Inteligência artificial customizada e automação de processos: por que o ChatGPT não serve para organizações?

Customized artificial intelligence and process automation: why ChatGPT is not suitable for organizations?

Inteligencia artificial personalizada y automatización de procesos: ¿por qué no es adecuado ChatGPT para las organizaciones?

Márcio Carneiro dos Santos

- Doutor pelo programa Tecnologias da Inteligência e Design Digital da Pontifícia Universidade Católica de São Paulo (TIDD/PUC-SP).
- Professor Associado do Departamento de Comunicação Social da Universidade Federal do Maranhão (UFMA), na área de Jornalismo em Redes Digitais.
- É coordenador do Laboratório de Convergência de Mídias (Labcom) e atualmente também coordenador do Programa de Pós-graduação Profissional em Comunicação (PPGCOMPRO) da UFMA.
- E-mail: márcio.carneiro@ufma.br

RESUMO

Discutem-se as características dos modelos de geração de linguagem da subárea da inteligência artificial conhecida como generativa (IAG) e seu possível uso no ambiente organizacional a partir do produto ChatGPT. Defende-se a hipótese de que em sua configuração padrão, o principal representante desse segmento não é adequado por razões diversas que são analisadas. Em contrapartida, propõe-se o conceito da IA customizada, operando através de interfaces de programação de aplicação (APIs) e aplicações específicas, que utilizam estratégias mais atualizadas para contornar essas limitações, aumentando seu controle.

PALAVRAS-CHAVE: INTELIGÊNCIA ARTIFICIAL • CHATGPT • RAG • AUTOMAÇÃO • ORGANIZAÇÕES.

ABSTRACT

This study discusses the characteristics of language generation models in generative artificial intelligence and their possible use in organizations based on ChatGPT. This research hypothesized that in its standard configuration, the main representative of this segment is unsuitable for organizations for reasons analyzed therein. In contrast, this study proposes the concept of a customized AI that operates by specific applications and application software that use strategies such as retrieval augmented generation to circumvent these limitations, increasing control.

KEYWORDS: ARTIFICIAL INTELLIGENCE • CHATGPT • RAG • AUTOMATION • ORGANIZATIONS.

RESUMEN

En este texto se discuten las características de los modelos de generación de lenguaje provenientes del subárea de la inteligencia artificial conocida como generativa (IAG) y su posible uso en el entorno organizacional basado en el producto ChatGPT. Se defiende la hipótesis de que en su configuración estándar el principal representante de este segmento no está apto por diversos motivos que se analizan. En cambio, se propone el concepto de IA personalizada, que funciona mediante interfaces de programación de aplicaciones (API) y aplicaciones específicas las cuales utilizan estrategias más actualizadas para superar estas limitaciones aumentando su control.

PALABRAS CLAVE: INTELIGENCIA ARTIFICIAL • CHATGPT • RAG • AUTOMATIZACIÓN • ORGANIZACIONES.

INTRODUÇÃO

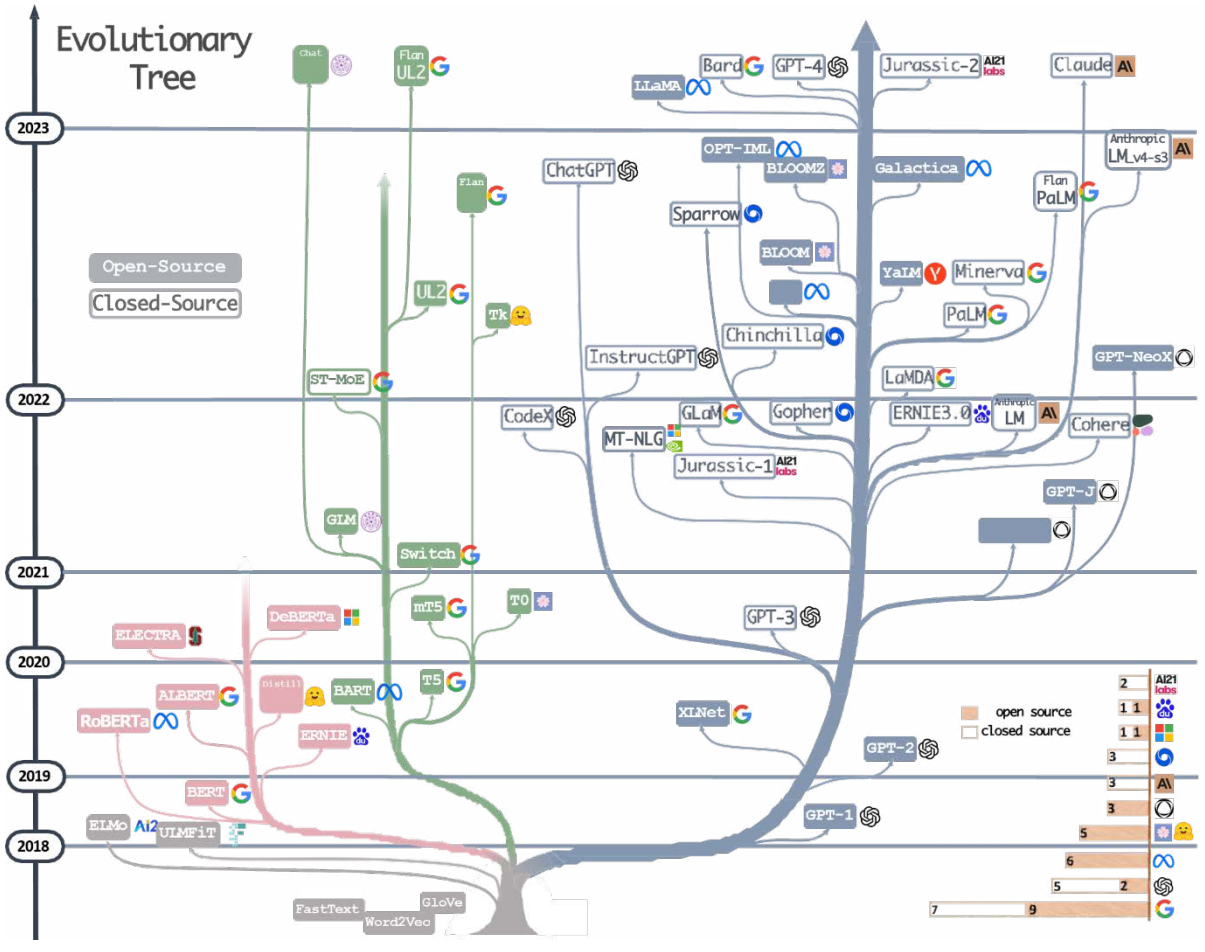
Parte 1

Lançado para uso público em outubro de 2022, o ChatGPT, aplicação de *inteligência artificial generativa* (IAG) que usa uma interface com formato de *bot* conversacional do *modelo de geração de linguagem GPT*, quebrou todos os recordes anteriores de velocidade de adoção tecnológica, ao conseguir, em uma semana, a marca de um milhão de usuários, segundo seus próprios criadores.

GPT, sigla para *Generative Pre-Trained Transformer*, é um *modelo de linguagem pré-treinado* (*Pre-trained Language Model, PLM*) desenvolvido pela OpenAI[2], capaz de, a partir de uma base para aprendizado de bilhões de parâmetros, operar com a linguagem humana inicialmente em formato de texto, tanto apreendendo sentido (ao seu modo), como também gerando conteúdo de forma bastante eficiente, a ponto de fazer muitas pessoas terem a ilusão de estarem conversando com um humano.

Transformers são um tipo específico de *rede neural*, constructo da área da Ciência da Computação que representa uma forma de treinar algoritmos (*machine learning*) usando o cérebro humano como modelo e metáfora. Foi o desenvolvimento desse tipo de tecnologia que tornou possível a geração dos *Large Language Models* (LLMs), modelos que são treinados a partir de imensas bases de dados que tem seu desenvolvimento iniciado em 2018 (Figura 1).

Figura 1: Árvore evolucionária dos sistemas LLMs



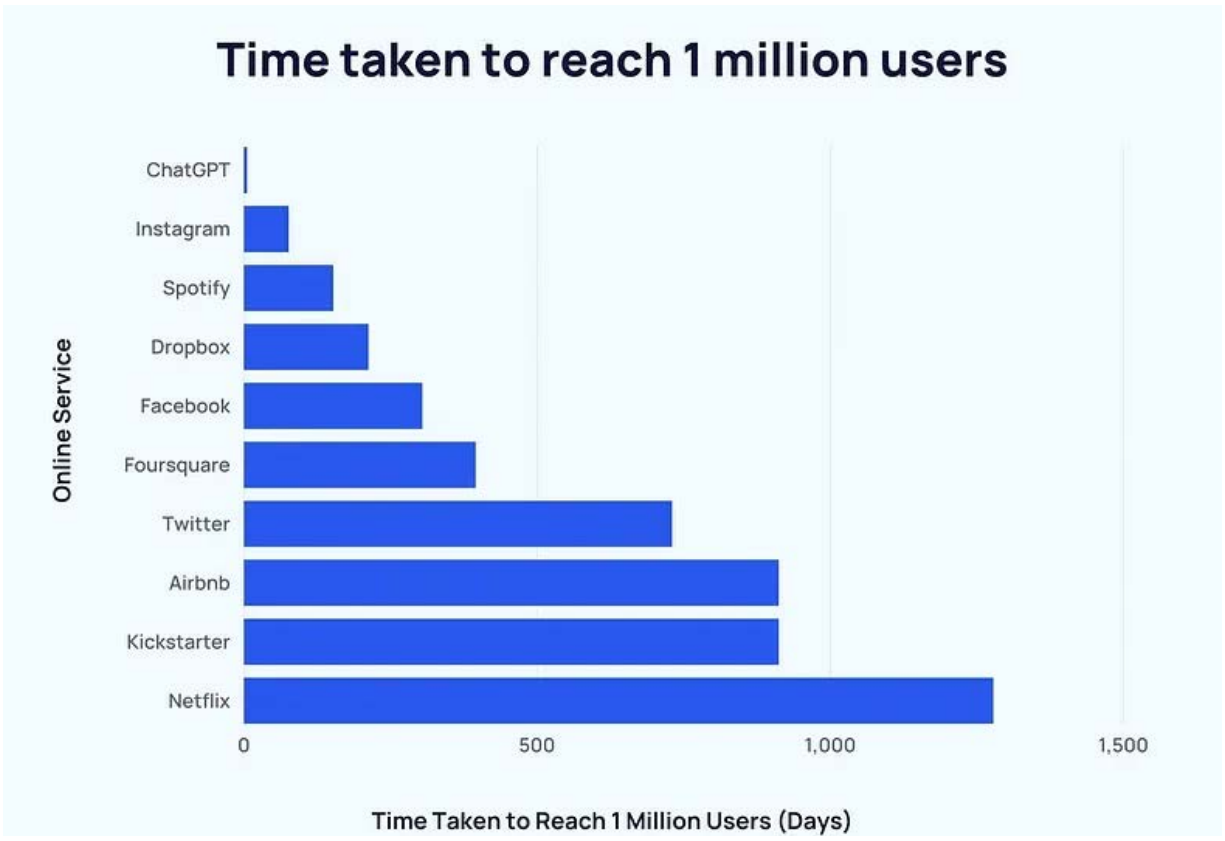
Fonte: Yang et al., 2023.

2 Empresa de tecnologia que desenvolve vários produtos usando inteligência artificial. Disponível em: https://openai.com/. Acesso em: 18 abr. 2024.

Parte 2

Segundo dados da própria empresa desenvolvedora, cerca de um ano depois, ao final de 2023, semanalmente existem 100 milhões de pessoas acessando a plataforma ChatGPT, consolidando um sucesso inesperado e despertando enorme atenção pública. Essa performance fez da ferramenta o mais veloz caso de adoção tecnológica recente (Figura 2).

Figura 2: Tempo em dias para conseguir um milhão de usuários



Fonte: Swansburg, 2023.

Inteligência Artificial (IA), termo antes apenas associado ao campo da Ciência da Computação, de uma hora para outra, passou a ser tema de conversa em mesa de bar, *trend* de busca no Google, fórmula para ganhar dinheiro fácil no YouTube e, para muitos também, a inimiga pública número um dos trabalhadores de diversos segmentos, capaz de materializar seu principal pesadelo: a ameaça de ter seu emprego roubado por uma máquina.

Tal situação, infelizmente, já aconteceu diversas vezes antes na história moderna, desde a revolução industrial e em tantos outros momentos nos quais a adoção tecnológica impôs seu inexorável poder de mudança sobre as coisas humanas. Contudo, a novidade sempre faz mais sucesso e, agora, realmente com ingredientes inéditos: o alcance e a velocidade que um ecossistema de meios digitais interconectados fornece à escala dos impactos das transformações geradas.

Parte 3

Apesar de ser um campo de conhecimento com décadas de desenvolvimento, com suas origens apontando para o tempo da Segunda Guerra Mundial e representando um grande guarda-chuva para diversas subáreas, a novidade da inteligência artificial, recém-descoberta por tantos, engendra várias contradições e provavelmente uma delas pode ser representada já nos primeiros parágrafos deste texto.

De que forma uma área tão específica e com anos de produção científica, aqui representada com uma pequena amostra, usando os termos que destaquei na Parte 1, pode ter se transformado nessa figurinha fácil de tantos especialistas, que sobre

ela falam com enorme desenvoltura, como se o passado não existisse e tudo fosse simples e fácil como parece ser, quando feito pelo seu novo professor, de onde tiram a maior parte de suas observações e conclusões, o ChatGPT?

Uma segunda e mais específica contradição, sobre a qual gostaríamos de nos debruçar neste artigo e que, sim, tem direta relação com a primeira, ou seja, o fato de que, mesmo com a fama e a curiosidade geral, o uso do ChatGPT em seu formato mais conhecido, de trabalho genérico e acesso gratuito, como um *bot* enciclopédia que parece saber sobre tudo e todos, no cenário das organizações não é o ideal e muito menos desejável, basicamente por suas características técnicas e *modus operandi*, tão solenemente desconhecidos por muitos.

É óbvio que no estágio em que estamos muitas organizações, ao tempo em que este texto está sendo escrito, estão atualmente testando a solução ChatGPT. Entretanto, mais cedo ou mais tarde, algumas constatações serão geradas, e esperamos que ajudem a corroborar a hipótese que guia este trabalho[3].

Assim, o objetivo deste texto é, a partir da descrição mais detalhada sobre o funcionamento das ferramentas de IAG, destacar suas limitações e riscos para uso organizacional, oferecendo em troca um constructo ora em desenvolvimento que chamamos de *Custom AI*, baseado em técnicas específicas tais como a *Retrieval Augmented Generation* (RAG), bem como ferramentas e frameworks que, mesmo não sendo tão famosos, efetivamente estão no centro do desenvolvimento de experiências de automação de processos organizacionais quando, eventualmente, isso pode ser feito usando a inteligência artificial generativa (IAG).

Sobre ela, já propomos em texto anterior (Santos, 2023) o seguinte conceito, que agora expandimos:

> A inteligência artificial generativa ou gerativa (*Generative AI*, em inglês) é uma subárea da inteligência artificial que se concentra em criar sistemas que tem como objetivo simular a própria criatividade humana, através da criação de imagens, sons, vídeos e texto. Esses sistemas são capazes de produzir conteúdo a partir de conjuntos de dados de treinamento baseados em bilhões de parâmetros e, por isso, pertencem à recente categoria dos LLMs (*Large Language Models*).[4] É essa característica essencial que os conecta à Comunicação e a operação das principais matrizes de linguagem, oferecendo aplicações e possibilidades de utilização em diferentes cenários e segmentos econômicos.

DE QUE IA ESTAMOS FALANDO?

Na área da Comunicação, o tópico da inteligência artificial, antes do ChatGPT, já tinha recebido atenção, através de precursores como Squirra (2016a; 2016b); Lima Jr. e Vergili (2018), Santos (2016; 2018; 2019) e, mais recentemente, Lemos (2020). Além dos pesquisadores brasileiros, poderíamos citar Reiter, Sripada e Robertson (2003), Jones (2013), Mayer-Schönberger e Cukier (2013), Manovich (2014), Latzer (2016), Guzman e Seth (2018) e Veel (2018), entre outros.

É comum ler nos dias de hoje referências do tipo “as IAs” que fazem isso ou aquilo e construções do gênero. Em tese, por tratar-se de um campo de conhecimento, a inteligência artificial existe no singular, enquanto suas implementações, produtos, modelos, aplicações ou, de um modo mais técnico, suas instanciações, são múltiplas e de existência plural, num mundo no qual seu potencial de utilização se expande a cada dia.

3 Uma metodologia recente para avaliação mais geral de ferramentas de IAG é a EEIF disponível em Santos (2023).

4 Para saber mais sobre IAG acesse um experimento específico sobre o tema em livro produzido com o auxílio do próprio ChatGPT, disponível em: https://www.labcomdata.com.br/iag.

Outra diferença importante na operação dos conceitos é que em muito material recente não se distingue de forma correta a subárea da IA Generativa (IAG), da IA campo de conhecimento maior, grande guarda-chuva de diversas subáreas, tais como o aprendizado de máquina (*machine learning*), a visão computacional (*computer vision*) e o processamento de linguagem natural (*natural language processing* ou NLP).

Tal diferenciação é importante porque a IAG, dos LLMs, opera basicamente por modelos estocásticos, ou seja, com parâmetros probabilísticos que funcionam dentro de margens de variabilidade. Isso significa que um modelo como o GPT, motor de geração de linguagem que faz o produto ChatGPT poder ser experimentado como um *bot* conversacional, vai entregar seus resultados a partir de um processo que, guardadas as devidas proporções, assemelha-se ao que vemos num aplicativo como o WhatsApp, que durante nossa digitação vai autocomplementando o que digitamos, a partir do que ele infere, probabilisticamente, que seja o mais provável para dar sequência ao que vamos escrevendo.

Um exemplo, ainda que muito mais simples do que o que realmente acontece num modelo como o GPT, seria imaginar a seguinte frase: “No domingo à tarde, para me divertir, irei ao ...”. Ao verificar correlações, na sua enorme base de dados, a partir da entrada, do *input* de texto fornecido (chamado de *prompt*) com situações de sentido semelhantes, o modelo irá verificar que as respostas mais comuns seriam “cinema” com 40% de incidência nos trechos detectados, 30% para “estádio” (de futebol), 20% para “teatro”, 9% para “shopping” e talvez, vamos citar aqui apenas a título pedagógico, 1% para “cemitério”, por mais estranho que pareça.

Com base nessas informações, em tese operando sem parâmetros de controle adicionais, o modelo devolverá aos usuários respostas diferentes ponderadas pelo peso, pela frequência da presença de cada uma delas na sua base de treinamento. Resumindo, ao responder diversas vezes à mesma questão, oferecerá como retorno, “cinema” para a maioria dos usuários; “estádio” para um número menor, mas significativo deles; como também indicará para alguns “teatro” e “shopping” e, ainda que para pouquíssimos, “cemitério”.

Para avançar um pouco mais, é bom saber que modelos de geração de linguagem, como o GPT, operam usando um parâmetro de controle chamado “*temperature*” que permite um ajuste mais fino sobre a variabilidade estocástica do conjunto possível de respostas. Por exemplo, variando numa escala de 0 a 1, a opção por *temperature(t)*=1 significaria deixar o modelo mais criativo, nos termos equivalentes ao significado disso na Comunicação, isto é apresentando respostas mais diversas e não usuais. De fato, o que realmente está acontecendo é que *t*=1, permite ao modelo operar com toda a variabilidade presente na sua base de treinamento original, ainda que respeitando a frequência ou peso de cada possibilidade dentro do seu universo de consulta. Com *t*=1, ainda que, muito raramente, “cemitério” será a resposta devolvida para alguns poucos.

Ao contrário, se o parâmetro de *temperature* tender a zero, mais rígido e limitado será o conjunto possível de respostas, fazendo o modelo agir de forma mais focada apenas no que é mais probabilisticamente importante na sua base, ou resumindo, com *t*=0, apenas “cinema” será a resposta para todas as diferentes solicitações semelhantes.

O entendimento desse mecanismo básico deve ser considerado para uma outra conclusão importante: a precisão não é o forte das ferramentas de IAG. Uma das possíveis consequências indiretas disso está ligada à discussão sobre o seu uso para geração de desinformação, por exemplo, como analisado em Corrêa e Santos (2023).

A não ser que o usuário possa definir individualmente parâmetros para controlar o nível de objetividade que deseja nas respostas, o que não está disponível na versão pública e famosa do ChatGPT, no qual o parâmetro *temperature* do modelo GPT, está configurado para operar um pouco acima de 0,5, oferecendo respostas com uma pitada adicional de criatividade, o caráter estocástico da solução permanece.

Apesar de ser uma óbvia simplificação, o modelo de operação ora apresentado nos permite inferir que a IAG, muito interessante como um assistente que oferece ganhos de produtividade em alguns processos através da geração volumosa, rápida e com variedade de suas respostas, dentro de muitas organizações, que precisam operar com maior nível de precisão, deve ter sua adoção avaliada com mais cuidado.

É óbvio que enquanto para indústrias criativas ou negócios, como publicidade e marketing, o parâmetro *temperature* alto é mais que desejável e pode representar um diferencial na atuação, para setores nos quais a qualidade da informação é crítica, como no jornalismo, na ciência, na educação, entre outros, há que se agir com cuidado e, principalmente, conhecimento.

Como a IAG é apenas uma entre as diversas subáreas do grande campo da IA, para estas últimas atividades existem, inclusive já operando há bastante tempo, diversas outras possibilidades e implementações da inteligência artificial. Definitivamente, o uso de IA nas organizações não começou com o ChatGPT e muito menos ficará restrito a ele ou a outras soluções da mesma categoria.

Outra conclusão importante oriunda da forma como a IAG opera é a necessidade de uma camada final de validação ou curadoria humana ao final de cada processo no qual ela é utilizada.

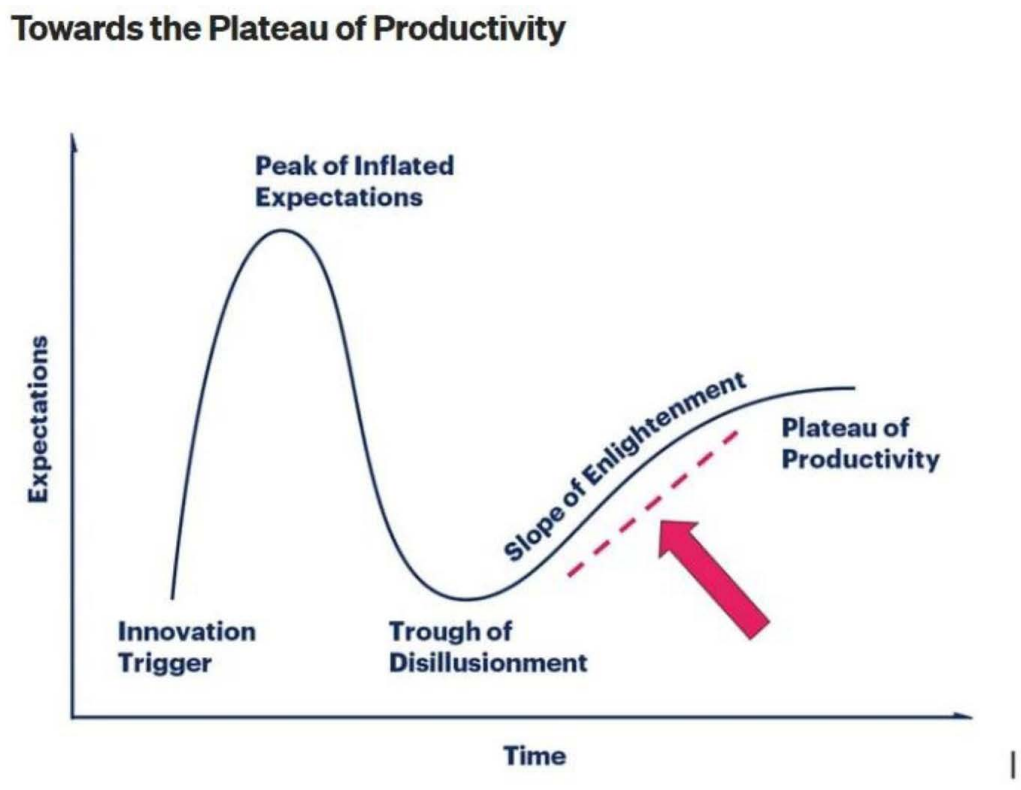
Apesar de talvez ser o exemplo mais comum e frequentemente associado a ganho de dinheiro fácil pelos gurus da internet que disseminam desinformação em canais do YouTube, por exemplo, não se tem notícia de nenhuma organização ou marca com grande nível de conhecimento público que esteja operando de forma totalmente automatizada, publicando direto posts escritos pelo ChatGPT em seus canais de mídias sociais, sem algum tipo de revisão.

Mesmo sendo essa uma tarefa que ele executa com razoável competência, o risco de um texto, mesmo que gramaticalmente correto, no qual uma palavra possa ter um duplo sentido ou um entendimento diferente do que foi inicialmente proposto e o consequente grave e rápido problema que tal mal-entendido poderia gerar numa rede social existe e de forma alguma compensaria automatizar tudo, excluindo a revisão humana do processo. Simplesmente porque o dano a uma marca numa situação assim, potencialmente geraria prejuízos muito grandes e, de forma alguma, compensaria a economia no RH, resultante da decisão de dispensar profissionais do setor.

Entendemos que a adoção tecnológica eventualmente segue um padrão como o da curva abaixo, gerando no início um pico de grandes expectativas, algumas delas que não se confirmarão na continuidade do processo, dando origem a um segundo momento de eventual descrença na tecnologia como um todo, para que, só depois de uso e experimentação intensivos, se estabeleça um novo patamar dessa adoção, baseado em conhecimento empírico sólido, indicando com maior precisão onde é possível ou não usar com produtividade adequada aquela tecnologia.

A falta de conhecimento sobre a IAG e suas especificidades nos coloca em grande parte ainda em algum ponto na curva de aprendizado antes do desejável patamar de produtividade assertiva, buscado pelas organizações.

Figura 3: Curva de expectativas sobre a adoção de tecnologias ao longo do tempo, baseado no modelo original “Gartner Hype Cycle”



Fonte: Gartner (2023).

QUAIS OS LIMITES DA IAG NAS ORGANIZAÇÕES

Além da questão da variabilidade inserida nas suas respostas ou entregas, característica importante nos modelos que operam a partir de grandes bases de conhecimento, como os LLMs, para as organizações também existem outras questões significativas.

Uma delas é o fato de que as respostas oferecidas são resultado de uma busca e processamento de informações feitos a partir da base original de treinamento, que mesmo sendo composta de bilhões de documentos, em sua maioria retirados da internet como conteúdo de sites, livros, notícias e publicações em plataforma de mídias sociais[5], pode eventualmente não conter nada a respeito de determinado tópico ou, principalmente, de temas importantes para organizações específicas.

Uma situação que poderíamos usar como exemplo seria pensar nos dados de sua operação ou cultura organizacional para ter utilidade real num chatbot de atendimento a clientes, entre outras. A menos que esse atendimento seja feito de forma muito genérica e de baixa efetividade, embutir o GPT, sem algum tipo de aperfeiçoamento numa ferramenta com esse propósito seria pouco produtivo sem as adequações necessárias.

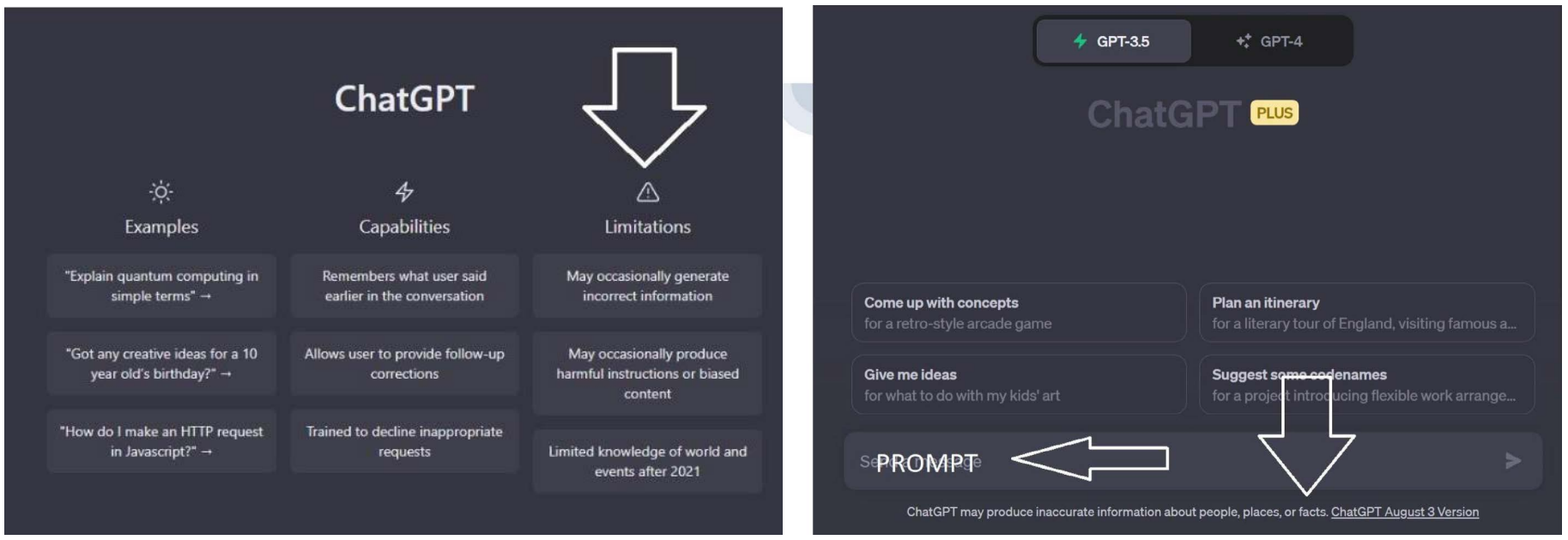
Além de serem limitadas, mesmo com bilhões de parâmetros que são incapazes de suprir as necessidades específicas da também numericamente enorme diversidade de organizações no mundo, as bases dos LLMs ainda são atualizadas em ciclos de tempo. A base de treinamento do ChatGPT, até a data da redação deste texto, só teve dois ciclos conhecidos, o primeiro que utilizou dados coletados até setembro de 2021 e o outro iniciado recentemente, com a base sendo atualizada até abril de 2023.

5 Esse fato tem sido objeto de crescente questionamento, inclusive judicial, baseado na questão de que os desenvolvedores dos modelos da IAG simplesmente se apropriaram desses documentos para treinar seus LLMs sem autorização prévia ou até conhecimento dos produtores de conteúdo.

Apenas para exemplificar, até essa última atualização, logo depois do seu lançamento e início de utilização pública, ao ser perguntado sobre quem havia ganhado o Campeonato Mundial de Futebol de 2022 (fato ocorrido após o primeiro ponto de atualização que foi até setembro de 2021), o ChatGPT polidamente explicava que não sabia, indicando inclusive a informação sobre o período de atualização da sua base, anterior ao fato sobre o qual estava sendo questionado. Se o usuário insistisse, ele simplesmente criava uma resposta, bastante detalhada, indicando os gols, os jogadores que os marcaram, o lugar da final e tudo o mais que pudesse construir a partir de informações de eventos semelhantes anteriores.

O ChatGPT nunca foi desenhado para oferecer respostas checadas, verificadas, precisas. Tal fato inclusive sempre foi informado na sua própria interface que, ao longo do tempo, sempre trouxe esse alerta.

Figura 4: Alertas nas interfaces do produto ChatGPT sobre sua possibilidade de imprecisão quando trata de dados



Fonte: OpenAI (2023).

Voltando então ao uso organizacional, usar o ChatGPT na sua versão genérica significa atrelar a utilização a uma base sobre a qual se tem o mínimo controle, com datas de atualização que também não se pode definir e até pior, com os resultados apresentados sem a indicação das fontes consultadas quando da entrega do que é pedido. Tal contexto para a maioria das organizações implica em enormes riscos.

Um experimento básico, realizado apenas como exemplo, que pode ser utilizado por outras organizações, foi feito através de uma consulta simples ao ChatGPT sobre o significado do termo SET, que no Brasil, entre outras possibilidades, refere-se à Sociedade Brasileira de Engenharia de Televisão[6].

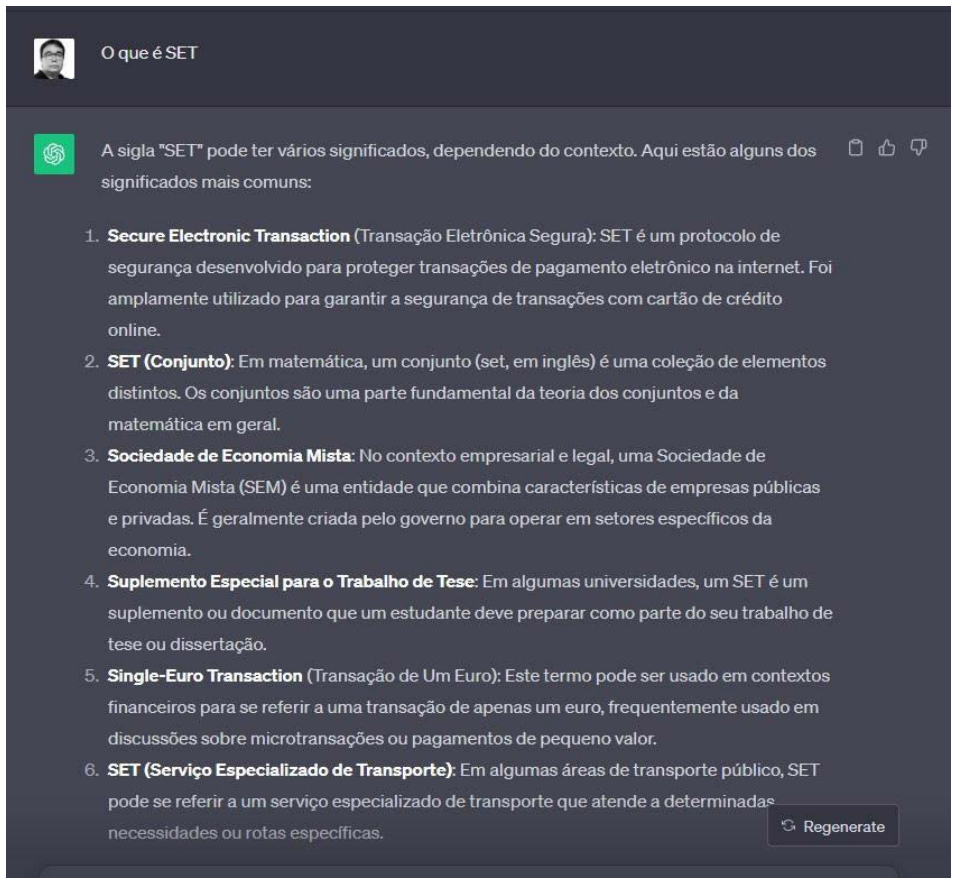
Os prints abaixo (Figuras 5 e 6) demonstram as respostas apresentadas pelo ChatGPT, inicialmente sem maior detalhamento de contexto e depois com a indicação do cenário brasileiro. O que se apreende é que, apesar da quantidade monstruosa de informações, nada ali se refere, ainda que de forma mínima, a essa organização específica.

É óbvio que, talvez, para milhares de outras organizações, com a probabilidade maior de estarem presentes nos tipos de documentos originais usados no treinamento do ChatGPT, principalmente em língua inglesa, haverá chance de se obter algum tipo de resposta correta, mas, ainda nesse caso, muito dificilmente com o nível de detalhamento ou aprofundamento

6 Disponível em: https://set.org.br/. Acesso em: 18 abr. 2024.

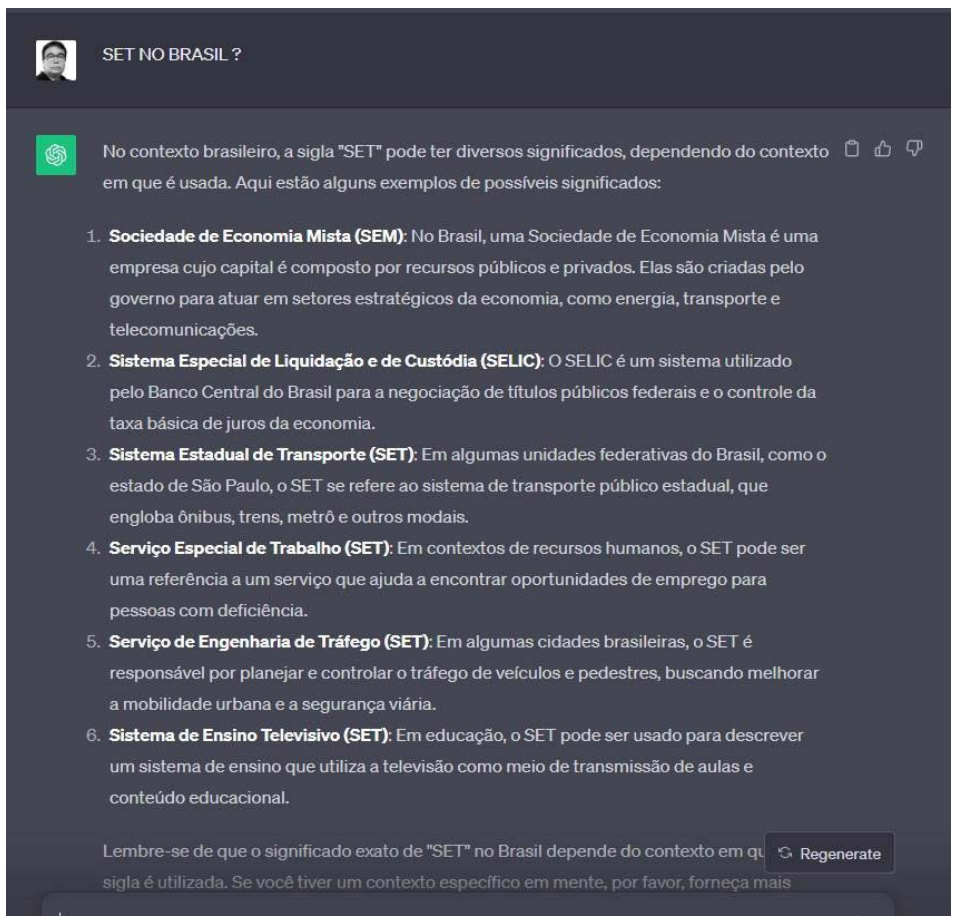
necessário para falar sobre tal organização de forma a contribuir num cenário como o relacionamento com clientes ou usuários, por exemplo.

Figura 5: Prompt solicitando ao CHATGPT para falar sobre a SET



Fonte: Elaborado pelo autor.

Figura 6: Prompt solicitando ao CHATGPT para falar sobre a SET indicando o contexto brasileiro

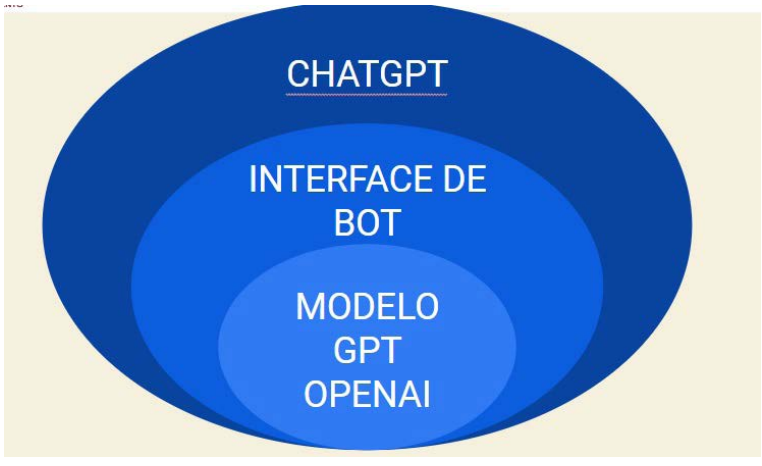


Fonte: Elaborado pelo autor.

INTELIGÊNCIA ARTIFICIAL CUSTOMIZADA – CUSTOM AI

É importante entender a estrutura interna do produto ChatGPT, uma aplicação que utiliza o motor de geração de linguagem GPT numa interface de *bot* conversacional, com parâmetros e condições muito específicas e pouco controle por conta do usuário, bastante útil em diversos cenários, mas com aplicação problemática na maioria das situações de uso corporativo ou organizacional.

Figura 7: Estrutura em camadas da aplicação ChatGPT que tem embutida para funcionar como *bot* conversacional o modelo GPT de geração de linguagem



Fonte: Elaborado pelo autor.

Assim, enquanto o produto ChatGPT é, por assim dizer, fechado, com pouco controle na configuração da sua operação interna, o modelo GPT, o motor que o movimenta, pode ser também inserido em diversos outros formatos, que oferecem maior controle, a partir do acesso via API[7], a interface de dados que está disponível para acesso aos servidores do desenvolvedor.

É importante lembrar que, ao usar a versão pública do ChatGPT, o processamento também está sendo feito nos servidores da OpenAI, que recebe os prompts do usuário, os processa e devolve o resultado, o qual vai sendo impresso na interface do *bot* de conversação, gerando ao final uma espécie de diálogo com solicitações e suas respectivas respostas.

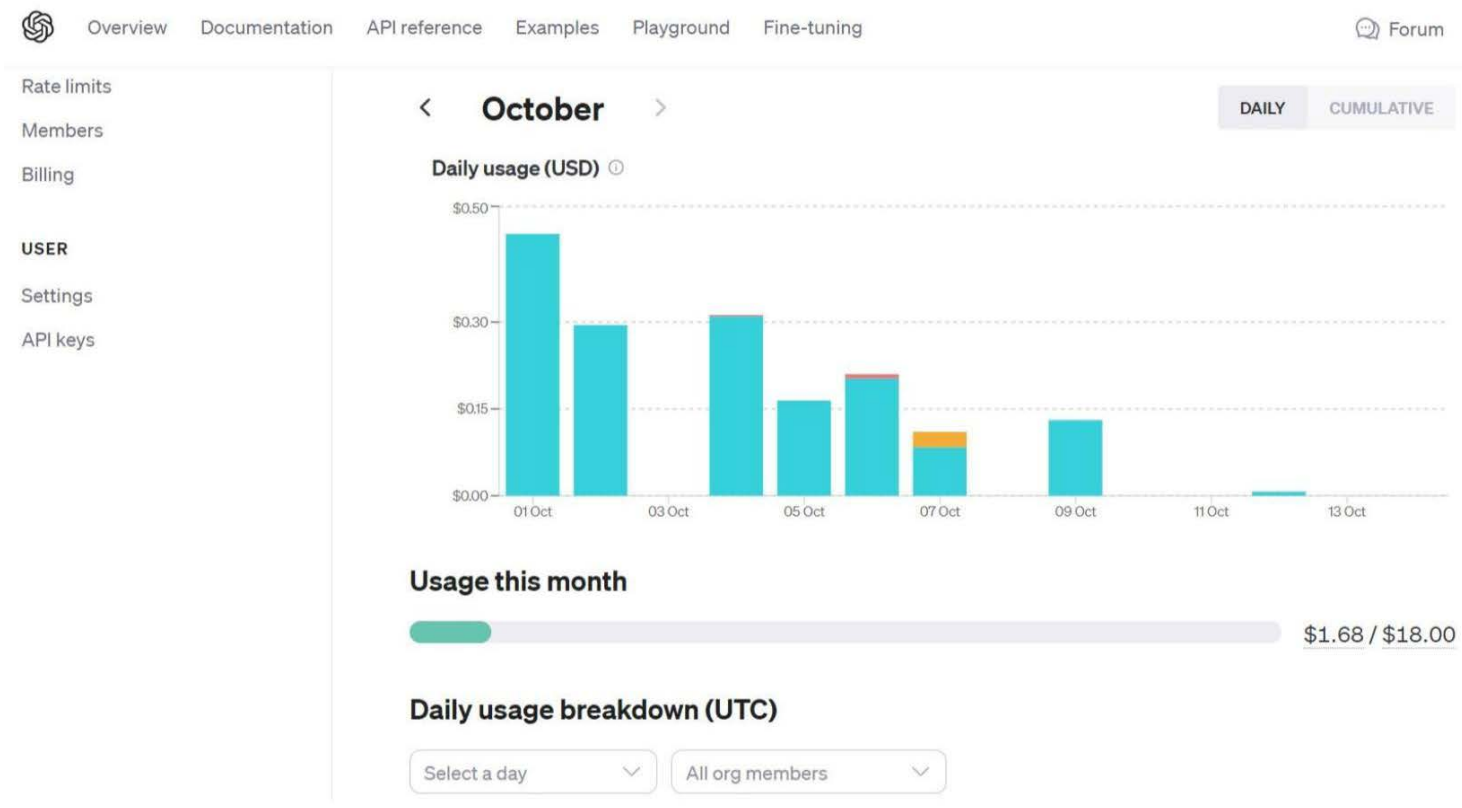
A diferença fundamental é que, como já descrevemos antes, as possibilidades de configuração do comportamento do modelo GPT dessa forma é mínima e se limita ao que podemos incluir no prompt[8].

Existem muitas formas de customizar uma aplicação de IAG, além de simplesmente pedir o que se quer com maior ou menor detalhamento no *prompt.* Ainda dentro da própria interface do ChatGPT, é possível, clicando no nome do usuário, acessar as instruções customizadas (*custom instructions*) que são dois campos adicionais em que é possível especificar coisas do tipo a sua localização, atividade profissional, áreas de interesse ou ainda o tom desejado para a resposta (mais ou menos informal, por exemplo).

7 API é a sigla em inglês para *Application Programming Interface*, ou interface de programação de aplicações. As APIs são conjuntos de ferramentas, definições e protocolos para a criação de aplicações de software, constituindo uma forma de comunicação entre sistemas. Elas permitem a integração entre dois sistemas, em que um deles fornece informações e serviços que podem ser utilizados pelo outro. O modelo de utilização via API é bastante comum em plataformas que oferecem serviços via web, como a OpenAI que disponibiliza vários de seus modelos de inteligência artificial dessa forma.

8 O resultado nesse processo pode ser aprimorado também quando utilizamos o que se convencionou chamar de técnicas de *prompt design*, isto é, procedimentos para garantir mais assertividade e diferenciação nas respostas que o modelo GPT nos retorna. Isso, ainda que de forma muito genérica, também poderia ser considerado como uma estratégia de customização.

Figura 8: Tela com o gráfico de consumo de serviços dos modelos de inteligência artificial através da API do desenvolvedor



Fonte: Elaborado pelo autor.

Para os usuários que são assinantes, isto é, que pagam mensalmente pelo uso e tem acesso, atualmente à versão do GPT 4, uma nova forma de customização foi disponibilizada no final do segundo semestre, permitindo que o usuário crie um *bot* para realizar funções específicas, que normalmente ele precisa usar com mais frequência. Essa funcionalidade chamada de GPTs é uma das estratégias para reforçar o ecossistema que a OpenAI está criando.

O mais importante é que o desenvolvimento é feito com a ajuda de um dos módulos do ChatGPT, chamado de *Code Interpreter*, que permite que um usuário totalmente leigo na área de programação consiga criar uma solução personalizada, sem ter que escrever uma linha de código sequer. Outros dois módulos, o *Web Browsing* e o modelo de geração de imagens da OpenAI, o DALL-E, também podem ser acionados para oferecer mais funcionalidades ao GPT que está sendo criado.

Nas imagens abaixo estão dois exemplos de desenvolvimento de GPTs. O primeiro dedicado a analisar ideias e fazer planejamento estratégico para determinado negócio que o usuário indica no *prompt* junto com as metas que deseja alcançar[9].

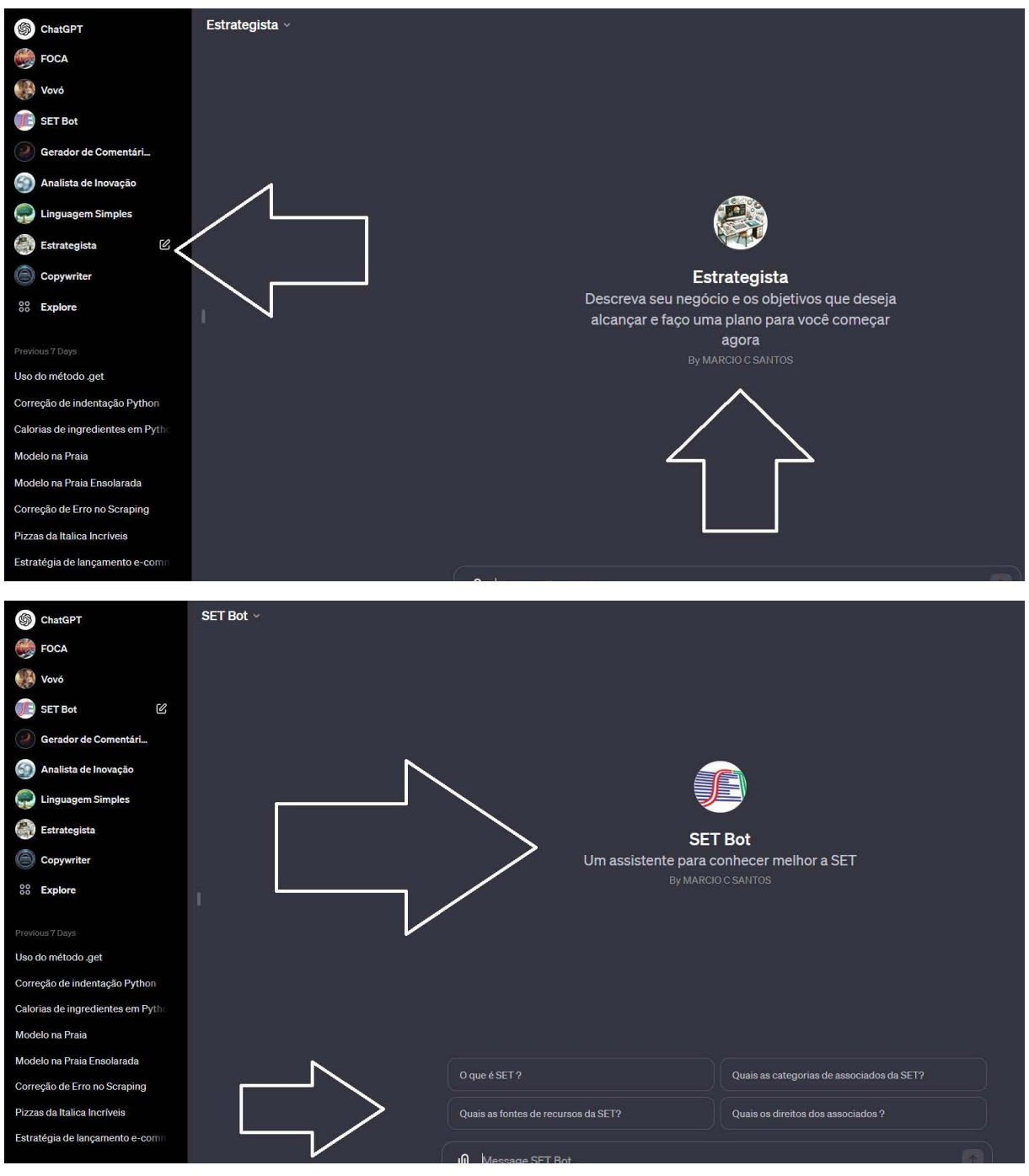
Já na outra imagem, está um GPT que resolve o problema que descrevemos antes nos testes com a SET, porque permite que o modelo, ao invés de olhar para sua enorme base de treinamento, para lidar com dados, utilize uma outra base, que é fornecida pelo usuário, ideal para a geração de respostas precisas.

A inteligência artificial customizada (*Custom AI*) conceito que ora propomos, e que entendemos ser a mais indicada para uso organizacional, em teoria pode ser definida como a estratégia de utilização da IAG quando operada pelo processamento na nuvem de servidores da empresa, com maior controle dos processos e parâmetros de execução das tarefas, através

9 Para saber mais sobre a funcionalidade dos GPTs, também conhecidos como *personal bots*, ver https://youtu.be/CD5RDPDOzPc?si=oG3MJT6DuOzsZLZi. Acesso em: 18 abr. 2024.

das possibilidades oferecidas pela API dos modelos que o desenvolvedor oferece[10] ou de soluções de terceiros[11], as quais potencializam o uso profissional da aplicação criada.

Figura 9: Exemplos de GPTs customizados que executam funções específicas definidas pelo usuário



Fonte: Elaborado pelo autor.

No caso da OpenAI, além do modelo de geração de texto, o próprio GPT em suas diferentes versões, há também outros modelos que podem ser usados, da mesma forma ou em conjunto, o que pode enriquecer ainda mais as soluções. Assim como o DALL-E que gera imagens, há também o modelo GPTVision, que reconhece elementos a partir de imagens a ele enviadas, bem como o Whisper, que converte áudio em texto, sendo muito útil para processos de transcrição.

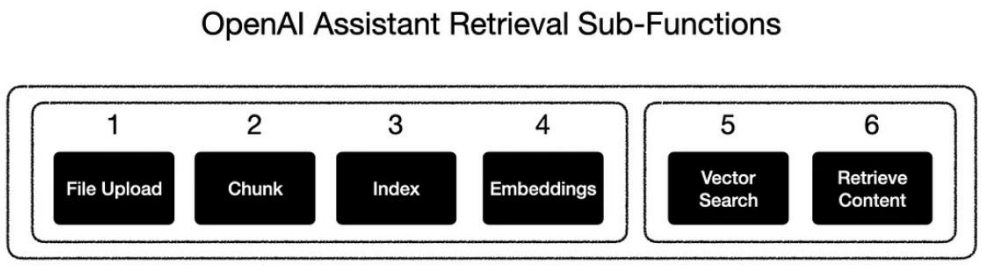
10 É importante entender que cada desenvolvedor de IAG pode oferecer os serviços de seus modelos através de sua nuvem de servidores e acesso via API, normalmente cobrando por esse consumo de alguma forma. No caso da maioria dos modelos da OpenAI, esse cobrança se dá em *tokens,* que são unidades, pedaços de texto que o modelo utiliza para operar e que, na língua inglesa, mantém uma relação com a unidade das palavras em torno de três para quatro, ou seja, três palavras contam em média quatro tokens.

11 Um exemplo é a solução LangChain, disponível em: https://python.langchain.com/docs/get_started/introduction. Acesso em: 18 abr. 2024.

Um maior nível de customização e controle é possível para as organizações quando desenvolvem suas próprias aplicações usando os modelos de inteligência artificial generativa fora da interface de *bot* original do ChatGPT.

Para tanto, uma das possibilidades é a utilização das soluções identificadas como RAG, sigla de *Retrieval Augmented Generation*. Apesar do detalhamento da operação de uma solução RAG estar além do escopo deste texto, podemos resumir sua vantagem justamente no sentido de superar as restrições que apontamos antes sobre a base de dados original de treinamento dos modelos LLMs, sobre os quais ninguém, além do próprio desenvolvedor, tem controle.

Figura 10: Etapas de processamento de documentos específicos para utilização como base de dados por um modelo LLM a partir da estratégia RAG



Fonte: Greyling (2023).

Usando RAG é possível criar uma nova base de dados complementar que dentro de uma solução customizada será utilizada como fonte, permitindo ao modelo a extração de informações corretas na geração de suas respostas. Tal diferença traz grandes vantagens principalmente em termos de precisão e assertividade, oferecendo para a aplicação de IAG a capacidade de falar sobre determinado assunto de interesse da organização com muito mais propriedade.

Figura 11: Exemplo de aplicação customizada, fora da interface tradicional do ChatGPT, criada para lidar especificamente com situações de crise e, a partir de uma pequena descrição oferecer sugestões de atuação

Fonte: Elaborado pelo autor.

CONSIDERAÇÕES FINAIS

Para as organizações, lidar com erros oriundos das características intrínsecas dos LLMs e suas enormes bases de dados pode implicar um nível de risco não aceitável. A maioria das organizações não precisa de uma enciclopédia digital de bilhões de verbetes e sim de uma base de conhecimento sobre seus próprios produtos, cultura, áreas de atuação, enfim, de informações sobre si mesmas que possam ser úteis para utilização em qualquer processo organizacional o qual se pretenda automatizar usando IA.

A IAG pode ser utilizada em cenários organizacionais, mas, preferencialmente, com um nível maior de controle sobre os resultados, além do que é permitido normalmente dentro da interface comum do seu maior expoente, o ChatGPT da OpenAI.

A proposta da IA customizada aqui estabelecida oferece inúmeras vantagens para a utilização da IAG nos cenários organizacionais e oferece diversos níveis de controle, dentro e fora da interface original do ChatGPT. Soluções baseadas em estratégias de customização como RAG podem conciliar a capacidade indiscutível de geração de textos que modelos como o GPT tem com a necessidade das organizações de lidarem com níveis de precisão mais altos e informações específicas dentro dos seus próprios interesses.

O conhecimento sobre as características fundamentais da operação dos LLMs e suas restrições é a base inicial para qualquer análise de viabilidade em projetos de automação com IAG. Além disso, a necessidade de uma camada final de controle ou curadoria, um layer humano de verificação, também são pré-condições importantes para uma implementação bem sucedida.

Talvez o ChatGPT não seja a solução ideal para as organizações, mas é possível aprender com ele, experimentando, testando extensivamente e identificando limites e cenários de aplicação recomendável.

REFERÊNCIAS

CORRÊA, Elizabeth Saad; SANTOS, Márcio Carneiro. Jornalismo, inteligência artificial e desinformação: avaliação preliminar do potencial de utilização de ferramentas de geração de linguagem natural, a partir do modelo GPT, para difusão de notícias falsas. *Estudios sobre el mensaje periodístico*, Madrid, n. 29, p. 783-794, 2023. doi: https://doi.org/10.5209/esmp.87965

GUZMAN, Andrea L; SETH C. Lewis. What is human-machine communication, anyway. In: GUZMAN, Andrea L. (ed.). *Human-machine communication: Rethinking communication, technology, and ourselves*. New York: Sage, 2018. p. 1-28.

JONES, Steven. *Against technology:* From the Luddites to neo-Luddism. New York: Routledge, 2013.

LATZER, Michael *et al.* The economics of algorithmic selection on the Internet. In: LATZER, Michael (ed.). *Handbook on the Economics of the Internet*. Cheltenham: Edward Elgar Publishing, 2016. p. 395-425.

LEMOS, André. Epistemologia da comunicação, neomaterialismo e cultura digital. *Galáxia*, São Paulo, p. 54-66, 2020.

MANOVICH, Lev. *El software toma el mando*. Barcelona: Editorial UOC, 2014.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. *Big data*: A revolution that will transform how we live, work, and think. New York: Houghton Mifflin Harcourt, 2013.

GARTNER HYPE CYCLE. Gartner. 2023. Disponível em: https://www.gartner.com/en/research/methodologies/gartner-hype-cycle. Acesso em: 2 jan. 2024.

GREYLING, Cobus. Knowledge Retrieval Via The OpenAI Playground. *Medium*, 8 nov. 2023. Disponível em: https://cobusgreyling.medium.com/knowledge-retrieval-via-the-openai-playground-8b04682ebe37. Acesso em: 2 jan. 2023.

LIMA JR, Walter Teixeira; VERGILI, Rafael. Digital Inclusion and Computational Thinking: New Challenges and Opportunities for Media Professionals. In: LIMA JR, Walter Teixeira; VERGILI, Rafael. (org.). *Information and Technology Literacy: Concepts, Methodologies, Tools, and Applications*. Hershey: IRMA, 2018. p. 759-773.

OPENAI. *ChatGPT*, 2023. Disponível em: https://chat.openai.com. Acesso em: 2 jan. 2023.

REITER, Ehud; SRIPADA, Somayajulu; ROBERTSON, Roma. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research*, v. 18, p. 491-516, 2003.

SANTOS, Márcio Carneiro. Internet das Coisas e sistemas inteligentes no jornalismo: o conceito de presença diluído entre as narrativas da complexidade urbana. *Comunicação & Inovação*, São Caetano do Sul, v. 17, n. 34, p. 21-39, 2016. doi: https://doi.org/10.13037/ci.vol17n34.3769

SANTOS, Márcio Carneiro. Inteligência híbrida e análise de sentimentos: integrando curadoria humana e coleta de dados automatizada para avaliar a comunicação de governo. *Conexão-Comunicação e Cultura*, Caxias do Sul, v. 17, n. 33, p.105-121, 2018.

SANTOS, Márcio Carneiro. A datificação de um campo de conhecimento: como algoritmos, números e abordagens quantitativas estão mudando a comunicação. *Organicom*, São Paulo, v. 16, n. 31, p. 145-157, 2019.

SANTOS, Márcio Carneiro. ENTREVISTANDO UM ROBÔ: notas sobre a aplicação experimental da metodologia EEAF usando a ferramenta ChatGPT de inteligência artificial.. *Comunicação & Inovação*, São Caetano do Sul, v. 24, p.1-17, 2023.

SQUIRRA, Sebastião Carlos. A tecnologia e a evolução podem levar a comunicação para a esfera das mentes. *Revista Famecos*, Porto Alegre, v. 23, n. 1, p.1-18, 2016a.

SQUIRRA, Sebastião Carlos. A informação essencial à vida, às máquinas e à comunicação. *Lumina*, Juiz de Fora, v. 10, n. 2, p.1-19, 2016b.

VEEL, Kristin. Make data sing: The automation of storytelling. *Big Data & Society*, Thousand Oaks, v.5, n.1, p.1-8, 2018. doi: https://doi.org/10.1177/2053951718756686.

SWANSBURG, Justin. How to Use LLMs to Build Better Clustering Models. *Medium*, 13 abr. 2023. Disponível em: https://medium.com/@swansburg.justin/how-to-use-llms-to-build-better-clustering-models-9b17a5491bb4 . Acesso em: 2 jan. 2024.

YANG, Jingfeng et al. *Harnessing the power of llms in practice*: A survey on chatgpt and beyond. New York: ACM, 2023.

Artigo recebido em 03.01.2024 e aprovado em 23.02.2024