

# 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). entrevista 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 50 <sup>th</sup> 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 90 <sup>o</sup> percentil de adultos <b>Nível</b> 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 99 <sup>o</sup> 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.

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

## 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).

