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[25]

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"[26].

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"[27].

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[28] 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[29] por Bengio et al. em 2003. Isso estabeleceu a base para inovações subsequentes na representação de palavras, principalmente a introdução de embeddings de palavras[30] por Mikolov et al. em 2013 por meio do Word2Vec. Os embeddings

[25]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.

[26]European Commission (2024).

[27]Wired (2023).

[28]Bréal (1883).

[29]Bengio (2003).

[30]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[31], 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"[32] 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 melhoram 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"[33], 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

[31] Parikh, A. P. (2016).
[32] Vaswani (2017).
[33] Euronews (2023).
[34] Adaptado de MindsDB (2024) e ampliado.

Figura 4. Alguns dos principais LLM e seus fornecedores[34].

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.	España
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.	Alemania
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.

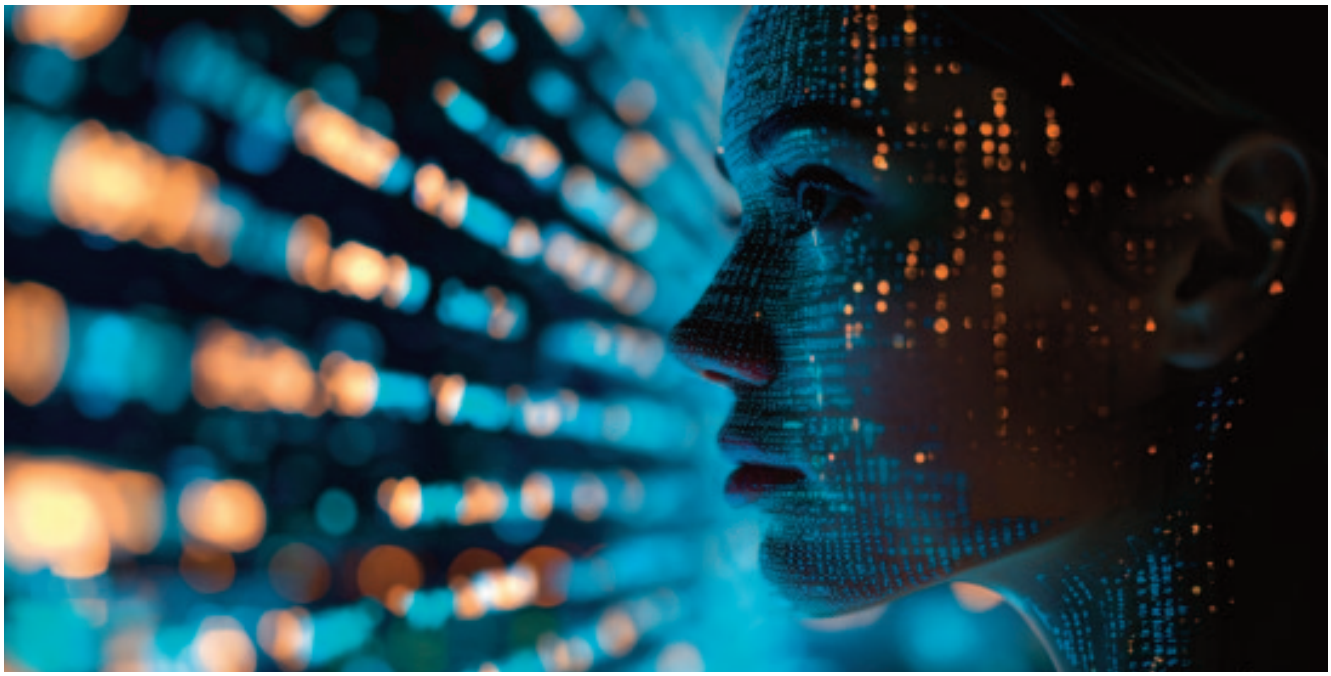
Por arquitetura

- **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 extrair 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 insumos 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 transparência, 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 transparência.

- **O guia sobre IA do NIST dos EUA**[35] **:** 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 transparência, segurança e imparcialidade, assinado por vários países.

[35] 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[36] 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.

[36] IAPP (2024).