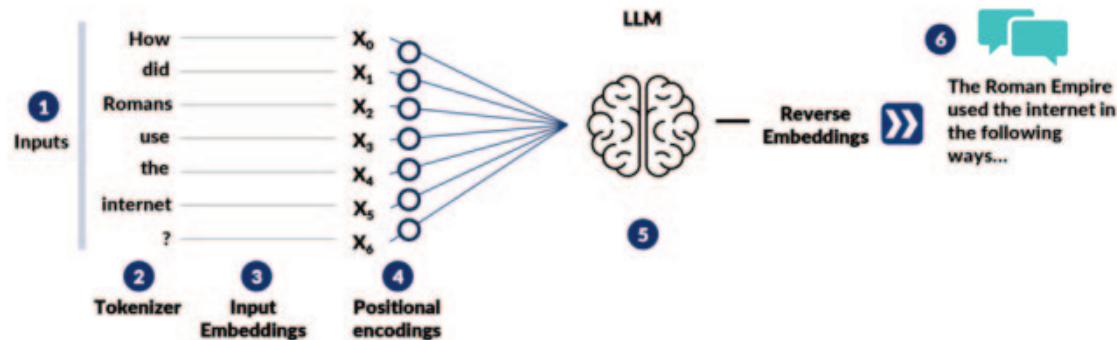
<sup>43</sup>Li (2024), Chu (2023).

<sup>44</sup>Alabdulmohsin (2024).

<sup>45</sup>Yogarajan (2023).

<sup>46</sup>Sachdeva (2024).

Figura 5. Estágios do processamento de dados de entrada do LLM..



selecionados e filtrados<sup>47</sup>, que incluem conteúdo de alta qualidade, como livros, artigos científicos e publicações respeitadas. Por exemplo, esses filtros podem ser limitados a um único idioma ou a um setor ou área temática, o que reduz drasticamente o tamanho do corpus. Essa estratégia pode resultar em LLMs com melhor desempenho e menos viés do que os modelos treinados em dados massivos não filtrados.

- ▶ Pré-processamento e rotulagem de dados<sup>48</sup>: antes de treinar ou fazer o fine-tuning de um LLM, os dados devem ser pré-processados e, em alguns casos, como o fine-tuning supervisionado ou o uso de um conjunto de dados específico, rotulados. O pré-processamento envolve a limpeza e a formatação dos dados<sup>49</sup>, a remoção de ruídos e de erros e a aplicação de técnicas como tokenização e normalização (por exemplo, LayerNorm<sup>50</sup> para Transformers).

### Tokenização e encoding

A tokenização refere-se ao processo de dividir um texto em unidades menores chamadas "tokens", que são as unidades processadas pelo LLM durante o treinamento e a inferência de resposta. Esses tokens podem ser palavras, partes de uma palavra (por exemplo, lemas) ou caracteres. Por exemplo, uma das maneiras mais simples de gerar tokens é separar o corpus de acordo com os espaços entre as palavras. O encoding é o processo de representar essas unidades de texto em formato numérico para que possam ser processadas pelo modelo.

Alguns pontos importantes sobre tokenização no LLM:

- ▶ Isso é feito com base no corpus de texto disponível, a fim de dividir o texto de origem em unidades menores de forma otimizada. O resultado final da tokenização é um encoding.

- ▶ As codificações têm um impacto significativo no desempenho do LLM<sup>51</sup>, pois definem a unidade mínima de processamento que receberão e determinam o vocabulário ao qual o LLM tem acesso.
- ▶ Há diferentes algoritmos de encoding no mercado<sup>52</sup> que diferem na forma como dividem o texto com base em palavras, frases ou sentenças, uso de espaços, capitalização ou formatação, aparência de caracteres em diferentes idiomas ou erros presentes no texto.
- ▶ As principais codificações<sup>53</sup> usadas são BytePairEncoding, SentencePieceEncoding e WordPieceEncoding.

O resultado da tokenização é usado como ponto de partida no modelo de embedding.

### Embedding

Embeddings são representações numéricas de palavras, frases, sentenças ou mesmo parágrafos que capturam seu significado semântico e as relações entre elas. Para isso, eles partem do corpus de entrada do LLM dividido em tokens. Elas são um componente fundamental dos LLMs e desempenham um papel crucial no pré-treinamento, no fine-tuning e no uso subsequente desses modelos.

#### Embeddings em LLMs:

- ▶ Eles são projetados para capturar as relações semânticas entre as palavras, de modo que as palavras com significados semelhantes tenham vetores semelhantes. Isso permite que o modelo compreenda a similaridade e as analogias entre palavras e conceitos.

<sup>47</sup>Tirumala (2023).

<sup>48</sup>Chen (2023).

<sup>49</sup>Wenzek (2019), Penedo (2023).

<sup>50</sup>Zhao (2023).

<sup>51</sup>Rejeleene (2024).

<sup>52</sup>Minaee (2024).

<sup>53</sup>Kudo (2018).

- ▶ Eles não são valores universais, mas variam entre modelos diferentes, dependendo do espaço vetorial em que são definidos.
- ▶ Eles são contextuais, o que significa que a representação de uma palavra pode variar de acordo com o contexto em que ela aparece. Isso possibilita a captura de nuances de significado e a desambiguação de palavras polissêmicas. Eles não são predefinidos, mas são aprendidos com dados de treinamento baseados no modelo de embeddings do LLM. Durante o pré-treinamento, o modelo ajusta os embeddings para maximizar sua capacidade de prever palavras no contexto (por exemplo, por meio de estruturas de embeddings, como SentenceTransformers). Entretanto, os embeddings por si só já são um modelo que precisa ser ajustado durante o processo.

### Pré-treinamento

O pré-treinamento é um estágio fundamental no desenvolvimento do LLM, durante o qual os modelos adquirem conhecimento geral e profundo do idioma a partir de grandes quantidades de dados não rotulados. Embora esse processo seja computacionalmente intensivo e caro, ele permite que o modelo seja adaptado a uma ampla gama de tarefas.

O principal objetivo do pré-treinamento é que o modelo adquira um conhecimento amplo e profundo do idioma, incluindo sua estrutura, semântica, sintaxe e contexto. Durante esse processo, o LLM aprende a prever palavras ou fragmentos de texto (ou seja, tokens) com base no contexto circundante, o que lhe permite capturar relações e padrões linguísticos complexos. Esse conhecimento geral torna-se a base sobre a qual o modelo pode ser adaptado para tarefas específicas por meio do fine tuning.

Há várias técnicas populares para o pré-treinamento do LLM, como:

- ▶ Modelagem autorregressiva da linguagem ou modelagem unidirecional (por exemplo, modelagem autorregressiva<sup>54)</sup>, que consiste em treinar o modelo para prever a próxima palavra ou fragmento de texto com base no contexto anterior. Essa tarefa permite que o modelo aprenda as probabilidades condicionais do idioma e gere um texto coerente. Exemplos são os modelos GPT e Claude.

## Tipos de embeddings

Os embeddings são usados em LLMs para estabelecer uma métrica que define a semelhança entre os significados das palavras e para incorporar informações sobre a posição das palavras em uma frase. Isso é fundamental, pois a ordem das palavras afeta o significado. Há três tipos principais de embeddings posicionais:

- ▶ *Embedding* posicional absoluto<sup>1</sup>: atribui a cada palavra – ou a cada unidade mínima de texto ou token – um vetor que representa sua posição exata na frase (por exemplo, primeira, segunda, terceira posição etc.).
- ▶ *Embedding* posicional relativa<sup>2</sup>: em vez de se basear em posições absolutas, ela representa a posição de uma palavra em relação a outras palavras (por exemplo, duas palavras antes, uma palavra depois etc.).
- ▶ *Embedding* posicional rotatório<sup>3</sup>: combina informações de posição absolutas e relativas, usando funções trigonométricas para criar representações vetoriais mais complexas.

Em um *transformer*, um *embedding* posicional simples para uma palavra em uma determinada posição pode ser representado matematicamente usando funções de seno e cosseno. Especificamente, um embedding posicional  $E$  para um token  $i$  com posição  $P$  pode ser representado matematicamente em sua forma mais simples como:

$$E(P, 2i) = \sin \frac{P}{2i} \\ 10000^{\frac{2i}{d}}$$

$$E(P, 2i + 1) = \cos \frac{P}{2i} \\ 10000^{\frac{2i}{d}}$$

onde  $P$  é a posição do token na sequência de entrada e  $d$  é a dimensão da camada oculta do *transformer*.

A escolha do tipo de *embedding* posicional pode afetar o desempenho do LLM, pois determina a quantidade e o tipo de informações posicionais disponíveis para o modelo durante o treinamento.

<sup>54</sup>Devlin (2018), Liu (2022).

<sup>1</sup>Vaswani (2017).

<sup>2</sup>Shaw (2018).

<sup>3</sup>Su (2021).

- ▶ O modelo não autorregressivo<sup>55</sup>, usado em modelos como o Gemini, no qual a resposta não é obtida sequencialmente palavra por palavra, mas é transformada e refinada como um todo.
- ▶ Modelagem de linguagem mascarada<sup>56</sup>, popularizada por modelos como o BERT, que consiste em mascarar aleatoriamente algumas palavras no texto de entrada e treinar o modelo para prever essas palavras mascaradas com base no contexto ao redor. Essa técnica permite o aprendizado bidirecional e uma melhor compreensão do contexto. Algumas arquiteturas de LLM (por exemplo, transformers bidirecionais) usam essa técnica.
- ▶ Modelagem de sequência para sequência<sup>57</sup> (por exemplo, seq2seq<sup>58</sup>), em que o modelo é treinado para gerar sequências de texto com base em outras sequências de entrada. Ele é usado em modelos como T5, BART ou ProphetNET.
- ▶ Pré-treinamento contrastivo<sup>59</sup>, usado em modelos como CLIP e ALIGN<sup>60</sup>, envolve o treinamento do modelo para identificar pares de texto-imagem semanticamente relacionados, o que permite que ele aprenda representações multimodais e transfira conhecimento entre diferentes modalidades<sup>61</sup>.

O pré-treinamento de um LLM é um processo computacionalmente intensivo que exige enormes quantidades de dados, tempo e recursos de hardware. Os maiores modelos podem ter cerca de 1 trilhão ( $10^{12}$ ) de parâmetros e exigem milhares de GPUs de última geração para semanas ou meses de treinamento. Isso torna o pré-treinamento extremamente caro e acessível apenas para algumas empresas e organizações no mundo com os recursos necessários.

### Quantificação

Durante o treinamento do LLM, os pesos dos neurônios são ajustados para fazer previsões mais precisas. Em geral, esses pesos são armazenados como números de alta precisão, o que pode resultar em modelos grandes e caros do ponto de vista computacional.

A quantização pós-treinamento é uma técnica<sup>62</sup> que permite que a precisão dos parâmetros do modelo seja reduzida sem afetar significativamente o desempenho do modelo. Por exemplo, as redes neurais que armazenam seus parâmetros usando números de ponto flutuante de 32 bits podem passar a usar apenas 16 bits ou 8 bits, dependendo do tipo de quantização. Isso resulta em modelos menores e mais rápidos, pois exigem menos memória e podem executar operações de forma mais eficiente com o hardware adequado.

Recentemente, houve uma tendência de desenvolver modelos de linguagem de pequena escala (SLMs), ou mesmo os chamados "tiny LLM"<sup>63</sup>, modelos que mantêm alto desempenho apesar de seu tamanho muito menor. Esses modelos compactos são obtidos por meio de uma combinação de técnicas, incluindo a quantização pós-treinamento.

Com a aplicação hábil dessas técnicas, os SLMs e os tiny LLMs estão, em alguns casos, alcançando desempenho comparável ao de modelos muito maiores<sup>64</sup>, o que os torna atraentes para aplicativos com restrições de recursos computacionais ou de memória.

<sup>55</sup>Xu (2021).

<sup>56</sup>Devlin (2019), Sinha (2021).

<sup>57</sup>Lee (2022).

<sup>58</sup>Sutskever (2014).

<sup>59</sup>Zeng (2023).

<sup>60</sup>Jia (2021).

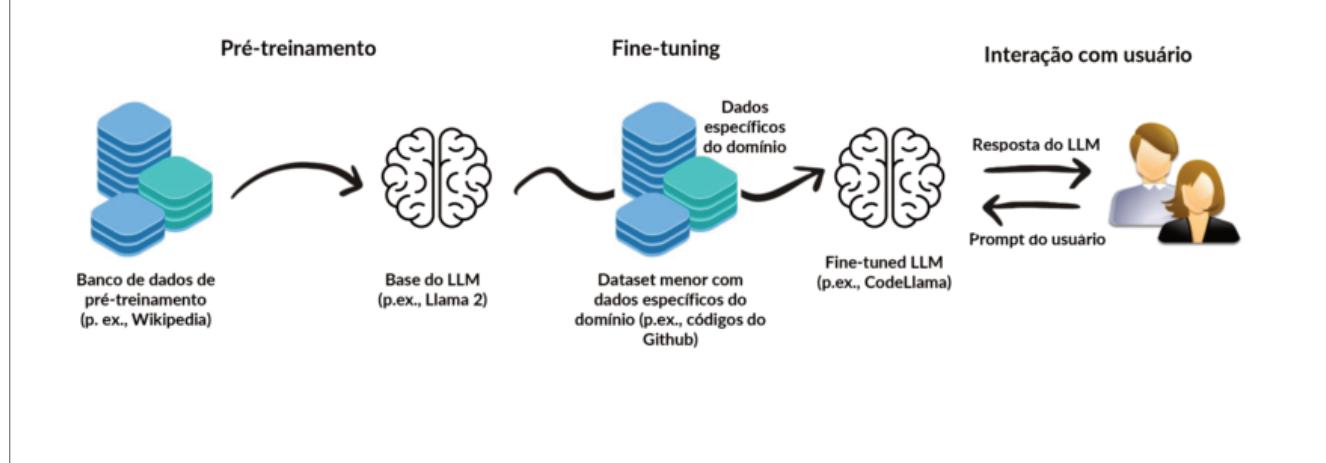
<sup>61</sup>Cui (2022).

<sup>62</sup>Li (2024).

<sup>63</sup>Tian (2024).

<sup>64</sup>Fu (2024).

Figura 6. Fine-tuning de un LLM.



## Fine-tuning, instruction-tuning y RAG

O fine-tuning é o processo de adaptação de um LLM pré-treinado a uma tarefa específica usando um conjunto de dados menor. Essa técnica permite aproveitar o conhecimento geral adquirido durante o pré-treinamento e especializá-lo para obter alto desempenho na tarefa-alvo.

O principal objetivo do fine-tuning (Fig. XX) é adaptar um LLM pré-treinado a uma tarefa específica, como classificação de sentimentos, resposta a perguntas, tradução automática ou geração de resumos. Durante esse processo, o modelo aprende a usar seu conhecimento geral de linguagem e a aplicá-lo de forma eficaz ao domínio específico e aos requisitos da tarefa em questão. Os LLMs disponíveis comercialmente, sejam eles proprietários ou de código aberto, geralmente são pré-treinados (e, portanto, de uso geral), mas não receberam fine-tuning, o que os adaptaria a uma finalidade específica.

O fine-tuning oferece vários benefícios significativos:

- ▶ **Aproveita o conhecimento prévio:** ao começar com um modelo pré-treinado, o fine-tuning permite aproveitar o vasto conhecimento geral do idioma adquirido durante o pré-treinamento, o que acelera o aprendizado e melhora o desempenho específico da tarefa.
- ▶ **Requer menos dados e recursos:** em comparação com o treinamento do zero, o fine-tuning requer muito menos dados rotulados e recursos computacionais, tornando-o mais acessível e econômico para uma ampla gama de organizações e aplicativos.
- ▶ **Permite a especialização:** o fine-tuning permite que os LLMs sejam adaptados a domínios e tarefas específicos, resultando em modelos altamente especializados e eficazes para aplicações específicas.
- ▶ **Facilita a transferência de aprendizado:** modelos fine-tuned podem receber fine-tuning adicionais para tarefas relacionadas, permitindo a transferência de aprendizado e a criação de modelos ainda mais especializados com relativamente poucos dados adicionais.

Apesar de seus benefícios, o fine-tuning também apresenta alguns desafios:

- ▶ **Super-especialização<sup>65</sup>:** se o modelo for fine-tuned em um conjunto de dados muito específico, ele poderá perder parte de sua generalização e ter um desempenho ruim com dados desconhecidos ou ligeiramente diferentes.

## Treinando um LLM: funções de perda

Os LLMs, como outros modelos de aprendizagem profunda, aprendem ajustando seus parâmetros para minimizar uma função de perda. Essa função mede a diferença entre as previsões do modelo e os resultados esperados, orientando o modelo para um melhor desempenho.

A escolha da função de perda depende do tipo de tarefa para a qual o LLM está sendo treinado. Por exemplo, para um modelo que prevê a próxima palavra em uma frase (modelagem de linguagem autorregressiva), uma função comum é a entropia cruzada. Essa função compara a distribuição de probabilidade das palavras previstas pelo modelo com a distribuição real observada nos dados de treinamento.

Matematicamente, a função de perda de entropia cruzada para um modelo autorregressivo pode ser expressa como uma soma dos logaritmos negativos das probabilidades atribuídas às palavras corretas em cada posição da sequência.

Especificamente, dada uma função de perda, como a entropia cruzada, e uma tipologia de treinamento, como a modelagem de linguagem autorregressiva, é possível definir a função de perda a ser minimizada como:

$$f_L(\varphi) = \sum_{i=1}^N -\log P(x_i | x_1, 1 \dots N, \varphi)$$

em que  $\varphi$  representa os parâmetros do modelo,  $i$  refere-se ao número de tokens em uma determinada sequência com  $N$  tokens,  $P$  é a probabilidade de prever o token  $i$  como uma função da sequência  $x$  de tokens anteriores.

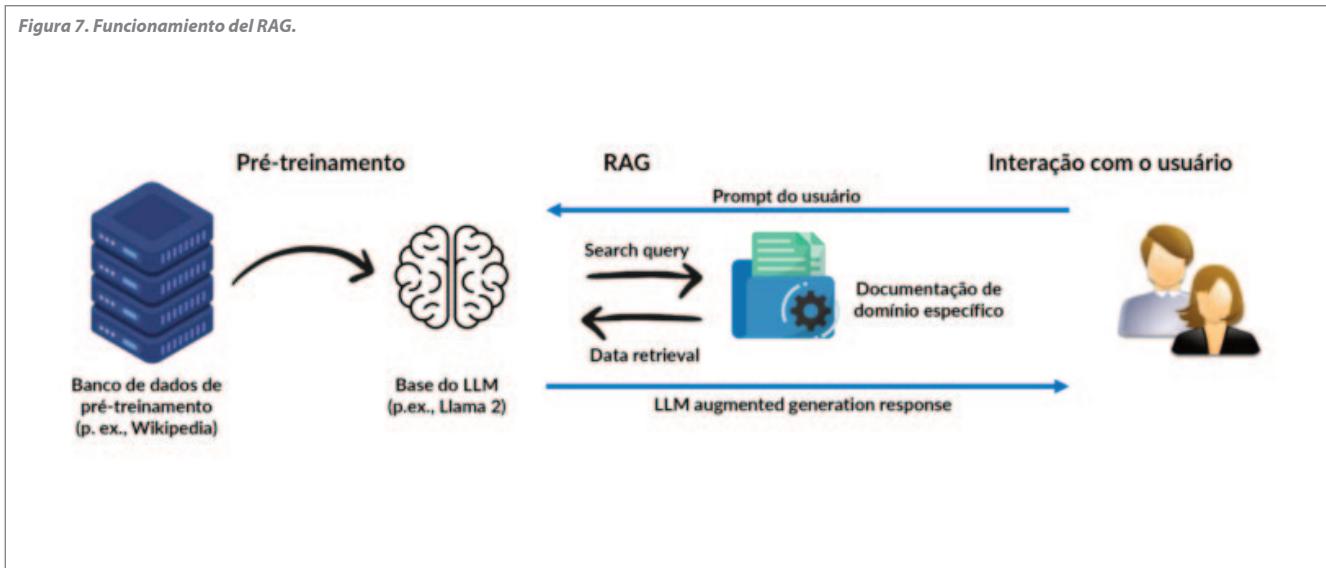
Durante o fine-tuning dos embeddings do modelo, funções de perda especializadas podem ser usadas para ajustar as representações vetoriais das palavras. As opções mais populares incluem:

- ▶ Perda de similaridade de cosseno: ajusta as incorporações para que palavras semelhantes tenham vetores mais semelhantes.
- ▶ Perda de erro quadrático médio: minimiza a diferença quadrática entre os embeddings previstos e esperados.
- ▶ Perda de classificação negativa múltipla: associa os embeddings de palavras relacionadas de modo que elas fiquem mais próximas do que as de palavras não relacionadas.
- ▶ Tripleto, Matryoshka ou perda contrastiva: variantes mais avançadas que consideram as relações entre trios ou grupos de embeddings.

A seleção cuidadosa da função de perda é fundamental para treinar LLMs eficazes e eficientes que possam capturar as nuances da linguagem natural.

<sup>65</sup>Wang (2024).

Figura 7. Funcionamento do RAG.



- ▶ **Esquecimento catastrófico<sup>66</sup>**: durante o fine-tuning, é possível que um modelo esqueça o conhecimento crítico aprendido anteriormente.
- ▶ **Instabilidade<sup>67</sup>**: o processo de fine-tuning pode ser sensível a fatores como inicialização de pesos, hiperparâmetros e seleção de dados, o que pode levar a resultados inconsistentes ou variações no desempenho.
- ▶ **Herança de vieses<sup>68</sup>**: os modelos que receberam fine-tuning podem herdar e ampliar os vieses presentes nos dados de pré-treinamento e de fine-tuning, o que exige consideração e mitigação cuidadosas.

Há vários tipos de fine-tuning a serem selecionados, dependendo de quanto o modelo inicial precisa ser modificado para se adequar a uma tarefa em um domínio mais específico. Os principais métodos são:

- ▶ **Fine-tuning supervisionado<sup>69</sup>**: requer conjuntos de dados de entrada e resposta rotulados do LLM, que são usados para melhorar sua resposta a tarefas específicas. Um método popular de fine-tuning supervisionado é o chamado instruction-tuning<sup>70</sup>, que consiste em alinhar as respostas do modelo às expectativas de seus usuários por meio de interações com o modelo.
- ▶ **Aprendizagem por reforço**: métodos baseados na aprendizagem por reforço que se concentram em melhorar a qualidade da resposta do LLM, neste caso, com base no feedback do usuário ou em modelos de recompensa (por exemplo, otimização direta de preferências<sup>71</sup>).
- ▶ **Fine-tuning não supervisionado<sup>72</sup>**: esse é um método que não exige conjuntos de dados rotulados, mas se baseia no retreinamento do modelo com as mesmas metodologias usadas durante o pré-treinamento (por exemplo, prever o próximo token).

- ▶ **Parameter efficient<sup>73</sup> fine-tuning (PEFT)**: outros métodos de fine-tuning visam aumentar a eficiência e reduzir o esforço necessário para treinar novamente o modelo. Por exemplo, técnicas baseadas em LoRA<sup>74</sup> (*low-rank adaptation*) como QLoRA ou LongLoRA<sup>75</sup>, que permitem o fine-tuning do modelo sem a necessidade de modificar seus pesos e armazenam o conhecimento aprendido durante o processo de fine-tuning em parâmetros adicionais do modelo.

Em muitos casos de uso de LLM, não é necessário empregar o fine-tuning para melhorar seus recursos em um domínio específico. A retrieval-augmented generation<sup>76</sup> (RAG) é uma técnica que melhora o desempenho do LLM por meio do uso de fontes de conhecimento externas ao modelo.

As técnicas RAG (Fig. 7) funcionam por meio da busca de documentos em um banco de dados que se assemelham ou se referem ao prompt de entrada. Essa pesquisa e seus resultados são adicionados à geração da resposta do LLM para enriquecer a, fornecendo um contexto específico.

<sup>66</sup>Luo (2024).

<sup>67</sup>Zhang (2024).

<sup>68</sup>Zhang (2024).

<sup>69</sup>Ovadia (2024).

<sup>70</sup>Zhang (2023).

<sup>71</sup>Rafailov (2023).

<sup>72</sup>Zhou (2023).

<sup>73</sup>Xu (2023).

<sup>74</sup>Dettmers (2023).

<sup>75</sup>Chen (2023).

<sup>76</sup>Lewis (2020) y Neelakantan (2022).



## Implementação e uso

Depois de treinado e validado, o LLM deve ser implantado em um ambiente de produção para uso em aplicações reais. Isso envolve a integração do modelo aos sistemas e fluxos de trabalho existentes, bem como a criação de interfaces e APIs para interagir com ele.

Esse processo envolve vários aspectos importantes, incluindo integração e monitoramento.

### Integração em sistemas e fluxos de trabalho

- ▶ **Infraestrutura<sup>77</sup>:** Os LLMs geralmente são modelos grandes e de computação intensiva, exigindo uma infraestrutura robusta para sua implementação. Isso pode envolver o uso de hardware especializado, como GPUs ou TPUs, e plataformas de computação em nuvem otimizadas para executar com eficiência o processo de inferência.
- ▶ **Interfaces e APIs<sup>78</sup>:** Para facilitar o uso do LLM em aplicativos e serviços, é necessário desenvolver interfaces e APIs que permitam que outros sistemas interajam com o modelo de forma eficiente e segura. Isso pode incluir endpoints, bibliotecas de clientes em várias linguagens de programação e interfaces gráficas de usuário para usuários não técnicos.
- ▶ **Integração com outros componentes:** em muitos casos, os LLMs fazem parte de um sistema maior que inclui outros componentes, como bancos de dados, serviços de processamento de linguagem natural e aplicativos para usuários finais. A integração suave e eficiente do LLM com esses componentes é fundamental para garantir o desempenho ideal e a experiência do usuário.

### Monitoramento e manutenção

- ▶ **Monitoramento do desempenho<sup>79</sup>:** uma vez implementado, é essencial monitorar de perto o desempenho do LLM em condições reais. Isso envolve o rastreamento de métricas, como latência, taxa de transferência, precisão e uso de recursos, e a definição de limites de consumo e custos de recursos, além de alertas para detectar e tratar qualquer degradação ou anomalia.
- ▶ **Atualização e retreinamento<sup>80</sup>:** À medida que novos dados se tornam disponíveis ou áreas de melhoria são identificadas, pode ser necessário atualizar ou retreinar o LLM. Isso requer um processo bem definido para coletar e preparar novos dados, realizar o fine-tuning e implementar a versão atualizada do modelo sem interrupções de serviço.
- ▶ **Gestão de versões<sup>81</sup>:** Com atualizações e aprimoramentos contínuos, é importante manter um rigoroso controle de versões do LLM e de seus componentes associados. Isso facilita a reproduzibilidade, a depuração e a capacidade de reverter para versões anteriores, se necessário.

29

Como se pode ver, o desenvolvimento e a implantação do LLM é um processo complexo e multifacetado que exige a consideração cuidadosa de vários aspectos, desde a seleção e a preparação dos dados até a implementação e o uso responsável do modelo. Um entendimento completo dos principais componentes, como pré-treinamento, fine-tuning e embedding, bem como a conscientização dos desafios e riscos associados, é essencial para a realização de todo o potencial da LLM de forma ética, sustentável e econômica, bem como alinhada aos objetivos de cada organização.

<sup>77</sup>Wan (2024).

<sup>78</sup>Abhyankar (2024).

<sup>79</sup>Goyal (2024).

<sup>80</sup>Lester (2021).

<sup>81</sup>Banerjee (2023).

## Arquitetura do LLM

A arquitetura do LLM refere-se à estrutura e à organização das redes neurais que compõem esses modelos. A escolha da arquitetura e de seus componentes tem um impacto significativo sobre o desempenho, a eficiência e os recursos do LLM. Esta seção explorará as principais arquiteturas usadas em LLMs e suas características, vantagens e limitações.

### Transformers: o estado da arte em LLM

Os transformers, apresentados<sup>82</sup> em 2017, surpreenderam-se a arquitetura dominante para LLMs. Diferentemente das arquiteturas anteriores baseadas em redes neurais recorrentes (RNNs) ou redes neurais convolucionais (CNNs), os transformers dependem exclusivamente de mecanismos de atenção para processar e gerar sequências de texto (Fig. 8).

A arquitetura do transformer consiste em dois componentes principais: o codificador (encoder) e o decodificador (decoder), e há transformers com somente codificador, somente decodificador ou ambos os componentes. O codificador processa a sequência de entrada e gera uma representação contextual para cada token, enquanto o decodificador gera a sequência de saída a partir da representação do codificador e das previsões anteriores.

O segredo dos transformers é o mecanismo de atenção, que permite que o modelo preste atenção a diferentes partes da sequência de entrada (atenção do codificador) e a previsões anteriores (atenção do decodificador) para gerar a próxima palavra ou token. Isso permite capturar dependências de longo prazo e gerar sequências coerentes.

Os transformers também introduzem o conceito de atenção multicabeça (multi-head attention), em que vários mecanismos de atenção operam em paralelo, permitindo que o modelo capture diferentes tipos de relações e padrões nos dados.

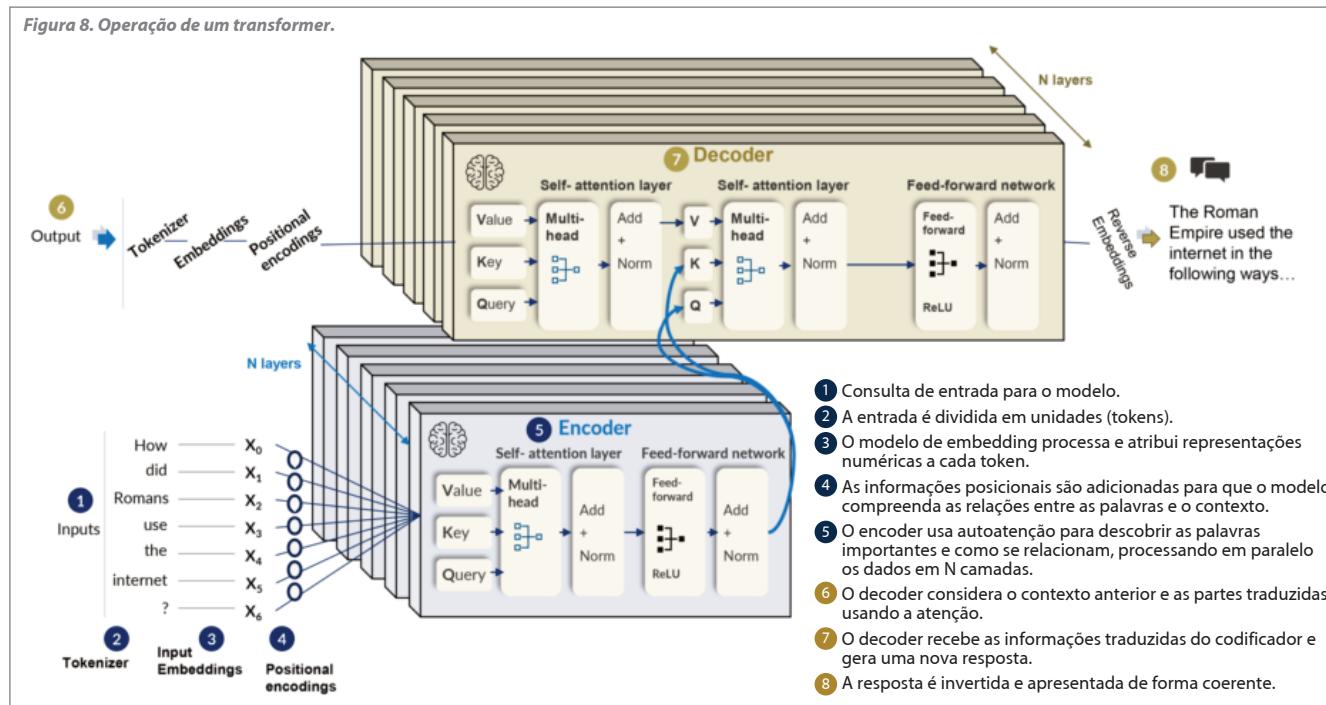
A arquitetura do transformer demonstrou excelente desempenho em uma ampla gama de tarefas de processamento de linguagem natural e foi adotada pela maioria dos LLMs de última geração.

### Variantes e extensões de transformers

Desde a introdução dos transformers, inúmeras variantes e extensões foram propostas para melhorar sua eficiência, escalabilidade e recursos de modelagem.

- ▶ Uma variante popular é o transformer bidirecional, que permite que o modelo atenda tanto ao contexto esquerdo quanto ao direito de cada token. Isso é obtido por meio do uso de um alvo de MLM (masked language modelling, modelagem de linguagem mascarada) de pré-treinamento, em que alguns tokens são mascarados aleatoriamente e o modelo deve prevê-los com base no contexto ao redor.
- ▶ Outra variante é o transformer gerativo, como o GPT, que usa uma abordagem de modelagem de linguagem unidirecional. Isso permite a geração de texto eficiente e consistente, pois o modelo só pode atender ao contexto esquerdo de cada token.

<sup>82</sup>Vaswani (2017).



## Prompt engineering em LLMs: princípios e melhores práticas

O *Prompt engineering* se refere ao processo de projetar e otimizar os prompts (entradas de texto) para obter os melhores resultados possíveis dos LLMs. Essa disciplina emergente contém vários princípios e práticas recomendadas que permitem a exploração de todos os recursos desses modelos. Eses princípios incluem:

- ▶ Seja claro e específico: as instruções dadas ao modelo devem indicar explicitamente o formato, a extensão e o nível de detalhes esperados na resposta. Por exemplo, em vez de simplesmente pedir "Analise a situação financeira da empresa X", é melhor dar uma instrução como "Escreva um relatório de 1.000 palavras sobre a situação financeira da empresa X, abrangendo sua lucratividade, liquidez, solvência e perspectivas futuras".
- ▶ Divisão de tarefas complexas: é útil dividir os problemas em subtarefas que sejam mais gerenciáveis para os LLMs. Por exemplo, ao invés de pedir "Desenvolva um plano estratégico para a empresa Y", podem ser solicitadas subtarefas como "Realize uma análise SWOT da empresa Y", "Defina os principais objetivos estratégicos para Y", "Proponha iniciativas para atingir cada objetivo", etc.
- ▶ Fornecer exemplos ilustrativos (few-shot learning): alguns exemplos bem escolhidos podem ajudar muito a comunicar a tarefa desejada. Digamos que se queira gerar propostas de valor para produtos; dois exemplos poderiam ser dados: "Nosso software de CRM permite que as equipes de vendas fechem negócios 50% mais rápido" e "Nosso aplicativo de bem-estar ajuda os funcionários a reduzir o estresse e aumentar sua produtividade em 25%".
- ▶ Peça um raciocínio passo a passo: instruir o LLM a verbalizar seu processo de pensamento geralmente leva a resultados mais robustos. Isso é especialmente útil para tarefas analíticas ou de solução de problemas em negócios. Por exemplo: "Descreva passo-a-passo como você calcularia o ROI deste projeto de investimento".
- ▶ Solicite as referências utilizadas: instrua o LLM a indicar em seu processo de argumentação as referências aos documentos que utilizou, incluindo citações do texto original ao qual ele tem acesso.
- ▶ Pedir ao LLM para adotar uma persona: antes da tarefa principal, você pode primeiro instruir o modelo a adotar uma determinada função, tom ou estilo. Por exemplo: "Aja como um analista financeiro especializado e forneça uma avaliação objetiva da empresa X". Isso ajuda a orientar seu comportamento.

- ▶ Aproveitar o conhecimento externo: o fornecimento de informações adicionais permite complementar a base de conhecimento do LLM. Por exemplo, para responder a perguntas sobre um setor específico, os relatórios setoriais relevantes podem ser recuperados e inseridos no modelo.
- ▶ Iterar e refinar sistematicamente: a avaliação contínua do desempenho do modelo permite identificar áreas de melhoria e ajustar os prompts de acordo. Métricas quantitativas e julgamentos qualitativos de especialistas no domínio podem orientar esse processo iterativo.

Ao aplicar esses princípios de engenharia imediata, é estatisticamente comprovado que os LLMs fornecem um resultado mais preciso e confiável.

Considerando tudo isso, uma sugestão ruim para um LLM escrever uma coluna sobre engenharia imediata seria: "Escreva um artigo sobre engenharia imediata".

E uma boa sugestão para escrever essa coluna seria:

"Aja como um especialista em inteligência artificial e escreva uma coluna de divulgação de 600 palavras sobre os princípios fundamentais do prompt engineering para obtenção dos melhores resultados dos LLMs. Estruture a coluna com uma introdução curta e envolvente, de 4 a 5 parágrafos que abordem os pontos principais (ser específico, dividir as tarefas, dar exemplos...) e uma conclusão com os benefícios da aplicação destas técnicas. Use um tom informativo, mas rigoroso, adequado para um público de negócios. Inclua exemplos concretos para ilustrar as ideias".

Fontes: Guia de prompt engineering da OpenAI<sup>1</sup>, suporte do Anthropic Claude Opus e elaboração própria.

- ▶ Também foram propostas extensões para tornar os transformers mais eficientes e dimensionáveis, como o transformer esparso, que usa atenção esparsa para reduzir a complexidade computacional, e o transformer comprimido, que usa técnicas de compressão para reduzir o tamanho do modelo.

### Comparação com arquiteturas anteriores

Antes dos transformers, as arquiteturas dominantes para a modelagem de sequências eram as redes neurais recorrentes (RNN), como a Long Short-Term Memory (LSTM) e a Gated Recurrent Unit (GRU), e as redes neurais convolucionais (CNN).

- ▶ As RNNs podem capturar dependências de longo prazo em sequências, mas sofrem com problemas como o desaparecimento do gradiente e a dificuldade de paralelizar o treinamento. Além disso, as RNNs têm dificuldade de capturar dependências muito longas devido à sua natureza sequencial e ao uso de recorrências de intervalo constante.
- ▶ As CNNs podem capturar padrões locais em sequências e são eficientes do ponto de vista computacional, mas têm dificuldades para modelar dependências de longo prazo e exigem um tamanho de contexto fixo.

Em comparação, os transformers superam essas limitações usando mecanismos de atenção que podem capturar dependências de longo prazo de forma eficiente e paralela. Além disso, os transformers são mais flexíveis em termos de manipulação de sequências de comprimento variável e podem ser pré-treinados em grandes quantidades de dados não rotulados.

A arquitetura do transformer revolucionou o campo do LLM e permitiu avanços significativos em uma ampla gama de tarefas

de processamento de linguagem natural. No entanto, ainda há desafios, como a escalabilidade, a interpretabilidade e a eficiência desses modelos. Com o avanço das pesquisas, é provável que surjam novas arquiteturas e técnicas que superem essas limitações e levem os LLMs a novos patamares de desempenho e capacidade.

### LLMops

MLOps (*Machine Learning Operations*) é uma metodologia e um conjunto de práticas projetadas para gerenciar o ciclo de vida completo dos modelos de aprendizado de máquina, desde o desenvolvimento e o treinamento até a implantação e a manutenção na produção.

Nos últimos anos, surgiu uma adaptação da metodologia MLOps especificamente voltada para LLMs, conhecida como LLMops (Large Language Model Operations). Essa disciplina se concentra no gerenciamento eficiente de todo o ciclo de vida do LLM, desde o desenvolvimento e o treinamento até a implementação e a manutenção em ambientes de produção.

O LLMops integra processos tradicionais de desenvolvimento de software com ferramentas e técnicas projetadas para enfrentar os desafios exclusivos apresentados por modelos de linguagem de grande escala. Alguns desses desafios incluem:

- ▶ **Gestão de grandes volumes de dados:** os LLMs exigem grandes quantidades de dados de treinamento, o que implica a necessidade de infraestruturas de armazenamento e processamento dimensionáveis e eficientes.





- ▶ **Dimensionamento de recursos computacionais:** o treinamento e a inferência de LLM exigem enormes recursos computacionais, tornando necessário o uso de técnicas de paralelização e distribuição, bem como a otimização do uso de hardware especializado, como GPUs e TPUs.
- ▶ **Monitoramento e manutenção:** uma vez implantados na produção, os LLMs devem ser monitorados de perto para detectar e corrigir problemas de desempenho, vieses, riscos, como alucinações, e degradação do modelo ao longo do tempo.
- ▶ **Controle de versões e reproduzibilidade:** considerando o tamanho e a complexidade dos LLMs, é fundamental manter um controle de versões rigoroso e maximizar a reproduzibilidade de experimentos e resultados.

Para enfrentar esses desafios, o LLMOps conta com várias ferramentas e estruturas específicas, como MLFlow<sup>83</sup>, CometML<sup>84</sup> e Weights & Biases<sup>85</sup>. Essas plataformas oferecem funcionalidades para rastreamento de experimentos, gestão de modelos, monitoramento de desempenho e colaboração em equipe.

Além disso, o LLMOps promove práticas como automação de processos, testes contínuos, documentação abrangente e governança de modelos. Isso não apenas melhora a eficiência e a qualidade do desenvolvimento dos LLMs, mas também garante seu uso ético e responsável.

## Desafios

O desenvolvimento e a implementação de LLMs apresentam uma série de desafios significativos que precisam ser enfrentados para garantir seu uso responsável, ético e seguro. Esta seção explorará vários dos principais desafios para as organizações na implantação e no uso da LLM.

### Vieses, alucinações e confiabilidade

Um dos maiores desafios para os LLMs é a presença de vieses e alucinações em seus resultados e previsões. Os vieses podem surgir de várias fontes, como vieses nos dados de treinamento, limitações das arquiteturas de modelos ou vieses humanos implícitos nas tarefas de anotação e avaliação. Por outro lado, as alucinações referem-se à geração de informações ou conteúdo que parecem plausíveis, mas não se baseiam em fatos reais ou no conhecimento adquirido durante o treinamento.

O viés no LLM pode se manifestar de várias maneiras, como a perpetuação de estereótipos de gênero, raça ou idade, a discriminação em tarefas de classificação ou a geração de conteúdo ofensivo ou inadequado. Esses vieses podem ter consequências graves, especialmente quando os LLMs são usados em aplicações sensíveis, como tomada de decisões jurídicas, financeiras ou médicas. As alucinações podem levar à disseminação de informações errôneas ou enganosas, o que pode ter um impacto negativo sobre a confiança dos usuários e a credibilidade dos aplicativos baseados em LLM.

Para enfrentar o desafio dos vieses, é necessário desenvolver técnicas robustas para detectar, medir e atenuar sua presença nos LLMs. Isso envolve a criação de conjuntos de dados de avaliação específicos para vieses, o uso de métricas de imparcialidade e a aplicação de técnicas de eliminação de vieses (debiasing) tanto no pré-treinamento quanto no fine-tuning. Além disso, é fundamental estabelecer processos contínuos de auditoria e monitoramento para garantir que os LLMs permaneçam imparciais ao longo do tempo.

<sup>83</sup>Zaharia (2018).

<sup>84</sup>CometML: <https://www.comet.com/>

<sup>85</sup>Weights and biases: <https://wandb.ai/site>



Para lidar com as alucinações em LLMs, estão sendo desenvolvidos vários métodos que se concentram no aprimoramento dos dados de treinamento, na aplicação de técnicas robustas de regularização e no uso de feedback humano para ajustar as respostas do modelo. Além disso, estão sendo investigadas alterações arquitetônicas nos modelos para torná-los inherentemente menos propensos a alucinações. Os métodos de geração de texto e o contexto de entrada também podem ser otimizados para reduzir as alucinações. O monitoramento humano e a avaliação rigorosa são essenciais para detectar e corrigir informações imprecisas. Além disso, o desenvolvimento de ferramentas específicas, como modelos de avaliação de alucinação e técnicas de ofuscação, pode contribuir para melhorar a precisão dos LLMs.

#### *Explicabilidade e responsabilidade*

Outro grande desafio dos LLMs é sua opacidade e falta de explicabilidade. Devido à sua complexidade e à natureza de suas arquiteturas, é difícil entender como esses modelos chegam aos seus resultados.

Essa falta de transparência gera problemas de responsabilidade, especialmente quando os LLMs são usados em contextos altamente sensíveis, em que as decisões têm um impacto significativo sobre os indivíduos (por exemplo, uso de LLMs na medicina, pesquisa farmacêutica, infraestrutura crítica ou acesso ao mercado de trabalho). Sem uma compreensão clara de como esses modelos funcionam, é difícil determinar a responsabilidade em caso de erros ou comportamento indesejado.

Para enfrentar esse desafio, é necessário desenvolver técnicas e ferramentas que permitam maior interpretabilidade e explicabilidade dos LLMs. Isso inclui métodos para visualizar e analisar os mecanismos internos de atenção, técnicas de atribuição para identificar as partes mais relevantes da entrada e abordagens para gerar explicações em linguagem natural das previsões do modelo.

Além disso, é importante estabelecer estruturas claras de prestação de contas que definam as responsabilidades dos desenvolvedores, implementadores e usuários de LLM, conforme proposto na Europa pelo AI Act. Isso pode envolver a criação de padrões e diretrizes para o desenvolvimento ético da LLM, mecanismos externos de monitoramento e auditoria e canais para que as partes interessadas manifestem suas preocupações.

#### *Confidencialidade e proteção das informações*

Os LLMs geralmente são treinados com grandes quantidades de dados que podem conter informações pessoais, sensíveis ou confidenciais. Além disso, quando implantados em aplicativos do mundo real, esses modelos podem ser expostos à entrada do usuário, que também pode incluir dados privados.

Isso representa desafios significativos de privacidade e segurança, pois os LLMs podem memorizar e reproduzir informações confidenciais de seus dados de treinamento ou ficar vulneráveis a ataques que tentam extraír dados privados por meio de consultas cuidadosamente elaboradas.

Para enfrentar esse desafio, é necessário desenvolver técnicas de preservação da privacidade no treinamento e na implantação do LLM (por exemplo, o Digger<sup>86</sup> para detectar informações protegidas, o uso de dados fictícios<sup>87</sup> durante o treinamento para detectar material protegido por direitos autorais).

Além disso, é fundamental estabelecer protocolos robustos de segurança e controle de acesso para proteger os LLMs e seus dados associados contra acesso não autorizado ou uso mal-intencionado. Isso pode envolver o uso de técnicas de autenticação e autorização, monitoramento de segurança e detecção de anomalias.

<sup>86</sup>Li (2024).

<sup>87</sup>Meeus (2024).

## *Uso racional de recursos*

O treinamento e a implementação do LLM exigem grandes quantidades de recursos computacionais, armazenamento e energia. Com modelos que chegam a centenas de bilhões ou até trilhões de parâmetros, o custo financeiro e ambiental do desenvolvimento e da operação desses sistemas pode ser muito significativo<sup>88</sup>.

Esse alto consumo de recursos apresenta desafios de eficiência, escalabilidade e sustentabilidade. Como a demanda por LLMs maiores e mais potentes continua a crescer, é necessário encontrar maneiras de otimizar seu desempenho e reduzir o consumo de recursos.

Para enfrentar esse desafio, várias direções de pesquisa estão sendo exploradas. Uma delas é o projeto de arquiteturas de modelos mais eficientes, como o uso de mecanismos de atenção esparsos ou técnicas de compressão que reduzem o tamanho e a complexidade computacional dos LLMs sem comprometer significativamente seu desempenho.

Também estão sendo feitas pesquisas para aprimorar o pré-treinamento contínuo<sup>89</sup> e o *fine-tuning* contínuo<sup>90</sup>, que visam integrar a capacidade de usar informações de diferentes domínios sem a necessidade de depender de um retreinamento extenso e caro com novos dados específicos. Também estão sendo feitos progressos no uso de sistemas inovadores e no projeto de algoritmos de IA ecológicos, que abordam os custos computacionais e ambientais associados à IA (por exemplo, o sistema GreenLightningAI da Qsimov Quantum Computing<sup>91</sup>, desenvolve retreinamento incremental e oferece interpretabilidade direta).

Outra direção é o desenvolvimento de infraestruturas e plataformas de computação mais sustentáveis, como o uso de hardware especializado de baixo consumo de energia, sistemas de resfriamento mais eficientes e fontes de energia renováveis para alimentar os data centers onde os LLMs são treinados e implantados.

Além disso, é importante promover práticas de uso racional e compartilhado de recursos, como a reutilização e a adaptação de modelos pré-treinados, em vez de treinar novos modelos do zero para cada tarefa, e o compartilhamento de recursos e conhecimentos entre organizações e comunidades de pesquisa.

## *Outros desafios*

Entre os muitos desafios adicionais que as organizações enfrentam no desenvolvimento, na implementação e no uso do LLM, vale a pena mencioná-los brevemente devido à sua importância:

- ▶ **Dependência e lock-in:** as organizações que dependem de LLMs fornecidos por terceiros podem enfrentar riscos de dependência e aprisionamento, especialmente se os

modelos forem baseados em dados ou infraestrutura proprietários. É importante considerar estratégias de diversificação e planos de contingência.

- ▶ **Riscos de segurança e uso malicioso**<sup>92</sup>: Os LLMs podem ser vulneráveis a ataques adversários, como injeção de dados envenenados ou engenharia reversa. Além disso, eles podem ser usados de forma maliciosa para gerar desinformação, spam ou conteúdo enganoso. É essencial estabelecer medidas de segurança robustas e projetar os modelos com salvaguardas contra o uso indevido.
- ▶ **Questões de propriedade intelectual e licenciamento:** o uso do LLM levanta questões sobre propriedade intelectual e licenciamento de dados de treinamento, modelos e resultados gerados. Além disso, há o risco de roubo de informações ou dados pessoais de usuários que lançam consultas ao LLM implantado em nuvens de terceiros. A conformidade com a regulamentação e as estruturas éticas é necessária para equilibrar os direitos dos criadores, dos usuários e do interesse público e, no caso das organizações, para evitar riscos legais e de compliance.
- ▶ **Escalabilidade da arquitetura LLM**<sup>93</sup>: Um desafio adicional é a escalabilidade dos transformers à medida que o tamanho das sequências e dos modelos aumenta. Os mecanismos de atenção têm uma complexidade quadrática com relação ao comprimento da sequência, o que limita sua aplicação a sequências muito longas.

<sup>88</sup>iDanae 1T24 (2024).

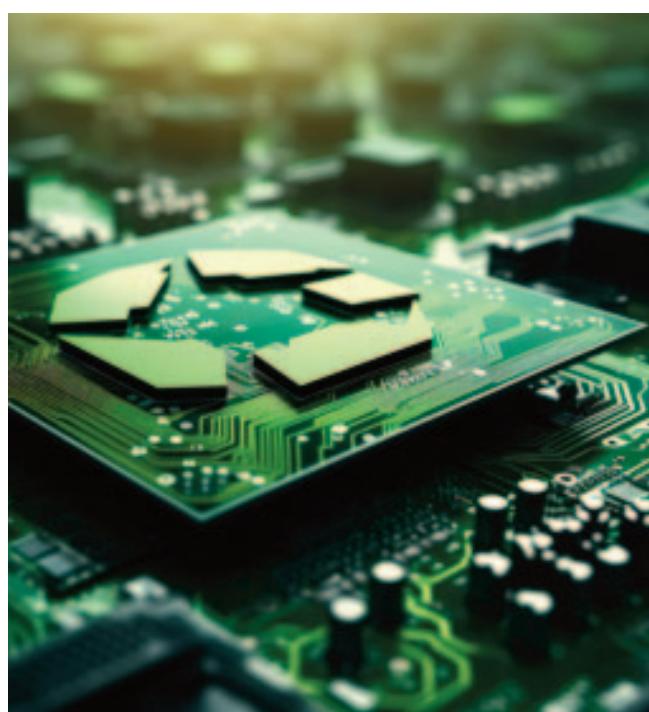
<sup>89</sup>Yıldız (2024).

<sup>90</sup>Mehta (2023).

<sup>91</sup>iDanae 1T24 (2024).

<sup>92</sup>Pankajakshan (2024).

<sup>93</sup>Rae (2021).



## Estrutura de validação de LLMs

*“As consequências de a IA dar errado são graves, portanto, precisamos ser proativos em vez de reativos”.*  
Elon Musk<sup>94</sup>



## Marco

Os modelos de linguagem de grande escala (LLMs) oferecem um grande potencial para transformar vários setores e aplicativos, mas também apresentam riscos significativos que precisam ser abordados. Esses riscos incluem a geração de desinformação ou alucinações, a perpetuação de vieses, a dificuldade de esquecer informações aprendidas, preocupações éticas e de imparcialidade, problemas de privacidade devido ao uso indevido, dificuldades de interpretação dos resultados e a possível criação de conteúdo malicioso, entre outros.

Dado o impacto potencial desses riscos, os LLMs precisam ser completamente validados antes de serem implantados em ambientes de produção. De fato, a validação dos LLMs não é apenas uma prática recomendada, mas também um requisito regulatório em muitas jurisdições. Na Europa, o AI Act proposto exige a avaliação e a mitigação de riscos dos sistemas de IA<sup>95</sup>, enquanto nos EUA, o framework de gestão de riscos de IA do NIST<sup>96</sup> e a AI Bill of Rights destacam a importância de entender e abordar os riscos inerentes a esses sistemas.

A validação dos LLMs pode se basear nos princípios estabelecidos na disciplina de risco de modelo, que se concentra<sup>97</sup> na avaliação e mitigação dos riscos decorrentes de erros, implementação inadequada ou uso indevido de modelos. Entretanto, no caso da AI e, particularmente, dos LLMs, é necessário adotar uma perspectiva mais ampla para englobar os outros riscos envolvidos. Uma abordagem abrangente da validação é essencial para garantir a implantação segura e responsável dos LLMs.

Essa abordagem holística está incorporada em uma estrutura de validação multidimensional para LLM, abrangendo aspectos fundamentais (Fig. 9), como risco de modelo, gestão de dados e privacidade, segurança cibernética, riscos legais e de compliance, riscos operacionais e tecnológicos, ética e reputação e risco de fornecedor, entre outros. Ao abordar todos esses aspectos de forma sistemática, as organizações podem

identificar e mitigar proativamente os riscos associados aos LLMs, estabelecendo a base para aproveitar seu potencial de forma segura e responsável.

Nos LLMs, essa avaliação de risco pode ser ancorada nas seguintes dimensões usadas na disciplina de risco de modelo, adaptando os testes de acordo com a natureza e o uso do LLM:

- ▶ **Dados de entrada:** compreensão de texto<sup>98</sup>, qualidade de dados<sup>99</sup>.
- ▶ **Solidez conceitual e projeto do modelo:** seleção do modelo e de seus componentes (por exemplo, metodologias de fine-tuning, conexões de banco de dados, RAG<sup>105</sup>), e comparação com outros modelos<sup>106</sup>.

<sup>99</sup>Elon Musk (nascido em 1971), CEO da X, SpaceX e Tesla. Empresário sul-africano, conhecido por fundar ou cofundar empresas como Tesla, SpaceX e PayPal, proprietário do X (antigo Twitter), uma rede social que tem seu próprio LLM, chamado Grok.

<sup>100</sup>European Parliament (2024) AI Act Art. 9: "Um sistema de gerenciamento de riscos deve ser estabelecido, implementado, documentado e mantido em relação aos sistemas de IA de alto risco. O sistema de gerenciamento de riscos [...] deverá [...] incluir [...] a estimativa e a avaliação dos riscos que podem surgir quando o sistema de IA de alto risco for usado de acordo com sua finalidade pretendida e em condições razoavelmente previsíveis de uso indevido".

<sup>101</sup>NIST (2023): "A decisão de comissionar ou implantar um sistema de IA deve ser baseada em uma avaliação contextual das características de confiabilidade e dos riscos, impactos, custos e benefícios relativos, e deve ser informada por um amplo conjunto de partes interessadas".

<sup>102</sup>Management Solutions (2014). Model Risk Management: Aspectos quantitativos e qualitativos.

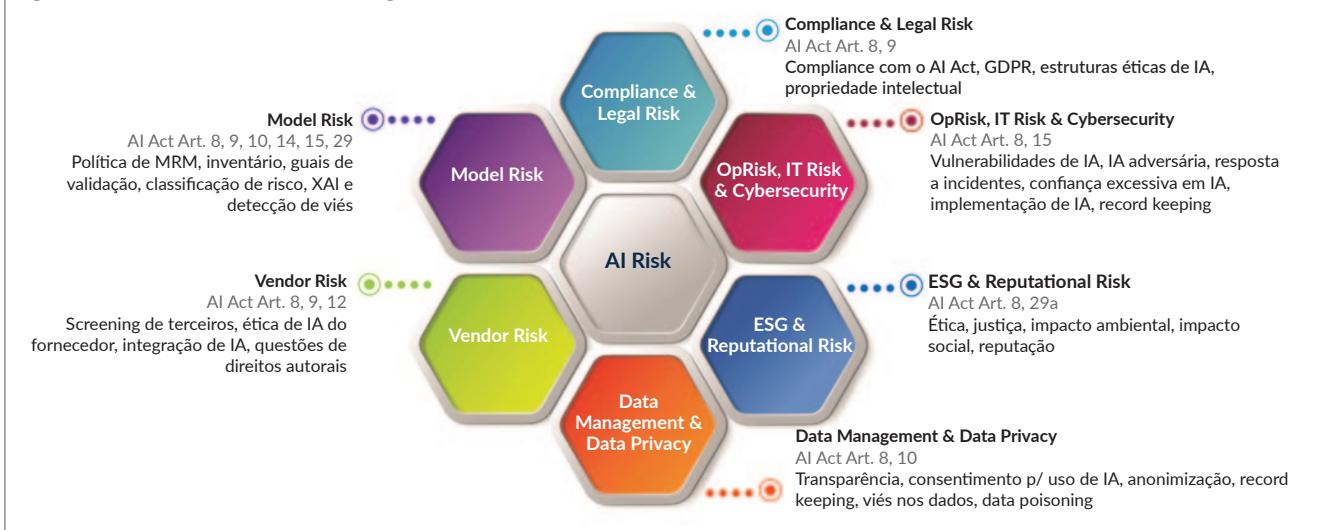
<sup>103</sup>Imperial et al. (2023).

<sup>104</sup>Wettig et al (2024).

<sup>105</sup>RAG (Retrieval-Augmented Generation) é uma técnica avançada na qual um modelo de linguagem busca informações relevantes de uma fonte externa antes de gerar o texto. Isso enriquece as respostas com conhecimento preciso e atual, combinando de forma inteligente a pesquisa de informações e a geração de texto. Ao integrar dados de fontes externas, os modelos RAG, como os modelos RAG-Token e RAG-Sequence propostos por Lewis et al. (2020), fornecem respostas mais informadas e consistentes, minimizando o risco de gerar conteúdo impreciso ou "alucinações". Esse avanço representa um passo significativo em direção a modelos de inteligência artificial mais confiáveis e baseados em evidências.

<sup>106</sup>Khang (2024).

Figura 9. Riscos associados à IA e referência regulatória na Lei de IA



- ▶ **Avaliação do modelo e análise de seus resultados:** privacidade e segurança dos resultados<sup>107</sup>, precisão do modelo<sup>108</sup>, consistência<sup>109</sup>, robustez<sup>110</sup>, adaptabilidade<sup>111</sup>, interpretabilidade (XAI)<sup>112</sup>, ética, vieses e imparcialidade<sup>113</sup>, toxicidade<sup>114</sup>, comparação com modelos desafiadores.
- ▶ **Implementação e uso:** revisão humana em uso (incluindo monitoramento de uso indevido), resolução de erros, escalabilidade e eficiência, aceitação do usuário.
- ▶ **Governança<sup>115</sup> e ética<sup>116</sup>:** estrutura de governança para IA generativa, incluindo LLM.
- ▶ **Documentação<sup>117</sup>:** integridade da documentação do modelo.
- ▶ **Compliance regulatório<sup>118</sup>:** avaliação dos requisitos regulatórios (por exemplo, AI Act).

Para garantir o uso eficaz e seguro dos modelos linguísticos, é essencial realizar uma avaliação de riscos que considere tanto o modelo em si quanto seu uso específico. Isso garante que, independentemente da origem (interna ou de um provedor) ou da personalização (fine-tuning), o modelo funcione adequadamente em seu contexto de uso, cumprindo as normas de segurança, éticas e regulamentares necessárias.

## Técnicas de validação

Quando uma organização está pensando em implementar um LLM para um caso de uso específico, pode ser benéfico adotar uma abordagem holística que englobe as principais dimensões do ciclo de vida do modelo: dados, design, avaliação, implementação e uso. Além disso, de forma transversal, é necessário avaliar a conformidade com os regulamentos aplicáveis, como o AI Act na União Europeia.

Em cada uma dessas dimensões, dois conjuntos de técnicas complementares permitem uma validação mais completa (Fig. 10):

- ▶ **Métricas de avaliação quantitativa (testes):** são testes quantitativos padronizados que medem o desempenho do modelo em tarefas específicas. São benchmarks e métricas predefinidos para avaliar diferentes aspectos do desempenho do LLM após o pré-treinamento ou durante os estágios de fine-tuning ou instruction-tuning (ou seja, técnicas de aprendizagem por reforço), otimização, engenharia de prompts ou recuperação e geração de informações. Os exemplos incluem precisão na criação de resumos, robustez a ataques adversários ou consistência da resposta a solicitações semelhantes.
- ▶ **Avaliação humana:** envolve o julgamento qualitativo de especialistas e usuários finais, por exemplo, a análise de uma amostra específica de prompts e respostas do LLM por um ser humano para identificar erros.

A validação de um uso específico de um LLM é, portanto, realizada por meio de uma combinação de técnicas quantitativas (testes) e qualitativas (avaliação humana). Para cada caso de uso específico, é necessário projetar uma abordagem de validação sob medida, que consiste em uma seleção de algumas dessas técnicas.

<sup>107</sup>Nasr (2023).

<sup>108</sup>Liang (2023).

<sup>109</sup>Elazar (2021).

<sup>110</sup>Liu (2023).

<sup>111</sup>Dun (2024).

<sup>112</sup>Singh (2024).d

<sup>113</sup>NIST (2023). Oneto (2020), Zhou (2021).

<sup>114</sup>Shaikh (2023).

<sup>115</sup>Management Solutions (2014). Model Risk Management.

<sup>116</sup>Oneto (2020).

<sup>117</sup>NIST (2023).

<sup>118</sup>European Parliament (2024). AI Act.

Figura 10. Testes de avaliação de LLMs.

Dimensões	Aspectos validados	Descrição	Métricas e abordagens de validação (exemplos)	Avaliação humana (exemplos)
1. Dados de entrada	1.1 Qualidade de dados	Grau de qualidade da modelagem ou dos dados de aplicação	<ul style="list-style-type: none"> <li>Flesch-Kinkaid Grade</li> </ul>	<ul style="list-style-type: none"> <li>Revisão caso a caso</li> </ul>
2. Desenho do modelo	2.1 Projeto do modelo	Escolha de modelos e metodologias apropriados	<ul style="list-style-type: none"> <li>Revisão dos elementos de LLM: RAG, filtros de entrada ou saída, definição de prompts, fine-tuning, otimização...</li> <li>Comparação com outros LLMs...</li> </ul>	<ul style="list-style-type: none"> <li>Testes A/B</li> </ul>
3. Avaliação do modelo	3.1 Privacidade e segurança	Respeito à confidencialidade e não regurgitação de informações pessoais	<ul style="list-style-type: none"> <li>Data leakage</li> <li>PII tests, K-anonymity</li> </ul>	<ul style="list-style-type: none"> <li>Registros</li> <li>Hacking ético</li> </ul>
	3.2 Precisão	Correção e relevância das respostas do modelo	<ul style="list-style-type: none"> <li>Q&amp;A: SummaQA, Word error rate</li> <li>Recuperação de informações: SSA, nDCG</li> <li>Resumo: ROUGE</li> <li>Tradução: BLEU, Ruby, ROUGE-L</li> <li>Outros: Sistemas de QA, nível de overrides, nível de alucinações...</li> <li>Benchmarks: XSUM, LogiQA, WikiData</li> </ul>	<ul style="list-style-type: none"> <li>Backtesting de forçagens</li> <li>Revisão caso a caso</li> </ul>
	3.3 Consistência	Respostas uniformes para consultas similares	<ul style="list-style-type: none"> <li>Cosine similarity measures</li> <li>Jaccard similarity index</li> </ul>	<ul style="list-style-type: none"> <li>Revisão caso a caso</li> <li>Testes A/B</li> </ul>
	3.4 Robustez	Resiliência a informações adversas ou enganosas	<ul style="list-style-type: none"> <li>Geração de texto adversarial (TextFooler), padrões Regex</li> <li>Benchmarks de ataques adversários (PromptBench), número de refusals</li> </ul>	<ul style="list-style-type: none"> <li>Hacking ético</li> <li>Simulações de incidentes</li> </ul>
	3.5 Adaptabilidade	Capacidade de aprender ou se adaptar a novos contextos	<ul style="list-style-type: none"> <li>Desempenho do LLM em novos dados por meio de Zero/One/Few-shot learning</li> </ul>	<ul style="list-style-type: none"> <li>Testes A/B</li> <li>Revisão caso a caso</li> </ul>
	3.6 Explicabilidade	Compreensão do processo de tomada de decisão	<ul style="list-style-type: none"> <li>SHAP Scores de explicabilidade</li> </ul>	<ul style="list-style-type: none"> <li>Hacking ético</li> <li>Focus groups</li> </ul>
	3.7 Vieses e imparcialidade	Respostas sem viés demográfico	<ul style="list-style-type: none"> <li>AI Fairness 360 toolkit</li> <li>WEAT Score, paridade demográfica, word associations...</li> <li>Benchmarks de vieses (BBQ...)</li> </ul>	<ul style="list-style-type: none"> <li>Hacking ético</li> <li>Focus groups</li> </ul>
	3.8 Toxicidade	Propensão a geração de conteúdo nocivo	<ul style="list-style-type: none"> <li>Perspective API, Hatebase API</li> <li>Toxicity benchmarks (RealToxicityPrompts, BOLD...)</li> </ul>	<ul style="list-style-type: none"> <li>Hacking ético</li> <li>Focus groups</li> </ul>
4. Implementação e uso	4.1 Revisão humana e segurança no uso	Evite sugestões prejudiciais ou ilegais e inclua uma pessoa no circuito.	<ul style="list-style-type: none"> <li>Protocolos de risco, avaliações de segurança</li> <li>Controle humano</li> </ul>	<ul style="list-style-type: none"> <li>Hacking ético</li> <li>Focus groups</li> </ul>
	4.2 Recuperação e tratamento de erros	Capacidade de se recuperar de erros e lidar com entradas inesperadas	<ul style="list-style-type: none"> <li>Testes de recuperação do sistema</li> <li>Métricas de processamento de erros</li> </ul>	<ul style="list-style-type: none"> <li>Simulações de incidentes</li> </ul>
	4.3 Escalabilidade	Manter o desempenho com mais dados ou usuários	<ul style="list-style-type: none"> <li>Stress testing do sistema, Apache Jmeter...</li> <li>Benchmarks de escalabilidade</li> </ul>	<ul style="list-style-type: none"> <li>Simulações de incidentes</li> <li>Testes A/B</li> </ul>
	4.4 Eficiência	Time-to-first-byte (TTFB), uso de GPU/CPU, inferência de emissões, memória, latência	<ul style="list-style-type: none"> <li>Time-to-first-byte (TTFB), uso de GPU/CPU, inferencia de emisiones, memoria, latencia</li> </ul>	<ul style="list-style-type: none"> <li>Simulações de incidentes</li> </ul>
	4.5 Aceitação do usuário	Teste de aceitação do usuário	<ul style="list-style-type: none"> <li>Checklist de requisitos do usuário, opt-out do usuário</li> <li>Satisfação do usuário (Net Promoter Score, CSAT)</li> </ul>	<ul style="list-style-type: none"> <li>Rastreamento de UX</li> <li>Testes A/B</li> </ul>

A seleção exata das técnicas dependerá das características específicas do caso de uso; em particular, há vários fatores importantes que devem ser levados em conta ao decidir sobre as técnicas mais adequadas:

- ▶ O nível de risco e a criticidade das tarefas a serem confiadas ao LLM.
- ▶ Se o LLM é aberto ao público (e, portanto, o hacking ético é de particular relevância) ou se seu uso é limitado ao uso interno da organização.
- ▶ Se a LLM processar dados pessoais.
- ▶ A linha de negócios ou serviço que o LLM usará.

Uma análise cuidadosa desses fatores permitirá a construção de uma estrutura de validação robusta adaptada às necessidades de cada uso de um LLM.

### *Métricas de avaliação quantitativa*

Embora seja um campo de estudo emergente, existe uma ampla gama de métricas quantitativas para avaliar o desempenho do LLM. Algumas dessas métricas são adaptações daquelas usadas em modelos tradicionais de aprendizado de máquina, como precisão, exaustividade (recall), pontuação F1 ou área sob a curva ROC (AUC-ROC). Outras métricas foram projetadas especificamente para avaliar aspectos exclusivos dos LLMs, como a coerência do texto gerado, a fidelidade factual ou a diversidade de idiomas.

Nesse sentido, já existem estruturas holísticas de teste quantitativo de LLM em ambientes de programação Python, que facilitam a implementação de muitas das métricas de validação quantitativa; por exemplo:

- ▶ **LLM Comparator**<sup>119</sup>: uma ferramenta criada por pesquisadores do Google para a avaliação e comparação automática de LLMs, que analisa a qualidade das respostas dos LLMs.
- ▶ **HELM**<sup>120</sup>: avaliação holística de modelos de linguagem, que compila métricas de avaliação ao longo de sete dimensões (precisão, calibração, robustez, imparcialidade, viés, toxicidade e eficiência) para vários cenários predefinidos.
- ▶ **ReLM**<sup>121</sup>: sistema de validação e consulta de LLM usando o uso do idioma, incluindo avaliações de modelos linguísticos, memorização, viés, toxicidade e compreensão do idioma.

No momento, algumas técnicas de validação, como os métodos de explicabilidade baseados em SHAP (XAI), algumas métricas como ROUGE<sup>122</sup> ou análises de imparcialidade usando paridade demográfica, ainda não têm limites predefinidos amplamente aceitos. Nesses casos, cabe à comunidade científica e ao setor continuar a pesquisa para estabelecer critérios claros para uma validação robusta e padronizada.

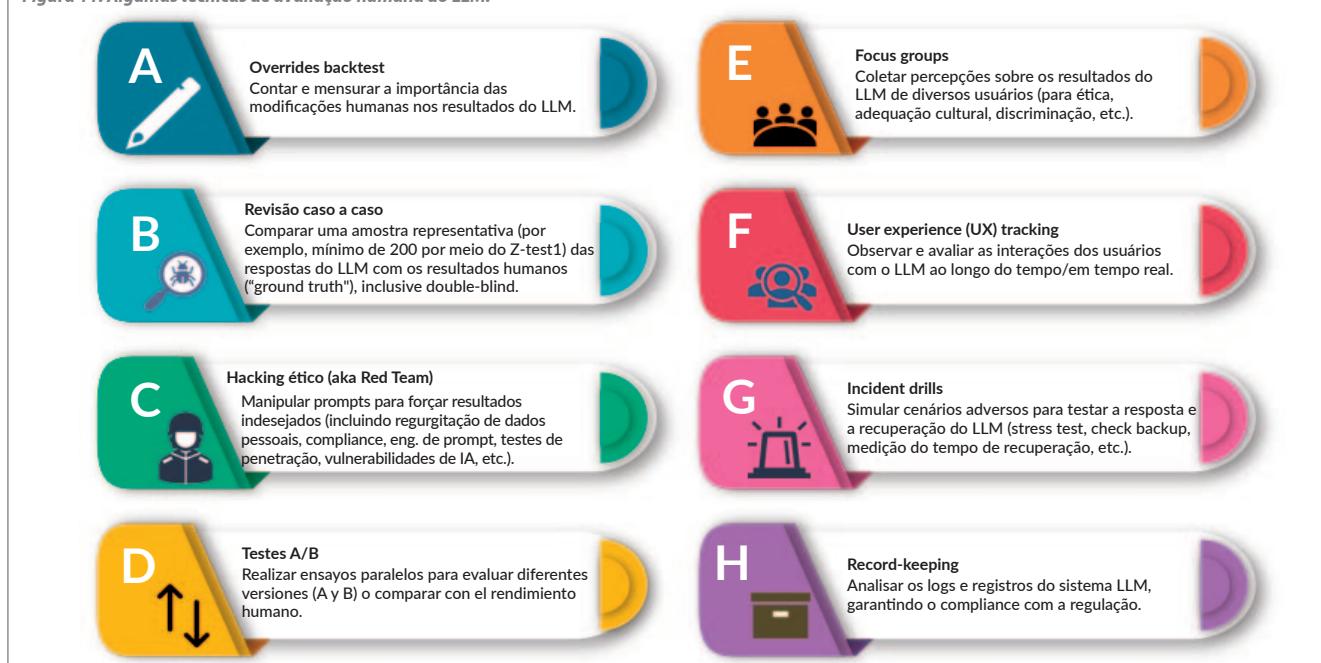
<sup>119</sup>Kahng (2024).

<sup>120</sup>Liang (2023).

<sup>121</sup>Kuchnik (2023).

<sup>122</sup>Duan (2023).

Figura 11. Algumas técnicas de avaliação humana do LLM.



## Técnicas de avaliação humana

Embora as métricas de avaliação quantitativa sejam mais diretamente implementáveis devido à grande quantidade de recursos e publicações on-line nos últimos anos, as técnicas de avaliação humana<sup>123</sup> são variadas e devem ser construídas de acordo com a tarefa específica<sup>124</sup> que está sendo executada pelo LLM, e incluem (Fig. 11):

- ▶ **Backtest do forçamento de usuários:** contabilizar e medir a importância das modificações humanas nos resultados do LLM (por exemplo, quantas vezes um gestor comercial precisa modificar manualmente os resumos de chamadas de clientes feitos por um LLM).
- ▶ **Revisão caso a caso:** comparar uma amostra representativa das respostas do LLM com as expectativas do usuário ("verdade básica").
- ▶ **Hacking ético (Red Team):** manipulação de prompts para forçar o LLM a produzir resultados indesejados (por exemplo, regurgitação de informações pessoais, conteúdo ilegal, testes de penetração, exploração de vulnerabilidades).
- ▶ **Teste A/B:** comparação para avaliar duas versões do LLM (A e B), ou um LLM contra um ser humano.
- ▶ **Focus groups:** obtenção de feedback de vários usuários sobre o comportamento do LLM, por exemplo, sobre ética, adequação cultural, discriminação, etc.
- ▶ **Experiência do usuário (UX tracking):** observar e avaliar as interações do usuário com o LLM ao longo do tempo ou em tempo real.
- ▶ **Simulações de incidentes:** simular cenários adversos para testar a resposta do LLM (por exemplo, teste de estresse, teste de backups, medição do tempo de recuperação, etc.).
- ▶ **Manutenção de registros:** analise os logs e registros do sistema do LLM, garantindo o compliance com os regulamentos e a trilha de auditoria.

## Benchmarks de avaliação do LLM

A maioria dos modelos de inteligência artificial generativa, incluindo os LLMs, é testada usando benchmarks públicos que avaliam seu desempenho em uma variedade de tarefas relacionadas à compreensão e ao uso da linguagem natural. Esses testes servem para medir como o LLM lida com tarefas específicas e refletem a compreensão humana. Alguns desses benchmarks incluem:

- ▶ GLUE/SuperGLUE: avalia a compreensão do idioma por meio de tarefas que medem a capacidade do modelo de entender o texto.
- ▶ Eleuther AI Language Model Evaluation Harness: realiza uma avaliação "few-shot" dos modelos, ou seja, sua precisão com pouquíssimos exemplos de treinamento.
- ▶ ARC (AI2 Reasoning Challenge): testa a capacidade do modelo de responder a perguntas científicas que exigem raciocínio.
- ▶ HellaSwag: avalia o senso comum do modelo por meio de tarefas que exigem a previsão do final coerente de uma história.
- ▶ MMLU (Massive Multitask Language Understanding): testa a precisão do modelo em uma ampla gama de tarefas para avaliar sua compreensão multitarefa.
- ▶ TruthfulQA: desafia o modelo a discernir entre informações verdadeiras e falsas, avaliando sua capacidade de lidar com dados verdadeiros.
- ▶ Winogrande: outra ferramenta de avaliação de senso comum, semelhante ao HellaSwag, mas com métodos e ênfase diferentes.
- ▶ GSM8K: avalia a capacidade lógico-matemática do modelo por meio de problemas matemáticos criados para os alunos.

<sup>123</sup>Datta, Dickerson (2023).

<sup>124</sup>Guzmán (2015).

## Novas tendências

O campo da validação de LLMs está em constante evolução, impulsionado pelos rápidos avanços no desenvolvimento de modelos de LLM e pela crescente conscientização da importância de garantir sua confiabilidade, imparcialidade e alinhamento com a ética e a regulamentação.

A seguir estão algumas das principais tendências emergentes nesse campo:

- ▶ **Explicabilidade dos LLMs:** à medida que os LLMs se tornam mais complexos e opacos, há uma demanda crescente por mecanismos para entender e explicar seu funcionamento interno. As técnicas de XAI (eXplainable AI), como SHAP, LIME ou a atribuição de importância a tokens de entrada, estão ganhando destaque na validação de LLMs. Embora para os modelos tradicionais haja uma variedade de técnicas post-hoc disponíveis para entender o funcionamento dos modelos em nível local e global<sup>125</sup> (por exemplo, Anchors, PDP, ICE), a definição e a implementação de modelos inherentemente interpretáveis por construção tenham proliferado, a implementação desses princípios para LLMs ainda não foi resolvida.
- ▶ **Uso de LLMs para explicar LLMs:** uma tendência emergente é usar um LLM para gerar explicações sobre o comportamento ou as respostas de outro LLM. Em outras palavras, um modelo de linguagem é usado para interpretar e comunicar de forma mais compreensível o raciocínio subjacente de outro modelo. Para enriquecer essas explicações, estão sendo desenvolvidas ferramentas<sup>126</sup> que também incorporam técnicas de análise post-hoc.

▶ **Técnicas de interpretabilidade post-hoc:** essas técnicas baseiam-se na interpretabilidade dos resultados no estágio pós-treinamento ou de fine-tuning e permitem identificar quais partes da entrada influenciaram mais a resposta do modelo (importância de características), encontrar exemplos semelhantes no conjunto de dados de treinamento (similaridade baseada em embeddings) ou projetar prompts específicos que orientem o modelo para explicações mais informativas (estratégias de prompts).

▶ **Pontuações de atribuição:** como parte da interpretabilidade post-hoc, estão sendo desenvolvidas técnicas<sup>127</sup> para identificar quais partes do texto de entrada têm maior influência na resposta gerada por um LLM. Elas ajudam a entender quais palavras ou frases são mais importantes para o modelo. Há diferentes métodos para calcular essas pontuações:

- Métodos baseados em gradiente: analisam como os gradientes (uma medida de sensibilidade) mudam para cada palavra à medida que ela se move para trás na rede neural.
- Métodos baseados em perturbações: modificam ligeiramente o texto de entrada e observam como a resposta do modelo muda.
- Interpretação de métricas internas: eles usam métricas calculadas pelo próprio modelo, como pesos de atenção em transformers, para determinar a importância de cada palavra.

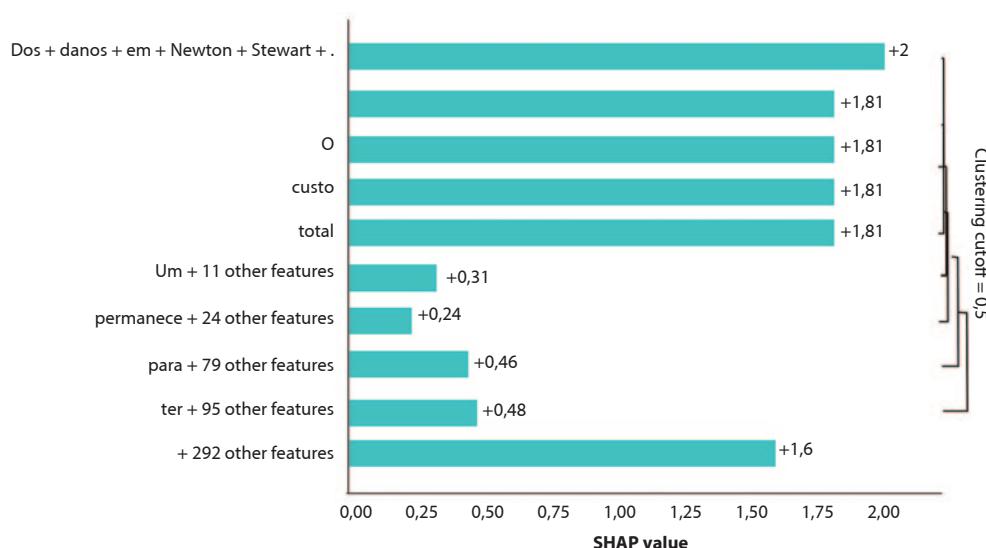
<sup>125</sup>Management Solutions (2023). Explainable Artificial Intelligence.

<sup>126</sup>Wang (2024).

<sup>127</sup>Sarti (2023).

Figura 12. Implementação de valores SHAP para resumo de texto.

Resumo da produção: "O custo total dos danos em Newton Stewart, uma das áreas mais afetadas, ainda está sendo avaliado. A Primeira Ministra Nicola Sturgeon visitou a área para inspecionar os danos. O vice-líder escocês do Partido Trabalhista, Alex Rowley, esteve em Hawick na segunda-feira para ver a situação em primeira mão. Ele disse que era importante acertar o plano de proteção contra enchentes".



Um exemplo de pontuação de atribuição é a aplicação da técnica SHAP para fornecer uma medida quantitativa da importância de cada palavra para o resultado do LLM, o que facilita sua interpretação e compreensão (Fig. 12).

- ▶ **Validação e monitoramento contínuos na produção:** além da avaliação oportuna antes da implantação, a prática de monitorar continuamente o comportamento dos LLMs quando eles estiverem em uso, como é feito com os modelos tradicionais, é bastante difundida. Isso possibilita a detecção de possíveis desvios ou degradações no desempenho ao longo do tempo, bem como a identificação de vieses ou riscos que não foram previstos inicialmente.
- ▶ **Validação colaborativa e participativa:** promove um maior envolvimento de diversas partes interessadas no processo de validação, incluindo não apenas especialistas técnicos, mas também usuários finais, órgãos reguladores, auditorias externas e representantes da sociedade civil. Essa participação pluralista permite a incorporação de diferentes perspectivas e promove a transparência e a responsabilidade.
- ▶ **Validação ética e alinhada à regulação:** além das métricas de desempenho, há um foco cada vez maior em avaliar se o comportamento do LLM é ético e alinhado aos valores humanos e à regulação. Isso envolve a análise de questões como imparcialidade, privacidade, segurança, transparência e o impacto social desses sistemas.
- ▶ **Machine unlearning:** essa é uma técnica emergente<sup>128</sup> que permite "desaprender" informações conhecidas de um LLM sem retreiná-lo novamente do zero. Isso é feito, por exemplo, adaptando os hiperparâmetros do modelo aos dados a serem desaprendidos. O mesmo princípio pode ser usado para remover as tendências identificadas. O resultado é um modelo que mantém seu conhecimento geral, mas que removeu as tendências problemáticas, melhorando sua imparcialidade e alinhamento ético de forma eficiente e seletiva. Vários métodos de machine unlearning estão sendo explorados atualmente, como o gradient ascent<sup>129</sup>, o uso de fine-tuning<sup>130</sup> ou a modificação seletiva de determinados pesos, camadas ou neurônios do modelo<sup>131</sup>.

## SHAP (SHapley Additive exPlanations) aplicado a um LLM

O SHAP é um método de explicabilidade post-hoc baseado na teoria dos jogos cooperativos. Ele atribui a cada recurso (token) um valor de importância (valor Shapley) que representa sua contribuição para a previsão do modelo.

Formalmente, seja  $x = (x_1, \dots, x_n)$  uma sequência de tokens de entrada. A previsão do modelo é denotada como  $f(x)$ . O valor de Shapley  $\varphi$  para o token  $x_i$  é definido como:

$$\varphi_i = \sum_{\{S \subseteq N_i\}} \frac{|S|!(n - |S| - 1)!}{\{n!\}} [f(S \cup \{i\}) - f(S)]$$

em que  $N$  é o conjunto de todos os tokens,  $S$  é um subconjunto de tokens e  $f(S)$  é a previsão do modelo para o subconjunto

Intuitivamente, o valor de Shapley  $\varphi_i$  captura o impacto médio do token  $x_i$  na previsão do modelo, considerando todos os subconjuntos possíveis de tokens.

Exemplo: Um LLM treinado para classificar e-mails corporativos como "importantes" ou "não importantes" é considerado. Dado o vetor de tokens de entrada:

$X = [O, \text{relatório}, \text{financeiro}, \text{do}, \text{Q2}, \text{mostra}, \text{um}, \text{aumento}, \text{significativo}, \text{na}, \text{receita}, \text{e}, \text{na}, \text{rentabilidade}]$ .

O modelo classifica a correspondência como "importante" com  $f(x) = 0,84$ .

Aplicando o SHAP, são obtidos os seguintes valores de Shapley:

$$\begin{aligned} \varphi_1 &= 0.01 (\text{O}) \\ \varphi_2 &= 0.2 (\text{relatório}) \\ \varphi_3 &= 0.15 (\text{financeiro}) \\ \varphi_4 &= 0.02 (\text{do}) \\ \varphi_5 &= 0.1 (\text{Q2}) \\ \varphi_6 &= 0.05 (\text{mostra}) \\ \varphi_7 &= 0.01 (\text{um}) \\ \varphi_8 &= 0.15 (\text{aumento}) \\ \varphi_9 &= 0.1 (\text{significativo}) \\ \varphi_{10} &= 0.01 (\text{na}) \\ \varphi_{11} &= 0.02 (\text{receita}) \\ \varphi_{12} &= 0.12 (\text{e}) \\ \varphi_{13} &= 0.01 (\text{na}) \\ \varphi_{14} &= 0.02 (\text{rentabilidade}) \end{aligned}$$

43

Interpretação: os tokens "relatório" (0,2), "financeiro" (0,15), "aumento" (0,15) e "receita" (0,12) têm as maiores contribuições para a classificação do e-mail como "importante". Isso sugere que o LLM aprendeu a associar esses termos à importância da mensagem em um contexto comercial.

<sup>128</sup> Liu (2024).

<sup>129</sup> Jang (2022).

<sup>130</sup> Yu (2023).

<sup>131</sup> Wu (2023).

# Caso prático: validação de um chatbot de políticas

*“A inteligência artificial atingirá níveis humanos até 2029”.*

*Ray Kurzweil<sup>132</sup>*

*“Acho que teremos uma IA mais inteligente do que qualquer ser humano provavelmente até o final de 2025”.*

*Perplexity<sup>133</sup>*



Para ilustrar a aplicação das técnicas de validação do LLM descritas acima, esta seção apresenta um estudo de caso da validação de um chatbot de políticas internas de uma empresa.

## Definição de caso

A empresa desenvolveu um chatbot baseado em um LLM de código aberto para responder a perguntas e fornecer informações sobre suas políticas internas. O principal objetivo desse chatbot é facilitar o acesso dos funcionários às políticas da empresa.

O chatbot foi criado usando uma infraestrutura de nuvem e foi alimentado com todas as políticas da empresa, que compreendem aproximadamente 1.000 páginas de documentação. Para melhorar sua capacidade de resposta, foram aplicadas técnicas de Retrieval-Augmented Generation (RAG), permitindo que o modelo recupere informações relevantes de sua base de conhecimento antes de gerar uma resposta. Inicialmente, foi considerado o fine-tuning do modelo, mas, após os testes iniciais, concluiu-se que a combinação do LLM básico com o RAG era suficiente para obter resultados satisfatórios.

Antes de sua implementação final, a empresa decidiu conduzir um processo de validação completo para avaliar a precisão, a segurança e a adequação do chatbot no contexto específico de seu uso pretendido. Esse processo de validação tem como objetivo identificar possíveis áreas de melhoria e garantir que o chatbot atenda aos padrões de qualidade e às expectativas da empresa.

A validação do chatbot de políticas será conduzida usando uma combinação de métricas quantitativas e técnicas de avaliação humana, seguindo a estrutura de validação multidimensional descrita na seção anterior. Os resultados desse processo serão usados para tomar decisões informadas sobre a implementação do chatbot e para estabelecer um plano de melhoria contínua.

## Desenho da abordagem de validação

Para validar de forma abrangente o chatbot de políticas, seguindo a estrutura de validação apresentada na seção anterior, foi desenhada uma abordagem de validação personalizada que abrange as principais dimensões do ciclo de vida do modelo: dados, design, avaliação, implementação e uso. Essa abordagem combina métricas quantitativas e técnicas de avaliação humana, com o objetivo de obter uma visão completa do desempenho e da adequação do chatbot no contexto específico da empresa.

Os testes e as técnicas selecionados para cada dimensão estão resumidos abaixo:

### Dados

- ▶ Métricas: a escala Flesch-Kincaid será usada para avaliar a legibilidade e a complexidade das políticas que alimentam o chatbot.
- ▶ Avaliação humana: uma amostra representativa das políticas será analisada para identificar possíveis inconsistências, erros ou ambiguidades.

### Desenho do modelo

- ▶ Métricas: elementos específicos do LLM serão modificados no código de desenvolvimento (por exemplo, a técnica RAG e seus hiperparâmetros, como o tamanho ou a estratégia de "chunking"<sup>134</sup>) que podem modificar seu desempenho de resposta, e os resultados serão comparados com o modelo original.

<sup>132</sup>Ray Kurzweil (nascido em 1948). Diretor de engenharia do Google, cientista da computação, inventor e futurista, conhecido pela invenção do OCR e por suas contribuições à IA.

<sup>133</sup>Elon Musk (nascido em 1971), CEO da X, SpaceX e Tesla. Empresário sul-africano, conhecido por fundar ou cofundar empresas como Tesla, SpaceX e PayPal, proprietário do X (antigo Twitter), uma rede social que tem seu próprio LLM, chamado Grok.

<sup>134</sup>Chunking refere-se ao processo de dividir o texto de entrada do LLM em unidades menores e mais gerenciáveis ("chunks") durante o uso ou a implementação.

- ▶ Avaliação humana: será realizada uma revisão completa dos componentes do chatbot, incluindo a configuração do RAG, filtros de entrada e saída, definição de prompts e otimização de hiperparâmetros. Além disso, serão realizados testes A/B para comparar o desempenho do chatbot com outros LLMs disponíveis no mercado.

### Avaliação do modelo

#### ▶ Privacidade e segurança

- Métricas: os testes de anonimização K serão aplicados para avaliar a proteção de dados pessoais nas respostas do chatbot, e os testes de PII (Personal Identifiable Information) serão aplicados para identificar atributos confidenciais nos dados, usando o PIIfilter.
- Avaliação humana: serão realizados testes de hacking ético para identificar possíveis vulnerabilidades e serão mantidos registros detalhados das interações do chatbot.

#### ▶ Precisão

- Métricas: as métricas Word Error Rate (WER) e ROUGE serão usadas para avaliar a precisão das respostas do chatbot em comparação com as políticas originais. Também serão usados benchmarks específicos do domínio, como um conjunto de perguntas e respostas criadas pelos especialistas em políticas da empresa.
- Avaliação humana: será realizada uma revisão caso a caso de uma amostra representativa das interações do chatbot para identificar possíveis erros ou imprecisões.

#### ▶ Consistência

- Métricas: a similaridade de cosseno e o índice de Jaccard serão aplicados para avaliar a consistência das respostas do chatbot a consultas semelhantes.
- Avaliação humana: testes A/B serão realizados para comparar as respostas do chatbot em diferentes cenários e uma revisão caso a caso será conduzida para identificar possíveis inconsistências.

#### ▶ Robustez

- Métricas: ferramentas como o TextFooler serão usadas para gerar textos adversários e avaliar a resistência do chatbot a informações enganosas. Além disso, será contado o número de rejeições de prompts maliciosos pelo chatbot.
- Avaliação humana: testes de hacking ético e incidentes simulados serão realizados para avaliar a capacidade do chatbot de lidar com situações adversas.

#### ▶ Adaptabilidade

- Métricas: o desempenho do chatbot será avaliado em relação a novas políticas ou atualizações usando técnicas de few-shot learning. Será avaliada a resposta do chatbot a idiomas não usados nas políticas ou solicitações de tradução para idiomas não incluídos no RAG (por exemplo, polonês).
- Avaliação humana: testes A/B e revisões caso a caso serão realizados para avaliar a capacidade do chatbot de se adaptar a novos cenários.

#### ▶ Explicabilidade

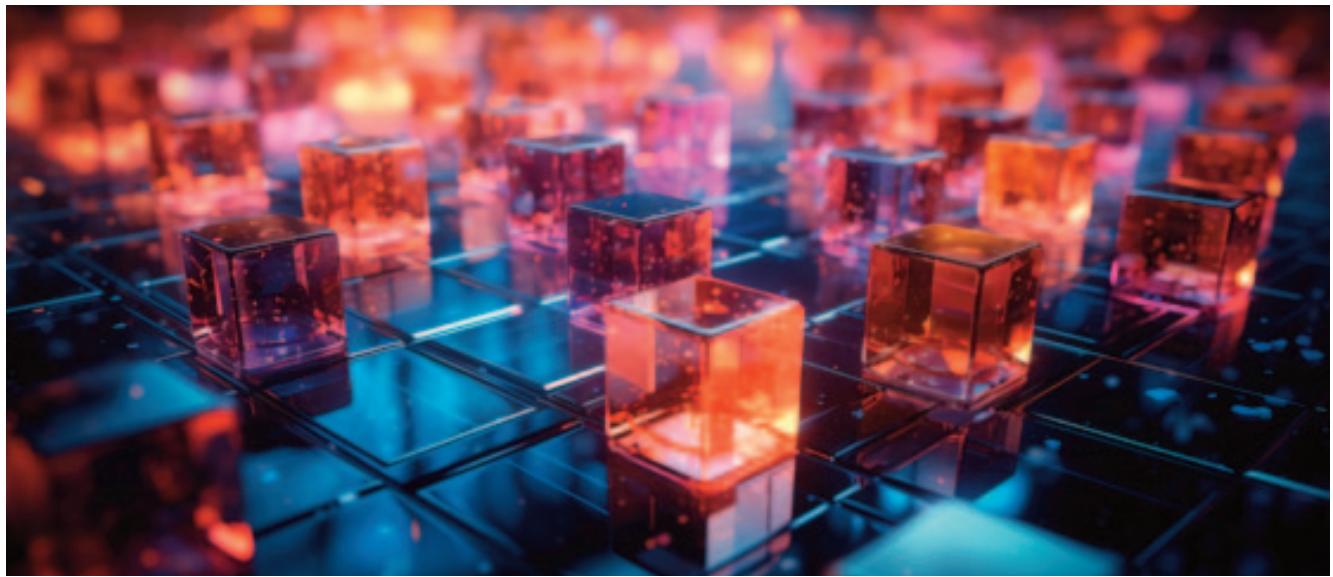
- Métricas: técnicas de explicabilidade, como o SHAP, serão aplicadas para entender o processo de tomada de decisão do chatbot. O módulo de interpretabilidade intrínseca do chatbot, que fornece uma explicação sobre a origem das informações na resposta ao usuário, será avaliado.
- Avaliação humana: o monitoramento da experiência do usuário (UX) e um focus group serão conduzidos para avaliar a percepção dos usuários sobre a transparência e a capacidade de explicação do chatbot.

#### ▶ Vieses e imparcialidade

- Métricas: o kit de ferramentas AI Fairness 360 será usado para avaliar possíveis vieses demográficos nas respostas do chatbot. Referências específicas, como a Bias Benchmark for QA (BBQ), também serão usadas para medir a imparcialidade no contexto das políticas da empresa.
- Avaliação humana: testes éticos de hacking e um focus group serão conduzidos para identificar possíveis vieses ou discriminação nas respostas do chatbot.

#### ▶ Toxicidade

- Métricas: as ferramentas da API Perspective e da API Hatebase serão aplicadas para avaliar a presença de linguagem tóxica ou inadequada nas respostas do chatbot. Além disso, benchmarks específicos, como o RealToxicityPrompts, serão usados para medir a toxicidade no contexto das políticas da empresa.
- Avaliação humana: serão realizados testes de hacking ético para identificar possíveis instâncias de linguagem ofensiva ou inadequada nas interações do chatbot.



## Implementação e uso

### ▶ Escalabilidade

- Métricas: o teste de estresse do sistema será realizado usando o Apache JMeter para avaliar o desempenho do chatbot sob altas cargas de trabalho.
- Avaliação humana: serão realizadas simulações para avaliar a capacidade do chatbot de lidar com um aumento inesperado no número de usuários ou consultas.

### ▶ Eficiência

- Métricas: o tempo de resposta (Time-to-First-Byte, TTFB), o uso de recursos (GPU/CPU, memória) e a latência serão medidos para avaliar a eficiência do chatbot.

### ▶ Aceitação do usuário

- Métricas: uma lista de verificação dos requisitos do usuário será estabelecida e a satisfação do usuário será medida usando indicadores como o Net Promoter Score (NPS) e o Customer Satisfaction Score (CSAT).
- Avaliação humana: o rastreamento da experiência do usuário (UX) será realizado para avaliar a aceitação e a satisfação do usuário com o chatbot.

Essa abordagem de validação personalizada permitirá que a empresa obtenha uma avaliação abrangente do chatbot de políticas, identificando áreas de melhoria e garantindo sua adequação ao uso pretendido. Os resultados desses testes e avaliações servirão como base para decisões informadas sobre a implementação e o refinamento contínuo do chatbot.

## Resultados

Depois de aplicar a abordagem de validação personalizada ao chatbot de políticas, foram obtidos resultados promissores, demonstrando sua adequação geral ao uso pretendido pela empresa (Fig. XX). Na maioria das dimensões avaliadas, o chatbot obteve um desempenho satisfatório, atendendo aos padrões de qualidade e às expectativas estabelecidas.

Em termos de qualidade dos dados de entrada, verificou-se que as políticas que alimentam o chatbot têm, em geral, um nível adequado de legibilidade e complexidade para os usuários entenderem. Além disso, a revisão humana não identificou inconsistências ou erros significativos no conteúdo das políticas.

O design do modelo também se mostrou apropriado para o caso de uso, com uma configuração ideal dos componentes do chatbot e desempenho superior em comparação com outros LLMs disponíveis no mercado.

Em termos de avaliação do modelo, o chatbot obteve resultados positivos na maioria das métricas e testes aplicados. Destacam-se a alta precisão das respostas, a consistência no tratamento de consultas semelhantes e a capacidade de adaptação a novos cenários. No entanto, foram identificadas algumas áreas de melhoria em aspectos como explicabilidade, detecção de vieses e resposta a perguntas muito específicas, em que é necessário um refinamento adicional do modelo. Na área de segurança cibernética, é necessária uma análise mais detalhada das vulnerabilidades específicas dos LLMs de código aberto usados para mitigar esse risco na produção.

Em termos de implementação e uso, o chatbot demonstrou boa escalabilidade e eficiência no tratamento de altas cargas de trabalho. Além disso, a satisfação do usuário foi alta, indicando uma boa aceitação da ferramenta no contexto da empresa.

Fig. 13. Resumo dos resultados de métricas e técnicas para avaliação humana do chatbot de políticas.

Dimensão	Teste	Resultado	Interpretação
<b>Dados</b>	<b>Flesch-Kincaid</b>	Legibilidade adequada (nota 8)	As políticas são compreensíveis para a maioria dos usuários
	<b>Revisão humana</b>	Não h in consistências significativas	As políticas são coerentes e não contêm erros graves
<b>Desenho do modelo</b>	<b>Modelos challenger</b>	Melhorias de parâmetros identificadas	É necessário adaptar os parâmetros do RAG ao contexto da política (ou seja, tamanho do bloco) para melhorar a captura de informações sobre perguntas muito específicas.
	<b>Revisão dos componentes</b>	Configuração ideal	O desenho do chatbot é apropriado para o caso de uso.
	<b>Testes A/B</b>	Desempenho superior ao de outros LLMs	O chatbot supera o desempenho de outros modelos disponíveis no mercado.
<b>Avaliação do modelo</b>	<b>K-anonimato</b>	Proteção adequada de dados pessoais	O chatbot não revela informações confidenciais em suas respostas
	<b>Hacking ético</b>	Vulnerabilidades menores identificadas	Ajustes necessários para fortalecer a segurança do chatbot
	<b>Word Error Rate (WER)</b>	WER < 5%	As respostas do chatbot são altamente precisas
	<b>ROUGE</b>	ROUGE-L > 0.8	As respostas do chatbot capturam adequadamente o conteúdo da política
	<b>Similaridade de cosseno / índice Jaccard</b>	Similaridade > 0.9	O chatbot fornece respostas consistentes para consultas semelhantes
	<b>TextFooler</b>	Resiliência moderada a textos contraditórios	O chatbot é moderadamente resistente a informações enganosas
	<b>Aprendizagem com poucos disparos</b>	Adaptabilidade satisfatória	O chatbot pode se adaptar a novas políticas ou atualizações com o mínimo de treinamento, mas é necessário monitorar e adicionar esses novos documentos ao RAG periodicamente.
	<b>SHAP</b>	Explicabilidade limitada	A capacidade do chatbot de explicar suas decisões precisa ser aprimorada, embora o componente RAG tenha sido desenvolvido de forma que o LLM dê uma resposta autoexplicativa.
	<b>AI Fairness 360 / BBQ</b>	Pequenos vieses demográficos identificados	O chatbot tem alguns vieses que precisam ser atenuados
<b>Implementação e uso</b>	<b>API do Perspective / API do Hatebase</b>	Baixa toxicidade (< 5%)	As respostas do chatbot raramente contêm linguagem tóxica ou inadequada.
	<b>Apache JMeter</b>	Escalabilidade satisfatória (até 1.000 usuários)	O chatbot pode lidar com altas cargas de trabalho sem degradação significativa do desempenho
	<b>TTFB / Uso de recursos / Latência</b>	Eficiência adequada (TTFB < 1s, uso moderado)	O chatbot responde rapidamente e usa os recursos de forma eficiente
	<b>NPS / CSAT</b>	Alta satisfação (NPS > 60, CSAT > 80%)	Os usuários estão muito satisfeitos com o chatbot e o recomendariam a outras pessoas.

Esses resultados indicam que o chatbot de políticas está no caminho certo para ser implementado na empresa, embora tenham sido identificadas algumas áreas específicas para melhorias adicionais. A seção a seguir abordará as principais conclusões e recomendações derivadas desse processo de validação.

## Principais conclusões

O processo de validação do chatbot de políticas mostrou que esse sistema baseado em LLM pode ser uma ferramenta valiosa para facilitar o acesso dos funcionários às informações relevantes da empresa. Os resultados obtidos nos vários testes e avaliações indicam que o chatbot atende amplamente aos requisitos de qualidade, segurança e eficiência definidos pela organização.

Entre os pontos fortes identificados, destacam-se a precisão e a consistência das respostas do chatbot, sua capacidade de se adaptar a novos cenários e sua escalabilidade para lidar com altas cargas de trabalho. Além disso, a satisfação do usuário com a ferramenta é alta, o que sugere uma boa aceitação e adoção pelos funcionários.

No entanto, o processo de validação também revelou algumas áreas de melhoria que precisam ser abordadas antes da implementação final do chatbot. Em particular, recomenda-se:

- 1. Melhorar a explicabilidade do modelo:** técnicas mais avançadas precisam ser desenvolvidas para que o chatbot possa fornecer explicações claras e comprehensíveis sobre seu processo de tomada de decisão. Isso aumentará a transparência e a confiança dos usuários na ferramenta. Embora o componente RAG tenha sido desenvolvido de forma que o LLM dê uma resposta autoexplicativa e faça referência à política correspondente, essa explicação não é totalmente clara para perguntas muito específicas.

**2. Mitigar os vieses identificados:** embora os vieses detectados sejam pequenos, é aconselhável aplicar técnicas de redução de vieses para garantir que as respostas do chatbot sejam justas e não discriminatórias. Sugere-se a revisão periódica dos vieses e a implementação de medidas corretivas quando necessário.

**3. Reforçar a segurança e a privacidade:** embora o chatbot esteja em conformidade com os padrões básicos de proteção de dados pessoais, recomenda-se realizar testes adicionais e recorrentes de hacking ético e adotar medidas de segurança mais robustas para evitar possíveis vulnerabilidades.

**4. Estabelecer um plano de monitoramento e melhoria contínua:** é essencial definir um processo de monitoramento e avaliação regulares do desempenho do chatbot para identificar oportunidades de melhoria e garantir seu funcionamento ideal a longo prazo. Esse plano deve incluir a coleta de feedback dos usuários, a atualização regular das políticas e a inclusão delas no banco de dados do chatbot, o monitoramento para melhorar os parâmetros usados no RAG e atualizá-los, além da incorporação de novas técnicas e tecnologias à medida que elas se tornarem disponíveis.

Em conclusão, o chatbot de políticas demonstrou potencial para melhorar a eficiência e a acessibilidade das informações na empresa. Com a implementação das melhorias sugeridas e o foco na melhoria contínua, esse sistema baseado em LLM pode se tornar uma ferramenta estratégica para o sucesso da organização. A recomendação final foi a de prosseguir com a implementação do chatbot, levando em conta as observações e recomendações derivadas desse processo de validação.



# Conclusão

*“Os LLMs são as únicas pessoas que podem escrever um romance, traduzi-lo para dez idiomas, e ainda assim não entender o enredo”.*

Perplexity<sup>135</sup>



Os modelos de linguagem de grande escala (LLMs) representam um avanço significativo no campo da inteligência artificial e estão revolucionando a maneira como interagimos com a tecnologia e aproveitamos o processamento de linguagem natural. Sua capacidade de processar e gerar textos coerentes e contextualizados abre uma ampla gama de aplicações em diversos setores, desde a criação de conteúdo e análise de sentimentos até a automação de tarefas e o aprimoramento da experiência do usuário.

Entretanto, o desenvolvimento e a implantação do LLM também trazem consigo uma série de desafios e considerações importantes. A presença de vieses e alucinações em seus resultados, a falta de transparência e explicabilidade de suas decisões, os desafios da privacidade e da segurança das informações e o alto consumo de recursos computacionais são alguns dos principais desafios que precisam ser abordados para garantir um uso responsável e ético desses sistemas.

Para enfrentar esses desafios, é essencial estabelecer uma estrutura robusta de governança de IA, especialmente no campo da IA generativa e do LLM. Essa estrutura deve abranger todos os principais aspectos, incluindo estratégia, apetite por riscos, governança, organização, framework de controle (políticas e procedimentos), dados, sistemas e relatórios. Somente uma abordagem abrangente e bem estruturada garantirá o desenvolvimento e o uso responsáveis dessas tecnologias.

Dentro dessa estrutura de governança, a validação desempenha um papel fundamental. Uma abordagem multidimensional é desejável, abrangendo todos os estágios do ciclo de vida do LLM, desde a qualidade dos dados de entrada e a robustez do projeto do modelo até a avaliação completa de seus resultados e a implementação e o uso adequados. Esse processo de validação deve combinar métricas quantitativas padronizadas com técnicas de avaliação humana adaptadas ao contexto específico de cada caso de uso.

Além disso, é necessário manter-se a par das últimas tendências e desenvolvimentos no campo da validação de LLM, como o desenvolvimento de técnicas de explicabilidade mais avançadas, o uso de LLM para explicar o comportamento de outros LLM, a validação e o monitoramento contínuos na produção e o alinhamento com princípios éticos e requisitos regulatórios.

O estudo de caso apresentado neste white paper ilustra como a aplicação de uma estrutura de validação personalizada pode ajudar as organizações a identificar os pontos fortes e as áreas de melhoria em seus sistemas baseados em LLM e a tomar decisões informadas sobre a implementação e a melhoria contínua.

Em resumo, os LLMs têm um potencial muito significativo para transformar a maneira como as empresas e a sociedade em geral se beneficiam da inteligência artificial. No entanto, para realizar todo o seu potencial de forma segura e responsável, é imperativo estabelecer uma estrutura robusta de governança de IA que aborde os desafios associados ao seu desenvolvimento e implantação, incluindo uma abordagem rigorosa e multidimensional de validação. Essa é a única maneira de garantir que esses sistemas sejam confiáveis, equitativos e alinhados com os valores e as metas das organizações e da sociedade.

<sup>135</sup>O Perplexity AI é um mecanismo de pesquisa e conversação baseado em LLM fundado em 2022 por Andy Konwinski, Denis Yarats, Johnny Ho e Aravind Srinivas (anteriormente OpenAI) que responde a consultas usando texto preditivo em linguagem natural.



## Glossário

**AGI (Artificial General Intelligence):** futura inteligência artificial hipotética que igualaria ou superaria a inteligência humana em qualquer domínio intelectual, sendo capaz de realizar qualquer tarefa intelectual que um ser humano possa fazer.

**Alucinações:** geração de informações ou conteúdo por um LLM que parece plausível, mas não se baseia em fatos ou conhecimentos reais adquiridos durante o treinamento, levando a imprecisões ou invenções nas respostas do modelo.

**Avaliação humana:** processo de revisão e avaliação qualitativa do comportamento e dos resultados de um sistema de IA por especialistas e usuários, que complementa as métricas quantitativas e permite a detecção de erros, vieses ou comportamentos indesejados que podem passar despercebidos em uma avaliação puramente automática.

**CNN (Convolutional Neural Network):** tipo de rede neural especializada no processamento de dados com uma topologia de grade, como imagens ou séries temporais. As CNNs usam camadas convolucionais para extrair automaticamente recursos locais e abstratos dos dados e são amplamente usadas em tarefas de visão computacional e processamento de sinais.

**Dados de treinamento:** conjunto de exemplos usados para treinar um modelo de aprendizado de máquina, incluindo as entradas (features) e, no caso do aprendizado supervisionado, os rótulos ou as respostas esperadas. A qualidade e a diversidade desses dados são fundamentais para o desempenho e a generalização do modelo.

**Efeito Eliza:** fenômeno psicológico em que os usuários tendem a atribuir capacidades cognitivas e emocionais semelhantes às humanas aos sistemas de conversação baseados em IA, apesar de esses sistemas não terem nenhum entendimento real de linguagem ou inteligência geral.

**Embeddings:** representações densas e contínuas de elementos discretos (como palavras, frases ou documentos) em um espaço vetorial de alta dimensão, em que elementos semelhantes têm representações próximas. Eles são usados em LLMs para capturar relações semânticas e sintáticas entre elementos de linguagem.

**Ética em IA:** disciplina que estuda os princípios morais, os valores e as diretrizes que devem orientar o desenvolvimento, a implantação e o uso de sistemas de inteligência artificial, com o objetivo de garantir que eles sejam benéficos, justos, transparentes e alinhados aos valores humanos.

**Explicabilidade (XAI, eXplainable AI):** propriedade de um modelo de IA que se refere à sua capacidade de fornecer explicações compreensíveis para humanos sobre seu funcionamento interno, o raciocínio por trás de suas previsões e os fatores que influenciam suas decisões.

**Few-shot learning:** capacidade de um modelo de aprendizado de máquina, especialmente LLMs, de aprender a executar uma nova tarefa com poucos exemplos (de um a algumas dezenas), aproveitando o conhecimento prévio adquirido durante o pré-treinamento em grandes quantidades de dados.

**Fine-tuning:** técnica para adaptar um modelo de linguagem pré-treinado a uma tarefa específica por meio de treinamento adicional com um conjunto de dados menor e mais especializado para essa tarefa. Ela permite aproveitar o conhecimento geral do modelo e ajustá-lo para obter alto desempenho em aplicativos específicos.

**Hacking ético:** prática de testar e desafiar um sistema de IA de forma controlada e autorizada, com o objetivo de identificar vulnerabilidades, falhas, vieses ou comportamentos indesejados e, em seguida, corrigi-los para melhorar a segurança e a robustez do sistema.



**Instruction tuning:** técnica de ajuste fino para LLM que consiste em fornecer ao modelo instruções, perguntas e exemplos de respostas esperadas, com o objetivo de alinhar seu comportamento com as expectativas e preferências dos usuários em um domínio específico.

**Inteligência artificial (IA):** campo da ciência e engenharia da computação dedicado ao desenvolvimento de sistemas capazes de realizar tarefas que normalmente exigem inteligência humana, como aprendizado, raciocínio, percepção, interação com linguagem natural e solução de problemas.

**Inteligência Artificial Generativa (GenAI):** subcampo da IA que se concentra na criação de modelos e algoritmos capazes de gerar conteúdo novo e original, como texto, imagens, vídeo, áudio, código-fonte ou projetos em 3D, aprendendo padrões e recursos de um conjunto de dados de treinamento.

**Large Language Models (LLM):** modelos de aprendizagem profunda especializados em processamento e geração de linguagem natural, treinados em grandes quantidades de texto e com um grande número de parâmetros (de milhões a bilhões), capazes de executar várias tarefas linguísticas com um alto nível de compreensão e coerência.

**LLMOps (Large Language Model Operations):** conjunto de práticas, ferramentas e processos para gerenciar de forma eficiente e dimensionável todo o ciclo de vida dos LLMs em ambientes de produção, abrangendo treinamento, implantação, monitoramento, atualização e governança desses modelos.

**Machine learning:** ramo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem que os sistemas aprendam e melhorem automaticamente por meio da experiência, sem serem explicitamente programados para isso.

**Machine unlearning:** conjunto de técnicas para remover ou "desaprender" seletivamente determinadas informações ou vieses indesejados de um modelo de aprendizado de máquina já treinado, sem a necessidade de treiná-lo novamente do zero, o que permite cumprir os requisitos de privacidade ou corrigir comportamentos indesejados.

**Métricas quantitativas:** medidas numéricas padronizadas usadas para avaliar de forma objetiva e consistente o desempenho de um modelo de IA em tarefas específicas, como exatidão, integridade, precisão ou eficiência.

**Modelo generativo:** tipo de modelo de aprendizado de máquina projetado para aprender a distribuição de probabilidade subjacente de um conjunto de dados e gerar novas amostras que sejam semelhantes aos dados de treinamento, podendo criar conteúdo novo e realista.

**Pré-treinamento:** estágio inicial do treinamento do LLM no qual um grande corpus de texto não estruturado e não rotulado é usado para que o modelo aprenda representações e padrões gerais do idioma, adquirindo um conhecimento amplo e robusto que pode ser adaptado a tarefas específicas por meio de ajuste fino.

**Privacidade diferencial:** técnica criptográfica usada para compartilhar informações agregadas sobre um conjunto de dados e, ao mesmo tempo, proteger a privacidade dos indivíduos presentes nesses dados, introduzindo ruído aleatório que dificulta a identificação de entradas individuais a partir dos resultados da análise.

**Prompt engineering:** disciplina que se concentra em projetar, otimizar e adaptar prompts (entradas de texto) para obter os melhores resultados possíveis dos LLMs em tarefas específicas, aproveitando técnicas como a inclusão de exemplos, especificação de formatos ou orientação passo a passo.

**Quantização:** técnica usada para reduzir o tamanho e acelerar a inferência de LLMs, reduzindo a precisão numérica dos pesos do modelo, passando de números de ponto flutuante para representações de menor precisão, como números inteiros ou de ponto fixo.

**Regulação da IA:** conjunto de leis, regulamentações, normas e diretrizes estabelecidas por governos e organizações para garantir que o desenvolvimento, a implantação e o uso de sistemas de inteligência artificial sejam conduzidos de forma responsável, segura, ética e alinhada com os valores e direitos fundamentais da sociedade.

**Retrieval-Augmented Generation (RAG):** técnica usada em LLMs que consiste em recuperar informações relevantes de uma base de conhecimento externa antes de gerar uma resposta, combinando assim a capacidade de acessar informações estruturadas com a geração de linguagem natural coerente e fluente.

**RNN (Recurrent Neural Network):** tipo de rede neural projetada para processar sequências de dados, como texto ou séries temporais. Diferentemente das redes neurais feedforward, as RNNs têm conexões recorrentes que lhes permitem manter o estado interno e capturar dependências temporais. Variantes como LSTM e GRU têm sido amplamente usadas em tarefas de processamento de linguagem natural antes do surgimento dos transformers.

**Segurança (AI safety):** disciplina que se concentra na identificação, prevenção e mitigação dos possíveis riscos associados ao desenvolvimento e ao uso de sistemas avançados de IA, tanto a curto quanto a longo prazo, incluindo riscos de segurança, vieses, erros, uso indevido ou consequências não intencionais.

**Testes A/B:** método experimental usado para comparar o desempenho de duas versões diferentes de um sistema de IA (A e B) ou entre um sistema de IA e uma abordagem alternativa (como um humano ou um modelo de base), a fim de determinar qual apresenta melhor desempenho de acordo com métricas predefinidas.

**Token:** unidade discreta na qual um texto é dividido para processamento por um modelo de linguagem. Os tokens podem ser palavras, subpalavras ou caracteres, e são a entrada básica para o treinamento e a inferência do LLM.

**Tokenização:** processo de conversão de texto em uma sequência de tokens. A escolha da estratégia de tokenização tem um impacto significativo sobre o desempenho e a eficiência do modelo.

**Transformers:** arquitetura de rede neural profunda que usa mecanismos de atenção para processar e gerar sequências em paralelo, em vez de sequencialmente como as RNNs. Ela permite capturar dependências contextuais e de longo prazo, sendo a arquitetura dominante para LLMs e definindo o estado da arte em várias tarefas de processamento de linguagem natural.

**Validação:** processo abrangente e multidisciplinar para avaliar um sistema de IA, especialmente o LLM, em termos de desempenho, robustez, segurança, imparcialidade, explicabilidade e alinhamento com requisitos e valores éticos e sociais, combinando métricas quantitativas e avaliação qualitativa por especialistas e usuários.

**Viés:** tendência sistemática de um modelo de aprendizado de máquina de produzir resultados que favorecem ou desfavorecem injustamente determinados grupos ou indivíduos, devido a características sensíveis, como gênero, etnia, idade ou orientação sexual, e geralmente é o resultado de vieses nos dados de treinamento ou de decisões abaixo do ideal durante o desenvolvimento do modelo.

# Referências



Abhyankar, R. et al. (2024). APIServe: Efficient API Support for Large-Language Model Inferencing.  
<https://arxiv.org/abs/2402.01869>. arXiv:2402.01869v1

Alabdulmohsin, I. et al. (2024). CLIP the Bias: How Useful is Balancing Data in Multimodal Learning?  
<https://arxiv.org/html/2403.04547v1.html>. arXiv:2403.04547v1

Banerjee, I., et al. (2023). MLOps with enhanced performance control and observability. <https://arxiv.org/abs/2302.01061>. arXiv:2302.01061v1

Bengio, Y. et al. (2003). A Neural Probabilistic Language Model. <https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf>

Bréal, M. (1883). Les lois intellectuelles du langage fragment de sémantique. Annuaire de l'Association pour l'encouragement des études grecques en France. Vol. 17 (1883), pp. 132-142. <https://www.jstor.org/stable/44253893>

Cambon, A. et al. (2023). Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity. A first update from Microsoft's research initiative on AI and Productivity.

Chen, D. et al. (2023). Data-Juicer: A One-Stop Data Processing System for Large Language Models.  
<https://arxiv.org/abs/2309.02033>. arXiv:2309.02033v3

Chen, Y. et al. (2023). LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models.  
<https://arxiv.org/abs/2309.12307>. arXiv:2309.12307v3

Chiang, C. et al. (2023). Can Large Language Models Be an Alternative to Human Evaluations?  
<https://arxiv.org/abs/2305.01937>. arXiv:2305.01937v1

Chu, T., Song, Z., Yang, C. (2023). How to Protect Copyright Data in Optimization of Large Language Models?  
<https://arxiv.org/abs/2308.12247>. arXiv:2308.12247v1

CIO (2023). Chief AI Officer: What it takes to land the C-suite's hottest new job. <https://www.cio.com/article/657977/chief-ai-officer-what-it-takes-to-land-the-c-suites-hottest-new-job.html>

Cui, Q. et al. (2022). Contrastive Vision-Language Pre-training with Limited Resources. <https://arxiv.org/abs/2112.09331>. arXiv:2112.09331v3

CommetML. <https://www.comet.com/site/>.

Datta, T. et al. (2023). Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook.  
<https://arxiv.org/abs/2303.06223>. arXiv:2303.06223v1

Dettmers, T. et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs <https://arxiv.org/abs/2305.14314>. arXiv:2305.14314v1

Devlin, J. et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.  
<https://arxiv.org/abs/1810.04805>. arXiv:1810.04805v2

Duan, J. et al. (2023). Shifting attention to relevance: towards the uncertainty estimation of large language models.  
<https://arxiv.org/abs/2307.01379>. arXiv:2307.01379v2

Dun, C. et al. (2024). Sweeping Heterogeneity with Smart MoPs: Mixture of Prompts for LLM Task Adaptation.  
<https://arxiv.org/abs/2310.02842>. arXiv:2310.02842v2

Elazar, Y. et al. (2021). Measuring and Improving Consistency in Pretrained Language Models.  
<https://aclanthology.org/2021.tacl-1.60/>.

Euronews (2023). 2023 was the year AI went mainstream. It was also the year we started to panic about it.  
<https://www.euronews.com/next/2023/12/27/2023-was-the-year-ai-went-mainstream-it-was-also-the-year-we-started-to-panic-about-it>

- European Parliament (2024). Artificial Intelligence Act / European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). <https://artificialintelligenceact.eu/>; <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>
- European Commission (2024). Knowledge Center on Interpretation. <https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/what-large-language-model>
- Fisher, M., Campagna, G., Choi, E., Lam, M. S., Freund, S. N., Yahav, E., (2021). DIY Assistant: A Multi-modal End-User Programmable Virtual Assistant. <https://dl.acm.org/doi/10.1145/3453483.3454046>.
- Gartner (2023). What is generative AI? <https://www.gartner.com/en/topics/generative-ai>
- Google DeepMind (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. Meredith Ringel Morris; Jascha Sohl-Dickstein; Noah Fiedel; Tris Warkentin; Allan Dafoe; Aleksandra Faust; Clement Farabet; and Shane Legg. [arXiv:2311.02462v1](https://arxiv.org/abs/2311.02462v1)
- Google + Implement (2023). The economic opportunity of generative AI in D9+. An Implement Consulting Group study commissioned by Google.
- Gozalo-Brizuela, R., y Garrido-Merchán, E.C. (2023). A survey of Generative AI Applications. <https://arxiv.labs.arxiv.org/html/2306.02781>
- Guo, Z. et al. (2023). Evaluating Large Language Models: A Comprehensive Survey. <https://arxiv.org/pdf/2310.19736.pdf>. [arXiv:2310.19736v3](https://arxiv.org/abs/2310.19736v3)
- Guzman, F. et al. (2015). How do Humans Evaluate Machine Translation. <https://aclanthology.org/W15-3059.pdf>.
- Fu, HY. et al. (2023). Estimating Large Language Model Capabilities without Labeled Test Data. <https://arxiv.org/abs/2305.14802>. [arXiv:2305.14802v2](https://arxiv.org/abs/2305.14802v2)
- Fu, X. et al (2024). Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization? <https://arxiv.org/abs/2402.00841>. [arXiv:2402.00841](https://arxiv.org/abs/2402.00841)
- Goyal, S. et al (2024). LLMGuard: Guarding Against Unsafe LLM Behavior. <https://arxiv.org/abs/2403.00826>. [arXiv:2403.00826v1](https://arxiv.org/abs/2403.00826v1)
- Hendrycks, D. et al (2021). Measuring Massive Multitask Language Understanding. <https://arxiv.org/abs/2009.03300>. [arXiv:2009.03300v3](https://arxiv.org/abs/2009.03300v3)
- Huang, L. et al. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. <https://arxiv.org/abs/2311.05232>. [arXiv:2311.05232v1](https://arxiv.org/abs/2311.05232v1)
- Hugging Face Datasets (2024). CodeParrot. <https://huggingface.co/codeparrot>.
- IAPP (2024). Global AI Law and Policy Tracker. <https://iapp.org/resources/article/global-ai-legislation-tracker/>
- iDanae 2T23 (2023): Large Language Models: una nueva era en la inteligencia artificial. Cátedra iDanae. Newsletter trimestral 2T23. <http://www.idanae-stem.com/>
- iDanae 1T24 (2024): Hacia una inteligencia artificial sostenible. Cátedra iDanae. Newsletter trimestral 1T24. <http://www.idanae-stem.com/>
- Imperial, JM., et al. (2023). Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. <https://arxiv.org/abs/2309.05454>. [arXiv:2309.05454v2](https://arxiv.org/abs/2309.05454v2)
- IndesIA (2024). Barómetro de adopción de la inteligencia artificial en las pymes españolas. <https://www.indesia.org/wp-content/uploads/2024/04/IndesIA.-Barometro-de-adopcion-de-la-inteligencia-artificial-en-las-pymes-espanolas-Edicion-2024.pdf>
- Jang et al. (2022). Knowledge unlearning for mitigating privacy risks in language models. <https://arxiv.org/abs/2210.01504>. [arXiv:2210.01504](https://arxiv.org/abs/2210.01504).
- Jia, C. et al (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. <https://arxiv.org/abs/2102.05918>. [arXiv:2102.05918v2](https://arxiv.org/abs/2102.05918v2)
- Kahng, M. et al. (2024). LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. <https://arxiv.org/abs/2402.10524>. [arXiv:2402.10524v1](https://arxiv.org/abs/2402.10524v1)
- Kuchnik, M. et al. (2023). Validating Large Language Models with Realm. <https://arxiv.org/abs/2211.15458>. [arXiv:2211.15458v2](https://arxiv.org/abs/2211.15458v2)
- Kudo, T., Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. <https://arxiv.org/abs/1808.06226>. [arXiv:1808.06226v1](https://arxiv.org/abs/1808.06226v1)
- Lam, M. (2018). <https://profiles.stanford.edu/monica-lam?tab=publications>. Keeping the Internet Open with an Open-Source Virtual Assistant.
- Lee, C. et al (2024). OrchestraLLM: Efficient Orchestration of Language Models for Dialogue State Tracking. <https://arxiv.org/abs/2311.09758>. [arXiv:2311.09758v2](https://arxiv.org/abs/2311.09758v2)

- Lee, J. et al. (2022). Seq2Seq-SC: End-to-End Semantic Communication Systems with Pre-trained Language Model. <https://arxiv.org/abs/2210.15237>. arXiv:2210.15237v2
- Lester, B. et al. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. <https://arxiv.org/abs/2104.08691>. arXiv:2104.08691v2
- Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. <https://arxiv.org/abs/2005.11401>
- Li, H. et al. (2024). Digger: Detecting Copyright Content Misuse in Large Language Model Training. <https://arxiv.org/abs/2401.00676>. arXiv:2401.00676v1
- Li, S. et al (2024). Evaluating Quantized Large Language Models. <https://arxiv.org/abs/2402.18158>. arXiv:2402.18158v1
- Li, Y. et al (2023). A Survey on Fairness in Large Language Models. <https://arxiv.org/abs/2308.10149>. arXiv:2308.10149.
- Liang, P. et al. (2023). Holistic Evaluation of Language Models. <https://arxiv.org/abs/2211.09110>. arXiv:2211.09110v2
- Liu, T. et al (2022). Autoregressive Structured Prediction with Language Models. <https://arxiv.org/abs/2210.14698>. arXiv:2210.14698v2
- Liu, Y. et al (2024). Datasets for Large Language Models: A Comprehensive Survey. <https://arxiv.org/abs/2402.18041>. arXiv:2402.18041v1
- Liu, Y. et al (2023). Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models. <https://arxiv.org/pdf/2308.07847.pdf>. arXiv:2308.07847v1
- Luo, Y. et al. (2023). An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. <https://arxiv.org/pdf/2308.08747.pdf>. arXiv:2308.08747v3
- Management Solutions (2023). Explainable artificial intelligence - Desafios na interpretabilidade de modelos.. <https://www.managementsolutions.com/pt-br/microsites/whitepapers/explainable-artificial-intelligence>
- Management Solutions (2022). Auto Machine Learning, rumo à automação dos modelos <https://www.managementsolutions.com/pt-br/publicacoes-e-eventos/informes-sectoriais/white-papers/auto-machine-learning-rumo-a-automacao-dos-modelos>
- Management Solutions (2014). Model Risk Management: Quantitative and Qualitative Aspects. <https://www.managementsolutions.com/pt-br/publicacoes-e-eventos/informes-sectoriais/white-papers/model-risk-management>
- Meeus, M. et al. (2024). Copyright Traps for Large Language Models. <https://arxiv.org/abs/2402.09363>. arXiv:2402.09363v1
- Mehta, S.V. et al. (2023). An Empirical Investigation of the Role of Pre-training in Lifelong Learning. <https://arxiv.org/abs/2112.09153>. arXiv:2112.09153v2
- Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. <https://arxiv.org/abs/1301.3781>. arXiv:1301.3781v3.
- Minaee, S. et al. (2024). Large Language Models: A Survey. <https://arxiv.org/abs/2402.06196>. arXiv:2402.06196v2
- MindsDB (2024). A Comparative Analysis of Leading Large Language Models. <https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models>
- Möckander, J. et al. (2023). Auditing large language models: a three-layered approach. [arXiv:2302.08500v2](https://arxiv.org/abs/2302.08500)
- Nasr, M., et al. (2023). <https://arxiv.org/pdf/2311.17035.pdf>. arXiv:2311.17035v1
- Neelakantan, A. et al. (2022). Text and Code Embeddings by Contrastive Pre-Training. <https://arxiv.org/abs/2201.10005>. arXiv:2201.10005v1
- NIST (2023). AI Risk Management Framework | NIST. <https://www.nist.gov/itl/ai-risk-management-framework>
- Oneto, L., Chiappa, S. (2020). Fairness in Machine Learning. [2012.15816.pdf](https://arxiv.org/pdf/2012.15816.pdf) (arxiv.org) arXiv:2012.15816v1
- OpenAI (2024). Prompt engineering. <https://platform.openai.com/docs/guides/prompt-engineering>
- Ovadia, O. et al (2024). Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. <https://arxiv.org/pdf/2312.05934.pdf>. arXiv:2312.05934v3
- Pankajakshan, R. et al (2024). Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal. <https://arxiv.org/html/2403.13309v1.html>. arXiv:2403.13309v1.
- Parikh, A. P., et al. (2016). A Decomposable Attention Model for Natural Language Inference. <https://arxiv.org/abs/1606.01933>. arXiv:1606.01933v2
- Penedo, G. et al (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. <https://arxiv.org/abs/2306.01116>. arXiv:2306.01116v1
- Pew Research Center (2023). Experts Predict the Best and Worst Changes in Digital Life by 2035.
- Project Gutenberg (2024). <https://www.gutenberg.org/>.

- Rae, JW, et al (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. <https://arxiv.org/abs/2112.11446>. arXiv:2112.11446
- Rafailov, R. et al (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. <https://arxiv.org/abs/2305.18290>. arXiv:2305.18290v2
- Rejeleene, R.; Xu, X.; Talburt, J.; (2024). Towards Trustable Language Models: Investigating Information Quality of Large Language Models. <https://arxiv.org/abs/2401.13086>. arXiv:2401.13086v1
- Risk.net. (2024). The bank quant who wants to stop gen AI hallucinating. <https://www.risk.net/risk-management/7959062/the-bank-quant-who-wants-to-stop-gen-ai-hallucinating>.
- Sachdeva, N., et al (2024). How to Train Data-Efficient LLMs. <https://arxiv.org/html/2402.09668v1>. arXiv:2402.09668v1
- Samsi, S., et al (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. <https://arxiv.org/pdf/2310.03003.pdf>. arXiv:2310.03003v1
- Sarti, G. et al (2023). Inseq: An Interpretability Toolkit for Sequence Generation Models. [2302.13942] Inseq: An Interpretability Toolkit for Sequence Generation Models (arxiv.org). arXiv:2302.13942v3
- Searle, J. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences*, vol. 3. Cambridge University Press. <https://web.archive.org/web/20010221025515/http://www.bbsonline.org/Preprints/OldArchive/bbs.searle2.html>
- Shaikh, O. et al. (2022). On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. <https://arxiv.org/abs/2212.08061>. arXiv:2212.08061v2
- SHAP documentation. <https://shap.readthedocs.io/>
- Shaw, P. et al (2018). Self-Attention with Relative Position Representations. <https://arxiv.org/abs/1803.02155>. arXiv:1803.02155v2
- Sherstinsky, A. (2018). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. <https://arxiv.org/abs/1808.03314>. arXiv:1808.03314v10
- Shi, W. et al (2024). Detecting pretraining data from large language models. <https://arxiv.org/abs/2310.16789>. arXiv:2310.16789v3
- Singh, C. et al (2024). Rethinking Interpretability in the Era of Large Language Models. <https://arxiv.org/abs/2402.01761>. arXiv:2402.01761v1
- Sinha, K. et al (2021). Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. <https://arxiv.org/abs/2104.06644>. arXiv:2104.06644v2
- Soskek (2019). BookCorpus. <https://github.com/soskek/bookcorpus>.
- Su, J., et al (2021). Roformer: Enhanced transformer with rotary position embedding. <https://arxiv.org/abs/2104.09864>. arXiv:2104.09864.
- Sutskever, I. et al (2014). Sequence to Sequence Learning with Neural Networks. <https://arxiv.org/abs/1409.3215>. arXiv:1409.3215v3
- The Next Web (2023). When will AGI arrive? Here's what our tech lords predict. <https://thenextweb.com/news/when-will-agi-arrive-tech-experts-predict-artificial-general-intelligence>
- Tian, Y. et al (2024). TinyLLM: Learning a Small Student from Multiple Large Language Models. <https://arxiv.org/abs/2402.04616>. arXiv:2402.04616
- Tirumala, K. et al. (2023). D4: Improving LLM Pretraining via Document De-Duplication and Diversification. <https://arxiv.org/abs/2308.12284>. arXiv:2308.12284v1
- UK Government (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>
- Vartziotis, T. et al (2024). Learn to Code Sustainably: An Empirical Study on LLM-based Green Code Generation. <https://arxiv.org/html/2403.03344v1>. arXiv:2403.03344v1.
- Vaswani, A. et al. (2017). Attention Is All You Need. <https://arxiv.org/abs/1706.03762>
- Wan, Z. et al (2024). Efficient Large Language Models: A Survey. <https://arxiv.org/pdf/2312.03863.pdf>. arXiv:2312.03863v3
- Wang, Q. et al (2024). LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools. [2401.12576] LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools (arxiv.org). arXiv:2401.12576v1
- Wang, Y. et al (2024). Two-stage LLM Fine-tuning with Less Specialization and More Generalization. <https://arxiv.org/html/2211.00635v3>. arXiv:2211.00635v3
- Wei, J. et al (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. <https://arxiv.org/abs/2201.11903>. arXiv:2201.11903v6

- Wenzek, G., et al (2019). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. <https://arxiv.org/abs/1911.00359>. arXiv:1911.00359v2
- Wettig, A. et al. (2024). QuRating: Selecting High-Quality Data for Training Language Models. <https://arxiv.org/abs/2402.09739>. arXiv:2402.09739v1
- Weights & Biases: The AI Developer Platform ([wandb.ai](https://wandb.ai/)). <https://wandb.ai/site>
- Wikipedia (2024). Dumps. <https://dumps.wikimedia.org/zhwiki/latest/>.
- Wired (2023). OpenAI's CEO Says the Age of Giant AI Models Is Already Over. <https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/>
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. <https://dl.acm.org/doi/10.1145/365153.365168>
- White House (2022). Blueprint for an AI Bill Of Rights. Making Automated Systems Work for the American People. <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>
- White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>
- Wu, X. et al. (2023). Depn: Detecting and editing privacy neurons in pretrained language models. <https://arxiv.org/abs/2310.20138>. arXiv:2310.20138.
- Xin Zhao, W., et al. (2023). A Survey of Large Language Models. <https://arxiv.org/abs/2303.18223>. arXiv:2303.18223v13
- Xu, L. et al. (2023). Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. <https://arxiv.org/pdf/2312.12148.pdf>. arXiv:2312.12148v1
- Xu, Y. et al. (2021). Non-Autoregressive Text Generation with Pre-trained Language Models. <https://aclanthology.org/2021.eacl-main.18/>
- Xu, Z. et al. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. <https://arxiv.org/abs/2401.11817>. arXiv:2401.11817v1
- Yang, J. et al. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. <https://arxiv.org/abs/2304.13712>. arXiv:2304.13712v2
- Yidiz, C. et al (2024). Investigating Continual Pretraining in Large Language Models: Insights and Implications. <https://arxiv.org/html/2402.17400v1>. arXiv:2402.17400v1
- Yu, C. et al. (2023). Unlearning bias in language models by partitioning gradients. <https://aclanthology.org/2023.findings-acl.375.pdf>.
- Yogarajan, V., et al (2023). Tackling Bias in Pre-trained Language Models: Current Trends and Under-represented Societies. <https://arxiv.org/pdf/2312.01509.pdf>. arXiv:2312.01509v1
- Zaharia, M. et al (2018). Accelerating the Machine Learning Lifecycle with MLflow. [https://people.eecs.berkeley.edu/~matei/papers/2018/ieee\\_mlflow.pdf](https://people.eecs.berkeley.edu/~matei/papers/2018/ieee_mlflow.pdf).
- Zeng, Y., et al (2023). CLIP2: Contrastive Language-Image-Point Pretraining from Real-World Point Cloud Data. <https://arxiv.org/abs/2303.12417>. arXiv:2303.12417v2
- Zhang, B. et al (2024). When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. <https://arxiv.org/abs/2402.17193>. arXiv:2402.17193v1
- Zhang, L. et al (2024). Enhancing Large Language Model Performance To Answer Questions and Extract Information More Accurately. <https://arxiv.org/html/2402.01722v1>. arXiv:2402.01722v1.
- Zhang, S. et al (2023). Instruction Tuning for Large Language Models: A Survey. [https://www.researchgate.net/publication/373263398\\_Instruction\\_Tuning\\_for\\_Large\\_Language\\_Models\\_A\\_Survey](https://www.researchgate.net/publication/373263398_Instruction_Tuning_for_Large_Language_Models_A_Survey).
- Zhang, Y. et al (2024). Bias Mitigation in Fine-tuning Pre-trained Models for Enhanced Fairness and Efficiency. <https://arxiv.org/html/2403.00625v1>. arXiv:2403.00625v1
- Zhao, B., et al (2023). Tuning LayerNorm in Attention: Towards Efficient Multi-Modal LLM Finetuning. <https://arxiv.org/abs/2312.11420>. arXiv:2312.11420v1
- Zhou, C. et al (2023). LIMA: Less Is More for Alignment. <https://arxiv.org/abs/2305.11206>. arXiv:2305.11206v1
- Zhou, N., et al (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. <https://arxiv.org/abs/2105.06558>. arXiv:2105.06558v1



**Nosso objetivo é superar as expectativas dos nossos clientes sendo parceiros de confiança**

A Management Solutions é uma empresa internacional de serviços de consultoria com foco em assessoria de negócios, riscos, organização e processos, tanto sobre seus componentes funcionais como na implementação de tecnologias relacionadas.

Com uma equipe multidisciplinar (funcionais, matemáticos, técnicos, etc.) de mais de 4.000 profissionais, a Management Solutions desenvolve suas atividades em 46 escritórios (20 na Europa, 22 nas Américas, 2 na Ásia, 1 na África e 1 na Oceania).

Para atender às necessidades de seus clientes, a Management Solutions estruturou suas práticas por setores (Instituições Financeiras, Energia e Telecomunicações) e por linha de negócio, reunindo uma ampla gama de competências de Estratégia, Gestão Comercial e de Marketing, Gestão e Controle de Riscos, Informação Gerencial e Financeira, Transformação: Organização e Processos, e Novas Tecnologias.

A área de P&D presta serviços aos profissionais e clientes da Management Solutions nos aspectos quantitativos necessários para a realização de projetos com rigor e excelência, por meio da aplicação das melhores práticas e da pesquisa contínua das últimas tendências em inteligência artificial, data science, machine learning, modelagem e big data.

**Javier Calvo Martín**

Sócio

*javier.calvo.martin@managementsolutions.com*

**Manuel Ángel Guzmán Caba**

Sócio

*manuel.guzman@managementsolutions.com*

**Luz Ferrero Peña**

Manager

*luz.ferrero@msgermany.com.de*

**Segismundo Jiménez Láinez**

Manager

*segismundo.jimenez@msspain.com*



**Management Solutions, serviços profissionais de consultoria**

**Management Solutions** é uma firma internacional de serviços de consultoria focada na assessoria de negócio, riscos, finanças, organização e processos

Para mais informações acesse [www.managementsolutions.com](http://www.managementsolutions.com)

Siga-nos em:

© Management Solutions. 2024

Todos os direitos reservados.

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña Málaga London Frankfurt Düsseldorf Paris Amsterdam Copenhagen Oslo Stockholm Warszawa Wrocław Zürich Milano Roma Bologna Lisboa Beijing İstanbul Johannesburgo Sydney Toronto New York New Jersey Boston Pittsburgh Atlanta Birmingham Houston Miami SJ de Puerto Rico San José Ciudad de México Monterrey Querétaro Medellín Bogotá Quito São Paulo Rio de Janeiro Lima Santiago de Chile Buenos Aires