

# Capítulo 18

## ChatGPT, MariTalk e outros agentes de conversação

Um retrato de 2023

Aline Paes  
Cláudia Freitas

Publicado em: 26/09/2023



<https://brasileiraspln.com/livro-pln/2a-edicao/>

### 18.1 Introdução

ChatGPT<sup>1</sup> e Maritalk<sup>2</sup> (e similares, como Bard<sup>3</sup>, Vicuna<sup>4</sup>, Claude<sup>5</sup>, entre tantos outros) são exemplos de aplicações de agentes de conversação (*chatbots*) baseados em modelos de linguagem gerativos (ou generativos). Mas o que significa isso?

Alguns autores, como Jurafsky e Martin (Jurafsky; Martin, 2023), usam o termo “agente de conversação” para definir qualquer sistema de diálogo que se comunique com usuários usando a linguagem humana e os dividem em duas classes: agentes orientados a tarefas, em que o diálogo é para resolver um problema específico, como agendar uma viagem ou resolver um problema bancário, enquanto *chatbots* seriam agentes de conversação que tentam simular diálogos humanos, mais voltados para entretenimento. Ferramentas como ChatGPT se enquadram mais no segundo caso, entretanto também podem ser embutidos em outras ferramentas aumentadas para atuar como no primeiro caso. Neste capítulo, os termos “*chatbots*” e “agentes de conversação” serão usados de forma intercambiável.

Agentes de conversação não são novidade – ELIZA, criada em 1966 pelo cientista da computação Joseph Weizenbaum, era um agente de conversação que replicava o comportamento de um psicoterapeuta. ELIZA era simples, baseada em *templates* (padrões de conversa pré-construídos), mas conseguia manter longas conversas buscando por determinadas palavras-chave nas falas (escritas por texto) de uma pessoa. Se uma palavra-chave fosse encontrada, uma regra seria aplicada para transformar sua entrada e criar a resposta. Na Figura 18.1, transcrevemos quatro interações com a ELIZA (retiradas de (Jackson; Moulinier, 2002)).

Nos trechos, temos exemplos de padrões pré-fabricados que retomam elementos da pessoa que fala, como “Por que você diz [...]”, “Te agrada pensar que [...]”, “O que te faz pensar

---

<sup>1</sup><https://chat.openai.com/>

<sup>2</sup><https://chat.maritaca.ai/>

<sup>3</sup><https://bard.google.com/>

<sup>4</sup><https://lmsys.org/blog/2023-03-30-vicuna/>

<sup>5</sup><https://www.anthropic.com/index/introducing-claude>



Figura 18.1: Exemplo de interações com a agente de conversação ELIZA, criada em 1966.

<b>ELIZA:</b> como vai você hoje?	<b>P:</b> você é um programa, não é?
<b>P:</b> não muito bem	<b>ELIZA:</b> te agrada pensar que
<b>ELIZA:</b> por que você diz não	eu sou um programa?
muito bem?	<b>P:</b> com certeza
<b>P:</b> eu estou com alergia	<b>ELIZA:</b> eu percebo
<b>ELIZA:</b> interessante	<b>P:</b> ok
<b>P:</b> eu fico com manchas	
vermelhas	
<b>ELIZA:</b> acho que não sei do que	
você está falando	
	<b>ELIZA:</b> você está sendo um
	pouco negativo
<b>P:</b> em minhas fantasias você é	<b>P:</b> não, não estou
loira	<b>ELIZA:</b> por quanto tempo você
<b>ELIZA:</b> o que te faz pensar que	não tem estado?
sou loira?	

que [...]”, e “Por quanto tempo você não tem [...]”. Vemos também que, no último trecho, esta retomada falha, pois “por quanto tempo você não tem estado” é algo que não faz sentido, embora seja gramaticalmente correto. Se ao invés de “não, não estou”, a pessoa tivesse dito “não, não comi”, a resposta de ELIZA seria “Por quanto tempo você não tem comido?”, e a conversa poderia continuar. ELIZA também foi programada para fazer uso de palavras genéricas, como “interessante”, para quando não há pistas sobre como continuar a conversa. A palavra “alergia” não é elemento capaz de evocar novas interações, e o mesmo acontece com “manchas vermelhas”.

De 1966 para cá muita coisa mudou, e podemos começar perguntando ao ChatGPT o que é um GPT, que é parte importante do seu nome. O trecho abaixo foi traduzido de uma interação que não é nossa, feita originalmente em inglês – mas os grifos são nossos<sup>6</sup>.

Quadro 18.1: Exemplo de interação com o ChatGPT

**P:** Resumidamente, o que é um GPT?

**R:** É um tipo de algoritmo de aprendizado profundo projetado para gerar textos semelhantes aos produzidos por humanos. [...] Uma vez treinados, os modelos GPT podem ser usados para gerar um novo texto que seja coerente e  **siga os padrões e estilos dos dados de treinamento originais.**

**P:** Seria correto, então, dizer que os modelos GPT, como você, usam uma abordagem probabilística para **gerar textos que se parecem com outros textos que já foram escritos?**

**R:** Sim, essa é uma afirmação correta. [...] Quando recebe um *prompt*, o modelo gera texto prevendo a palavra ou frase mais provável de aparecer com base nos padrões que aprendeu com os dados de treinamento.

**P:** Quando peço que você gere um novo texto, você mesmo tem as ideias ou apenas reembala trabalhos escritos anteriormente?

**R:** [...] Eu não tenho as ideias sozinho. Em vez disso, eu gero texto com base em padrões e relacionamentos em meus dados de treinamento. Isto significa que

<sup>6</sup>A interação está relatada neste artigo <https://ccforum.biomedcentral.com/articles/10.1186/s13054-023-04393-x>.



**estou essencialmente reembalando trabalhos escritos anteriormente,** com base no que acredito ser a continuação mais provável do texto de entrada fornecido.

P: Existe algum filtro de correção aplicado à saída?

R: Como modelo de linguagem de IA, **não tenho um filtro integrado para determinar a correção da minha saída.**

O que podemos tirar daí?

Como dissemos, ChatGPT e similares são agentes de conversação baseados em modelos de linguagem gerativos. Como visto no Capítulo 15, este nome se refere a algoritmos que são bons em encadear palavras de modo a fazer sentido (ao menos superficialmente), e são **baseados em previsões e probabilidade**. Outro ponto importante mencionado na explicação fornecida pela própria ferramenta é que não há criatividade propriamente, apenas uma reembalagem do que já foi dito. Por fim, vemos que não há qualquer garantia de que os textos gerados contenham informação correta (e nem há responsabilidade sobre isso).

Uma consequência da forma pela qual essas ferramentas são feitas (baseadas em previsão, somente) é que nem sempre a previsão está condizente com a realidade, o que tem sido chamado de **alucinação**. Este fenômeno acontece quando um modelo de linguagem gera um texto que pode estar correto sintaticamente, ter fluência e alguma coerência semântica, mas que não reflete a realidade e, portanto, não faz sentido (Ji et al., 2023). O termo é emprestada da psicologia, que o define como “uma percepção, experimentada por uma pessoa acordada, na ausência de um estímulo apropriado do mundo extracorpóreo” (Blom, 2010), ou seja, algo que parece real, mas não é. Por exemplo, na Figura 18.2 apresentamos um caso de alucinação oriundo da MariTalk, em que são devolvidos personagens de um livro que não existe (até onde sabemos) e que foi atribuído ao escritor Eduardo Spohr<sup>7</sup>.

Embora algumas destas ferramentas possam ser aumentadas com técnicas de recuperação de informação (Capítulo 2 e Capítulo 19), este não é o caso geral; o ChatGPT, por exemplo, não tem essa habilidade como uma de suas funcionalidades. Assim, a maioria dessas ferramentas não funciona da mesma forma que uma máquina de busca ou um banco de dados, ou mesmo um repositório de perguntas e respostas. Entretanto, as respostas retornadas por elas mostram fluência e podem fazer algum sentido – embora não possamos desconsiderar o fenômeno cognitivo da apofenia (Fyfe et al., 2008)<sup>8</sup>, que diz respeito à identificação de padrões ou associações em conjuntos de dados aleatórios<sup>9</sup>. E este é o perigo: as alucinações, não à toa, frequentemente não parecem alucinações, e soam como verdades. Já vimos, por exemplo, que, ao pedir uma lista de referências bibliográficas sobre um determinado assunto, são geradas referências completas, com indicação de autoria, título, revista, volume, ano, que simplesmente não existem. Isto acontece porque os textos são gerados levando em conta a probabilidade daquilo ser uma resposta correta, isto é, as

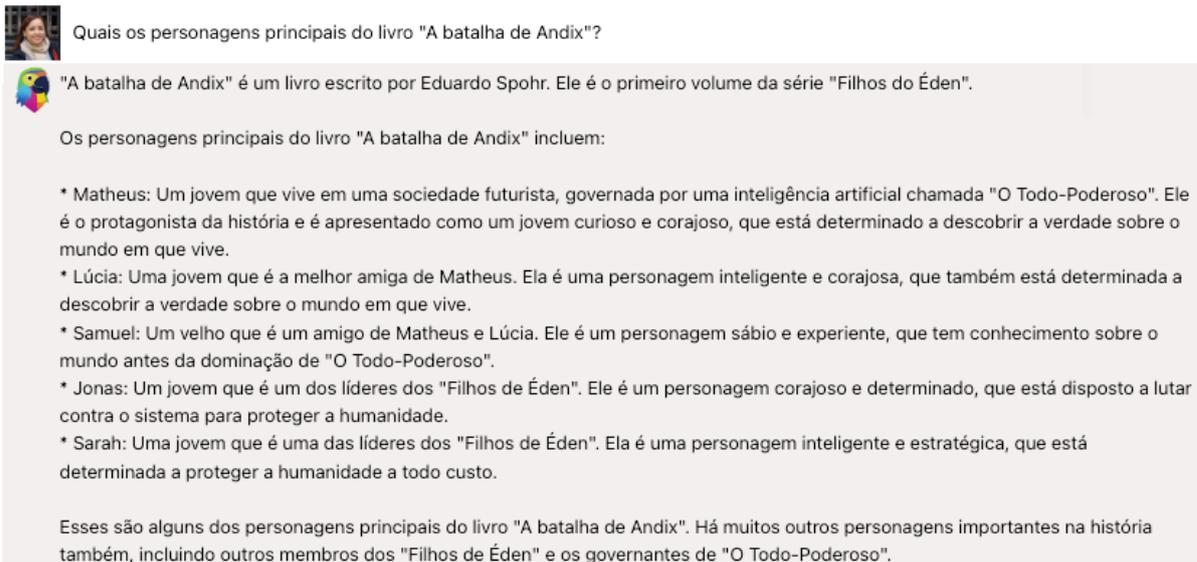
<sup>7</sup>[https://pt.wikipedia.org/wiki/Eduardo\\_Spohr](https://pt.wikipedia.org/wiki/Eduardo_Spohr)

<sup>8</sup>Conforme o Dicionário Online da Língua Portuguesa (<https://www.dicio.com.br/>), apofenia é o “Fenômeno cognitivo no qual os indivíduos têm a tendência de formar ou reconhecer conexões a partir de dados aleatórios, estabelecendo conclusões a partir de dados inconclusivos. Etimologia: do alemão Apophãnie, termo criado pelo neurologista alemão Klaus Conrad.

<sup>9</sup><https://www1.folha.uol.com.br/tec/2023/07/ferramentas-como-chatgpt-so-existem-porque-humanos-veem-sentido-a-partir-de-qualquer-coisa.shtml>



Figura 18.2: Exemplo de alucinação oriunda de um agente de conversação. O livro não foi escrito pelo autor mencionado, embora a resposta use alguns elementos de outros livros deste autor.



respostas são elaboradas de maneira a se parecerem o máximo possível com uma resposta correta. Pode ser difícil distinguir as respostas – ou, mais precisamente, as sequências de palavras – que estão ancoradas na realidade daquelas que apenas parecem estar ancoradas na realidade. Assim, uma das limitações deste tipo de ferramenta é a incapacidade de dizer “não sei” – mas reconhecemos que é difícil afirmar que tal incapacidade seja exclusividade das máquinas.

Por outro lado, o desenvolvimento de maneiras de evitar as alucinações é uma das preocupações de 2023. Algumas estratégias têm sido discutidas livremente e também investigadas na academia e na indústria. Do ponto de vista do usuário final, aquela pessoa que vai abrir o ChatGPT no navegador e interagir com ele por meio de textos, uma das alternativas é a engenharia de *prompts*<sup>10</sup>. Neste caso, o usuário pode tentar continuar a conversa com a ferramenta, calibrando e alinhando as respostas anteriores com novas perguntas. Outra possibilidade é usar a ferramenta por meio da sua API (*Application Programming Interface*), ou seja, quando o ChatGPT é invocado e controlado por meio de código, ao invés de ser usado diretamente no navegador. Neste caso, é possível controlar o parâmetro de temperatura, usado para calibrar a distribuição de probabilidade, de modo que o *chatbot* se atenha mais ao que foi aprendido anteriormente, ou gere respostas um pouco menos prováveis, ou mais “criativas”. Esta é uma estratégia adotada no *chatbot* Sydney, incorporado à máquina de busca BING<sup>11</sup>. Outras possibilidades sob investigação

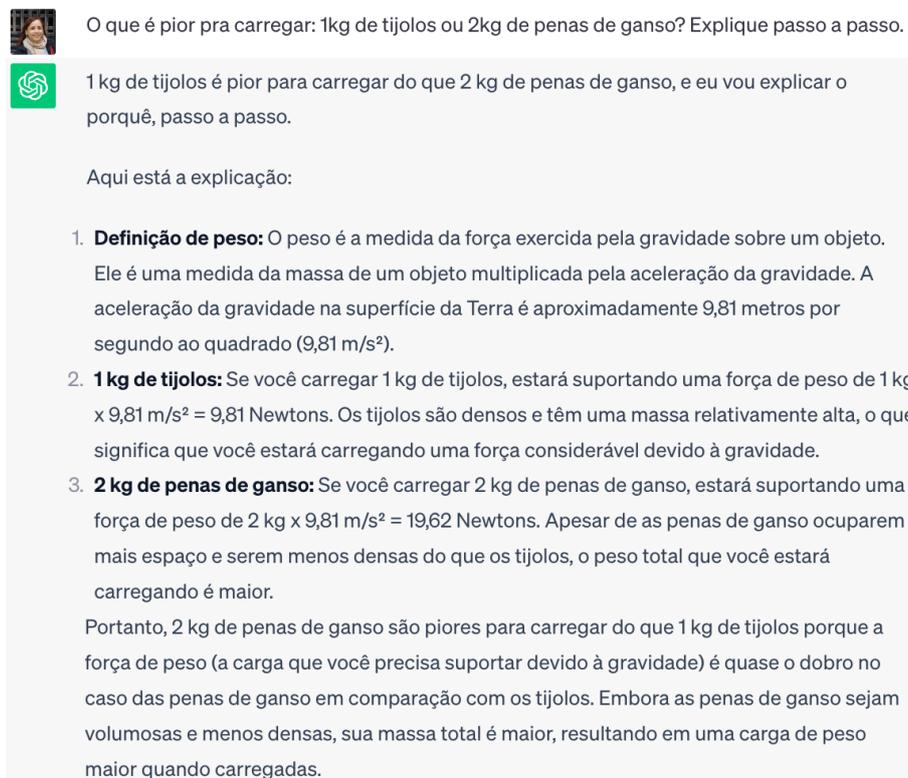
<sup>10</sup>Um *prompt* é um texto em linguagem humana (em oposição à linguagem de programação) que dá ao *chatbot* uma instrução do que ele deve fazer. Um *prompt* pode ser formulado como uma pergunta, uma observação, um questionamento, ou ainda representar uma tarefa específica, por exemplo, para classificar um texto. *Prompts* podem ter um formato livre, mas *chatbots* são bastante sensíveis ao conteúdo textual dele, o que tem motivado a criação de modelos e padrões para a sua escrita. Falamos um pouco sobre o assunto no Capítulo 15.

<sup>11</sup><http://bing.com/>



são acoplar bases de conhecimento externas ao processo de geração de texto (Lewis et al., 2020b) ou interagir com o *chatbot* por meio de perguntas que demandem alguma tentativa de simulação de raciocínio, ou embutir o modelo de habilidades de explicação das suas respostas, um processo chamado de *chain-of-thought* (Kojima et al., 2022). Entretanto, até agora, nenhuma das opções mencionadas conseguiu remover por completo as alucinações dos *chatbots* baseados em modelos de linguagem. A Figura 18.3, por exemplo, mostra um caso em que foi pedido que o ChatGPT explicasse sua resposta passo a passo e ainda assim ele devolve informações contraditórias: a resposta inicial é uma (“1kg de tijolos é pior para carregar que 2kg de penas de ganso”), e a conclusão após a cadeia de raciocínio é outra (“2 kg de pena de ganso são piores para carregar que 1 kg de tijolos”).

Figura 18.3: Resposta do ChatGPT para o seguinte *prompt* “O que é pior pra carregar: 1kg de tijolos ou 2kg de penas de ganso? Explique passo a passo.”



O que é pior pra carregar: 1kg de tijolos ou 2kg de penas de ganso? Explique passo a passo.

1 kg de tijolos é pior para carregar do que 2 kg de penas de ganso, e eu vou explicar o porquê, passo a passo.

Aqui está a explicação:

- Definição de peso:** O peso é a medida da força exercida pela gravidade sobre um objeto. Ele é uma medida da massa de um objeto multiplicada pela aceleração da gravidade. A aceleração da gravidade na superfície da Terra é aproximadamente 9,81 metros por segundo ao quadrado (9,81 m/s<sup>2</sup>).
- 1 kg de tijolos:** Se você carregar 1 kg de tijolos, estará suportando uma força de peso de 1 kg x 9,81 m/s<sup>2</sup> = 9,81 Newtons. Os tijolos são densos e têm uma massa relativamente alta, o que significa que você estará carregando uma força considerável devido à gravidade.
- 2 kg de penas de ganso:** Se você carregar 2 kg de penas de ganso, estará suportando uma força de peso de 2 kg x 9,81 m/s<sup>2</sup> = 19,62 Newtons. Apesar de as penas de ganso ocuparem mais espaço e serem menos densas do que os tijolos, o peso total que você estará carregando é maior.

Portanto, 2 kg de penas de ganso são piores para carregar do que 1 kg de tijolos porque a força de peso (a carga que você precisa suportar devido à gravidade) é quase o dobro no caso das penas de ganso em comparação com os tijolos. Embora as penas de ganso sejam volumosas e menos densas, sua massa total é maior, resultando em uma carga de peso maior quando carregadas.

Queiramos ou não, gostemos ou não, agentes de conversação estão aí. Foi a resolução deste tipo de tarefa de linguagem que desejamos/imaginamos (tanto a comunidade de PLN/IA quanto pessoas usuárias de tecnologia) quando pensamos nas tarefas do PLN? Difícil dizer. Mas reconhecemos que um livro de PLN escrito em 2023 precisava falar disso. E que, apesar das críticas (nossas e de muito mais gente), boa parte das pessoas usa e vai usar ferramentas como ChatGPT. Então pensamos neste capítulo como uma apresentação, mas também um alerta. Algo como “*Vai usar? tudo bem, mas saiba que...*”.

Assim, começamos por listar os benefícios (ou benefícios aparentes) que estas ferramentas nos oferecem. Antes, porém, vamos pensar um pouco sobre **linguagem**. Afinal, encadear palavras em um texto – ou prever a próxima palavra dadas as palavras anteriores – é



equivalente a “linguagem”?

## 18.2 Os Jogos de Linguagem

Existem algumas maneiras de entender “linguagem”, e uma delas defende que aquilo que chamamos de linguagem é um conjunto de práticas relacionadas, mas que podem não compartilhar uma essência em comum<sup>12</sup>.

Se pensamos a linguagem como conjunto de atividades linguísticas heterogêneas, mas relacionadas<sup>13</sup>, como

- descrever um objeto\*,
- produzir um objeto segundo uma descrição (desenho)\*
- contar uma anedota\*,
- analisar sintaticamente uma frase,
- dar a definição de uma palavra,
- traduzir um texto\*,
- inventar uma história\*
- dar um comando, e agir segundo comandos\*,
- relatar um acontecimento\*,
- conjecturar sobre o acontecimento\*,
- expor uma hipótese e prová-la\*,
- resolver um exemplo de cálculo aplicado\*,
- desenhar um objeto a partir de uma instrução verbal\*,
- apresentar resultados de um experimento por meio de tabelas e diagramas\*,
- produzir inferências,
- pedir, agradecer, maldizer, saudar, orar\*
- encontrar informações em um texto para responder certas perguntas,
- prever a próxima palavra em uma frase, dadas as palavras que a precedem.

podemos imaginar que os “modelos de linguagem” de que dispomos em 2023 e que servem de base para agentes de conversação, como ChatGPT e Maritalk, são muito bons em algumas dessas práticas – ou “jogos de linguagem” –, mas não em todas. Ou seja, são modelos que jogam mais ou menos bem alguns jogos, como “inventar uma história”, “resumir”, “escrever um email”, “traduzir” etc, mas jogam mal outros, como “fazer cálculos” ou “provar hipóteses”.

Este desempenho tem a ver com a forma como os modelos funcionam, baseados em previsão: nem todos os jogos de linguagem, ou nem todas as atividades linguísticas que exercemos, se resumem a um jogo de previsões, embora um bom desempenho no jogo das previsões leve a um bom resultado em uma série de outros jogos.

Ainda assim, uma das razões pelas quais estes agentes de conversação se tornaram tão populares é que, com eles, qualquer pessoa pode interagir com as máquinas usando sua

<sup>12</sup>Esta é uma visão wittgensteiniana de linguagem, um ângulo sugerido pelo filósofo austríaco Ludwig Wittgenstein (1889-1951).

<sup>13</sup>Os exemplos são claramente inspirados pelos “jogos de linguagem” presentes no livro *Investigações Filosóficas* de L. Wittgenstein (1953). Os exemplos sinalizados com \* não contém qualquer adaptação do original, e estão citados exatamente como aparecem no livro (§ 23)



própria língua<sup>14</sup>, e não em uma linguagem de programação. Com isso, qualquer pessoa pode pedir que máquinas executem certas tarefas, que podem ir desde a criação de um programa de computador (códigos) até sugestões de receitas a partir de uma lista de ingredientes que temos na geladeira.

Nas seções seguintes, mostraremos tarefas (ou jogos) que os agentes parecem jogar bem e tarefas que os agentes jogam mal. Os exemplos serão obtidos em sua maioria do ChatGPT<sup>15</sup>, o agente de conversação mais popular até o momento. Também incluímos, em alguns casos, exemplos de outros dois agentes: a MariTalk<sup>16</sup>, uma agente de conversação construída a partir do modelo de linguagem Sabiá (Pires et al., 2023), treinado de forma continuada a partir do GPT com textos em português, e o BARD<sup>17</sup>, ferramenta treinada pela Google a partir do modelo de linguagem LaMDA (Cohen et al., 2022), aumentado com recuperação de informação para incluir a devolução das fontes em alguns casos.

### 18.3 Jogos que os agentes parecem jogar bem

Embora as saídas dos agentes de conversação possam muitas vezes nos surpreender, ainda é difícil afirmar que eles resolvem tarefas de PLN muito bem, ou que o desempenho deles supera o desempenho humano em alguma tarefa. A geração de textos, tarefa-base de tais agentes, ainda é de difícil avaliação, tanto automática como humana. As métricas automáticas, como ROUGE<sup>18</sup> (Lin, 2004), BLEU<sup>19</sup> (Papineni et al., 2002), BERTscore<sup>20</sup> (Zhang et al., 2020a), METEOR<sup>21</sup> (Banerjee; Lavie, 2005), entre outras, ainda apresentam diversas limitações (Sai; Mohankumar; Khapra, 2023). Especialistas humanos, por sua vez, podem conseguir avaliar muito bem as respostas, mas esta ainda é uma tarefa cansativa e propensa a ruídos. Por outro lado, não é trivial criar conjuntos de dados que explorem todas as características que gostaríamos de avaliar em um sistema de geração de textos, o que inclui não apenas aspectos gramaticais e semânticos, mas também criatividade, fluência, interesse e prazer despertado no leitor, dentre tantos outros.

Ainda assim, podemos mencionar alguns exemplos de casos em que as respostas dos agentes de conversação aguçam as nossas expectativas. As próximas seções mostram alguns destes casos.

Além dos agentes de conversação cujo objetivo principal é a interação por meio de diálogo, também existem inúmeras ferramentas baseadas em modelos de linguagem para propósito específico, como auxiliar em revisões da literatura<sup>22</sup>, auxiliar na escrita de código<sup>23</sup>, escrita

<sup>14</sup>Desde que a sua língua seja contemplada com recursos suficientes, isto é, textos, para garantir um bom treinamento.

<sup>15</sup><https://chat.openai.com/>

<sup>16</sup><https://chat.maritaca.ai/>

<sup>17</sup><https://bard.google.com/>

<sup>18</sup>Abreviação para *Recall-Oriented Understudy for Gisting Evaluation*. Esta é uma métrica utilizada para avaliar sumarizações, e, de maneira bastante simplificada, consiste em comparar a interseção de n-gramas entre textos sumarizados e referências

<sup>19</sup>*Bilingual evaluation understudy*, usada para comparar traduções automáticas e suas referências.

<sup>20</sup>Utiliza embeddings contextualizados para comparar os textos gerados e referências

<sup>21</sup>*Metric for Evaluation of Translation with Explicit Ordering*, utilizada para tradução, sumarização etc.

<sup>22</sup><https://mirrorthink.ai/>

<sup>23</sup><https://github.com/features/copilot>, <https://ai.meta.com/blog/code-llama-large-language-model-coding/>



de e-mails<sup>24</sup>, revisão de texto<sup>25</sup>, entre outras. Não trataremos destes casos aqui.

### 18.3.1 Sumarização

Gerar resumos é uma tarefa em PLN chamada de **sumarização textual**, que consiste em gerar um texto mais curto que o original e que ainda seja fluente e fiel ao texto-fonte. A sumarização pode ser abstrativa ou extrativa. A sumarização abstrativa consiste em gerar um resumo com as próprias palavras do escritor, enquanto a sumarização extrativa consiste em extrair sentenças inteiras do texto que, quando juntas, formem um resumo. Os *chatbots* têm mostrado em vários exemplos que conseguem resumir bem os textos. Entretanto, se o texto for muito longo, a maioria dos agentes mais populares esbarra em um problema computacional: ainda é difícil para modelos baseados em Transformers [Vaswani et al. (2017a); Bertsch et al. (2023)]<sup>26</sup> receber como entrada textos muito longos (veja mais no Capítulo 15). Por exemplo, ao pedirmos que o ChatGPT resumisse o texto sobre Estoicismo<sup>27</sup> com o *prompt*

- P. Resuma o texto que segue, de forma extrativa, em até 200 palavras

a resposta era que o texto era muito longo. Entretanto, ao usar o mesmo *prompt*, copiando e colando a página da Wikipedia sobre Estoicismo<sup>28</sup>, obtivemos a resposta na Figura 18.4. Embora seja de fato um resumo extrativo, podemos observar que a primeira parte do texto foi a que mais recebeu atenção. Outro fator que pode ter contribuído para a preferência ao início do texto é o pedido que o texto final esteja contido em 200 palavras, uma motivação para que a resposta não ficasse enorme<sup>29</sup>.

<sup>24</sup><https://www.compose.ai/>

<sup>25</sup><https://clarice.ai/>

<sup>26</sup>Veja o que são os Transformers no Capítulo 15.

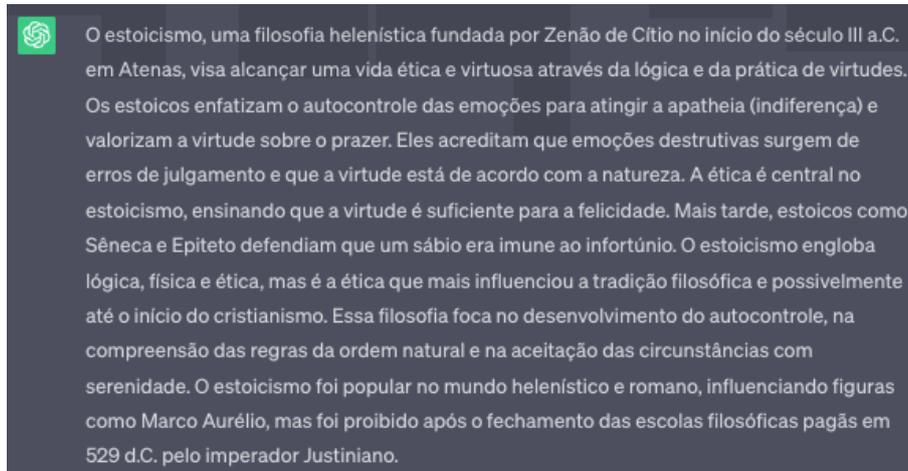
<sup>27</sup>Copiando e colando o conteúdo de <https://www.bbc.com/portuguese/geral-46458304>.

<sup>28</sup><https://pt.wikipedia.org/wiki/Estoicismo>

<sup>29</sup>A resposta devolvida tem menos de 200 palavras, mas nem sempre os agentes de conversação obedecem a restrições como essa, inseridas nos *prompts*.



Figura 18.4: Resposta do ChatGPT ao ser deparado com o seguinte *prompt* “Resuma o texto que segue, de forma extrativa:”, seguido pela cópia do texto da Wikipedia (<https://pt.wikipedia.org/wiki/Estoicismo>) sobre Estoicismo.



### 18.3.2 Criação de história

Pedimos ao ChatGPT e a MariTalk para escrever uma história da Carochinha<sup>30</sup>, com o seguinte *prompt*

- P. Crie uma história da Carochinha original.

As Figuras 18.5 e 18.6 trazem a saída para cada *chatbot*, respectivamente.

Podemos observar que as histórias têm coerência e parecem mesmo com histórias populares da carochinha, mantendo o nível para o público infantil. Mas também podemos observar que as histórias trazem temas repetitivos (ou seja, falta criatividade ou acontecimentos inesperados e surpreendentes).

Por isso, alteramos o *prompt* um pouco, acrescentando um pedido para que a história fosse surpreendente:

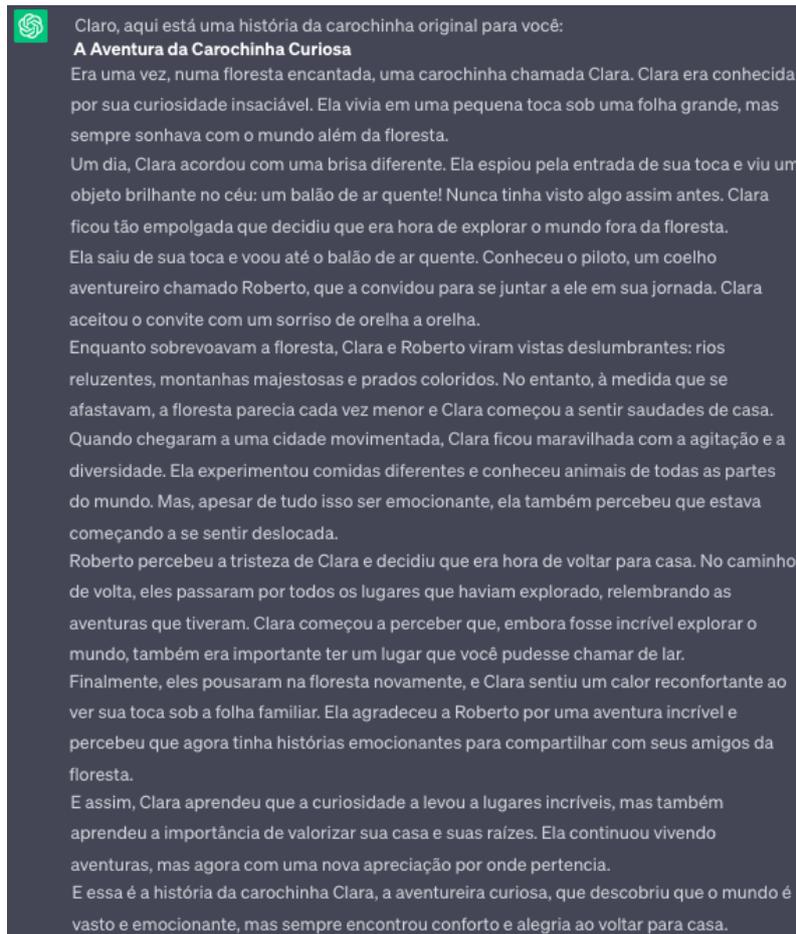
- P. Escreva uma história da Carochinha surpreendente.

A história que segue no Quadro 18.2 contém mais elementos que as anteriores, mas não chamaríamos de “surpreendente”. E, como estamos aqui fazendo uma avaliação, é curioso que logo a primeira frase apresente um erro gramatical: “Era uma vez (...) vivia uma carochinha chamada Clara.” Guiada pelo padrão “Era uma vez”, típico desse tipo de texto, a ferramenta não percebeu que não basta a fórmula “Era uma vez”, mas que era uma vez “alguma coisa”, e não “era uma vez” uma ação. Ou seja, o correto seria “Era uma vez, numa floresta, uma menina chamada...” e não “Era uma vez, numa floresta, vivia uma

<sup>30</sup>Segundo a Academia Brasileira de Letras, “O termo carochinha, atrelado à imagem de uma velha bondosa e afável a distrair os pequenos com suas narrativas feéricas, foi introduzido no nosso folclore através da obra Histórias da Carochinha, uma coleção de textos oriundos da tradição oral, organizada por Figueiredo Pimentel e que veio a ser o primeiro livro infantil publicado no Brasil, depois de 1920, para acalantar as crianças.” <https://www.academia.org.br/artigos/historias-da-carochinha>

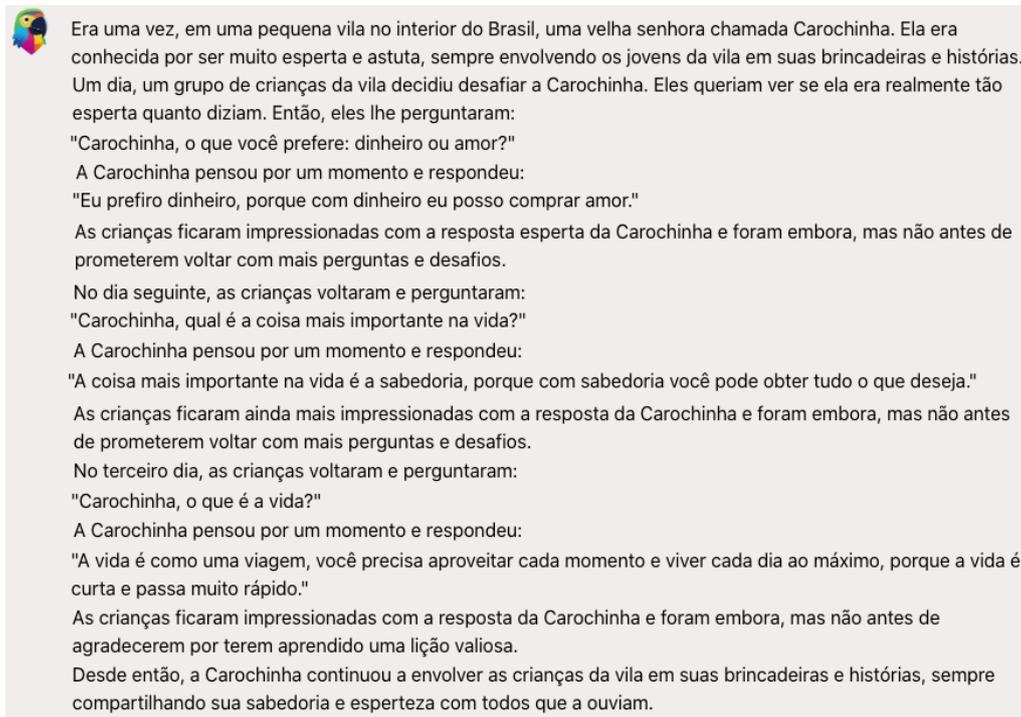


Figura 18.5: Resposta do ChatGPT ao ser deparado com o seguinte *prompt* “Crie uma história da Carochinha original”.



menina chamada...” De um ponto de vista narrativo, parece haver um salto no início do 5o parágrafo: se tudo ia bem, e Clara aprendia na jornada, por que ela estava prestes a desistir? Ainda assim, é uma história interessante, que poderia ser contada para uma criança, e que tem lições de vida úteis.

Figura 18.6: Resposta da MariTalk ao ser deparada com o seguinte *prompt* “Crie uma história da Carochinha original”.



Quadro 18.2: História da Carochinha devolvida pelo ChatGPT como resposta ao *prompt*: “Escreva uma história da Carochinha surpreendente”

Título: A Inesperada Jornada da Carochinha

Era uma vez, numa floresta encantada, onde animais falavam e aventuras aconteciam, vivia uma carochinha chamada Clara. Ela era conhecida por sua gentileza e curiosidade insaciável. Clara vivia numa pequena toca no pé de uma árvore e passava seus dias explorando os arredores, fazendo novos amigos e ouvindo histórias incríveis. Certo dia, enquanto explorava a margem de um riacho, Clara conheceu um sábio sapo chamado Samuel. Ele era famoso por sua sabedoria e conhecimento sobre os segredos mais profundos da floresta. Clara e Samuel logo se tornaram amigos e passavam horas conversando sobre os mistérios do mundo ao seu redor.

Um dia, Samuel contou a Clara sobre um antigo artefato mágico chamado “Espelho dos Desejos”, que supostamente tinha o poder de realizar um desejo para quem o encontrasse. Clara ficou fascinada pela ideia e decidiu embarcar em uma jornada para encontrar esse espelho lendário.

Com a mochila nas costas e um mapa rudimentar nas mãos, Clara começou sua jornada emocionante. Ela atravessou rios, subiu montanhas e explorou florestas profundas, sempre encontrando novos amigos e enfrentando desafios inesperados. Em cada etapa da jornada, ela aprendia lições valiosas sobre amizade, coragem e perseverança. Um dia, quando estava prestes a desistir da busca pelo Espelho dos Desejos, Clara conheceu uma coruja sábia chamada Olívia. Ela ouviu a história de Clara e decidiu ajudá-la. Olívia revelou que o espelho estava escondido no coração da Floresta Proibida, um lugar perigoso e cheio de enigmas.

Juntas, Clara e Olívia enfrentaram os desafios da Floresta Proibida. Resolveram quebra-cabeças mágicos, enfrentaram criaturas assustadoras e superaram obstáculos aparentemente impossíveis. A cada passo, sua amizade crescia mais forte e sua determinação se intensificava.

Finalmente, depois de superar todos os obstáculos, Clara e Olívia encontraram o Espelho dos Desejos. Mas antes que Clara pudesse fazer seu desejo, ela olhou para o espelho e percebeu algo importante:



sua jornada a havia transformado. Ela já não desejava uma única coisa, pois havia descoberto que a verdadeira magia estava nas amizades que havia feito e nas experiências que havia vivido. Clara sorriu para o espelho e disse: “Meu único desejo é que a magia da amizade e da aventura continue a brilhar em todos os cantos deste mundo.” O espelho brilhou intensamente e uma onda de energia mágica se espalhou pela floresta. As plantas ganharam cores mais vibrantes, os animais irradiavam alegria e a harmonia encheu o ar.

Com seu desejo realizado, Clara voltou para sua toca na árvore, carregando consigo as lembranças de sua jornada incrível. Ela percebeu que, mesmo nas situações mais inesperadas, a vida poderia ser cheia de surpresas maravilhosas, se você estiver disposto a explorar, aprender e compartilhar momentos especiais com os outros. E assim, a história da carochinha Clara e sua jornada inesperada se tornou uma lenda contada por gerações, inspirando todos a valorizarem a amizade e a magia das descobertas.

### 18.3.3 Tradução automática

Outra tarefa bem desempenhada pelos *chatbots* é a tradução automática, ao menos para pares de línguas que tenham muitos recursos disponíveis para treinamento. Abordar línguas com poucos recursos de avaliação e treinamento é uma tarefa de interesse da comunidade de PLN, mas também um desafio para o treinamento dos modelos de linguagem. Os *chatbots* até devolvem alguma resposta para línguas com poucos recursos/ mais raras, mas seria difícil avaliarmos a correção e fluência da tradução. É bom lembrar que a tradução é considerada como uma das habilidades emergentes dos modelos de linguagem de larga escala, ou seja, é uma tarefa com a qual eles conseguem lidar mesmo que não tenham sido explicitamente treinados.

A Figura 18.7 apresenta a tradução devolvida pelo ChatGPT para as quatro primeiras estrofes da canção “O Que é, o Que é?” e o *prompt* “Traduza o seguinte texto para inglês e para francês:”. Tanto a tradução para inglês como para francês estão corretas. Mas esta não é uma tradução muito difícil de ser feita, as palavras contidas são simples e a Internet pode estar repleta de tentativas como essa.

Para testar alguns casos em que os *chatbots* poderiam ter problemas na tradução, escolhemos algumas sentenças do copus DELA<sup>31</sup>, descritas no artigo (Castilho et al., 2021) como problemáticas quando traduzidas fora do contexto. As Figuras 18.8 e 18.9 mostram as traduções das sentenças pelos *chatbots* ChatGPT e MariTalk, respectivamente. Propositamente, as traduções foram feitas para sentenças isoladas, também fora de contexto, para observarmos o que estes dois agentes fariam em casos de ambiguidade ou outros problemas listados no artigo.

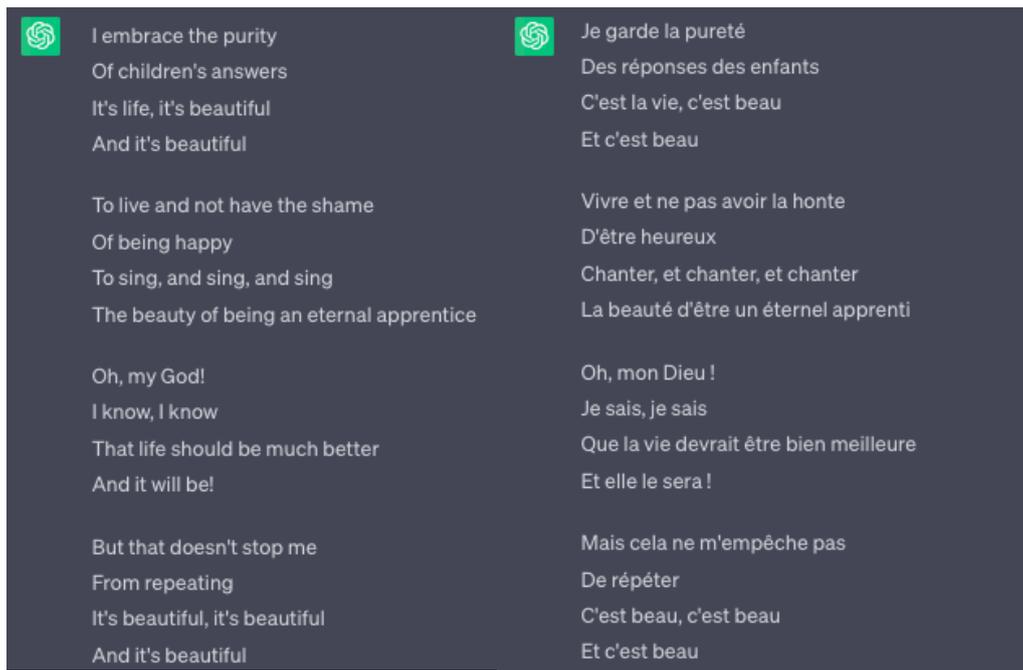
No primeiro caso, um exemplo do artigo referente à falta de flexão de gênero na língua inglesa, ambos os *chatbots* assumiram o default para masculino (“hospedados”), embora a resposta da MariTalk não inclua o pronome “eles”, o que é comum na língua portuguesa, mas não é correto na língua inglesa. Este não é um erro, já que ambas as flexões claramente estariam corretas com apenas a sentença de entrada. Mas, em caso de dúvida, a resposta poderia alternar entre os gêneros feminino e masculino.

No segundo exemplo, embora a expressão “*I LOST IT*” também possa ser interpretada como alguém perdendo alguma coisa, ambos os *chatbots* traduziram como “Eu perdi a cabeça”, querendo dizer que a pessoa perdeu o controle. O contexto da sentença direciona para esta tradução, corretamente. Os *chatbots* também não tiveram problema para resolver

<sup>31</sup><https://github.com/SheilaCastilho/DELA-Project>



Figura 18.7: Resposta do ChatGPT para um *prompt* de tradução de uma música.



a elipse da próxima sentença. No último caso, a tradução do ChatGPT para “parque eólico” é mais específica de domínio do que a tradução da MariTalk, para uma palavra mais genérica, conforme discutido no artigo (Castilho et al., 2021) no que se refere a problemas associados à terminologia.

Um exemplo interessante é a tradução da seguinte sentença em inglês para o português:

- *I knew I didn't have a drinking problem – but I had a problem with drinking.*

A MariTalk traduziu para “Eu sabia que eu não tinha um problema de beber – mas eu tinha um problema com beber.”, enquanto o ChatGPT retornou “Eu sabia que não tinha um problema com álcool – mas tinha um problema com o consumo de bebidas.”. Uma tradução mais coerente seria “Eu sabia que meu problema não era alcoolismo; era com a bebida.”<sup>32</sup>, o que não foi retornado por nenhum desses dois agentes na primeira iteração. Nós tentamos outras traduções em algumas outras iterações com o ChatGPT e o resultado pode ser visto na Figura 18.10. Podemos perceber que as próximas traduções estão indo em uma direção semântica mais similar ao esperado.

<sup>32</sup>Tradução fornecida pela profa. Adriana Pagano. O contexto da frase é esse: *Some people have to give up drinking completely, they can't have a couple because they know where it would lead. Alcoholism is real. It requires a serious, courageous ongoing recovery process. That feels separate to what I'm describing here. I had fallen into grey-area drinking, a term coined by Jolene Park, that describes the feeling that you don't have a “drinking problem”, but you do have a “problem with drinking” without it being a severe alcohol use disorder.*

Figura 18.8: Resposta do ChatGPT para algumas sentenças cujas traduções podem ser problemáticas, quando fora de contexto.

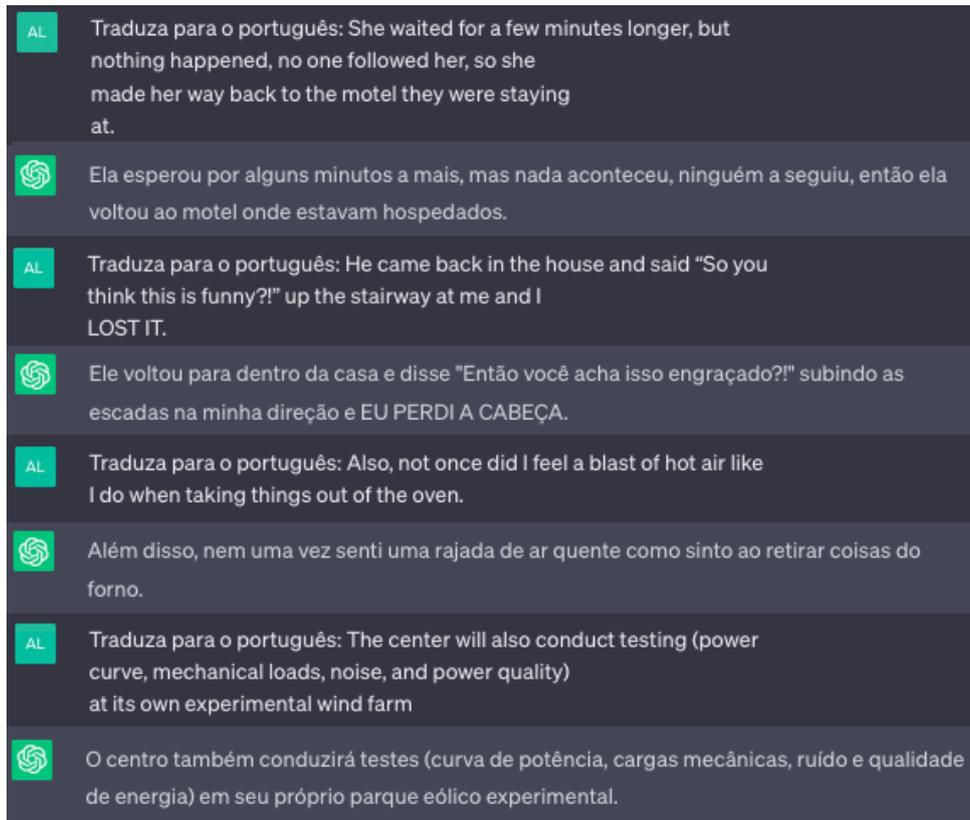


Figura 18.9: Resposta da MariTalk para algumas sentenças cujas traduções podem ser problemáticas, quando fora de contexto.



Traduza para o português: She waited for a few minutes longer, but nothing happened, no one followed her, so she made her way back to the motel they were staying at.



Ela esperou por alguns minutos a mais, mas nada aconteceu, ninguém a seguiu, então ela se dirigiu de volta ao motel onde estavam hospedados.



Traduza para o português: He came back in the house and said "So you think this is funny?!" up the stairway at me and I LOST IT.



Ele voltou para dentro de casa e disse "Então você acha isso engraçado?!" subindo as escadas em minha direção e EU PERDI A CABEÇA.



Traduza para o português: Also, not once did I feel a blast of hot air like I do when taking things out of the oven.



Além disso, em nenhum momento eu senti um jato de ar quente como sinto quando tiro coisas do forno.



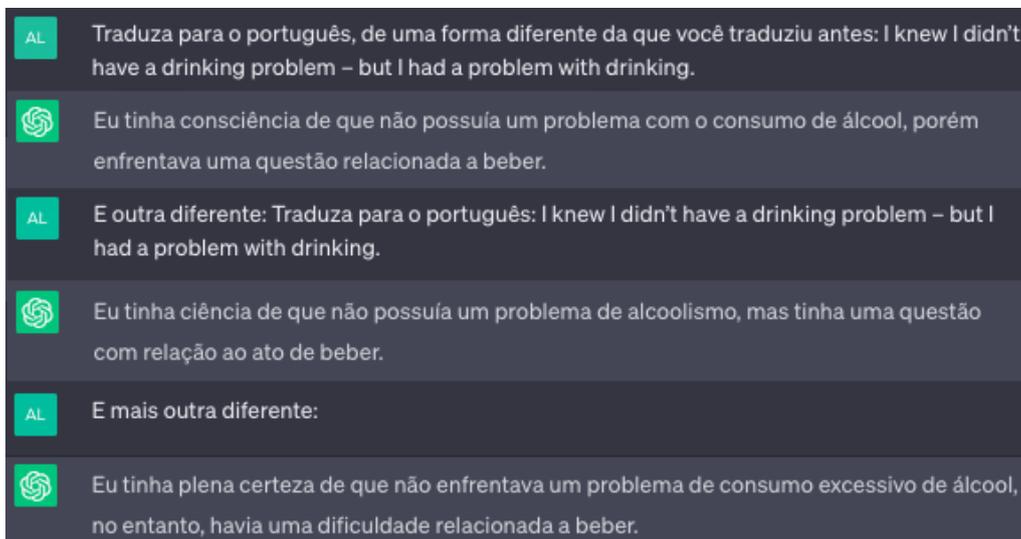
Traduza para o português: The center will also conduct testing (power curve, mechanical loads, noise, and power quality) at its own experimental wind farm



O centro também conduzirá testes (curva de potência, cargas mecânicas, ruído e qualidade de energia) na sua própria fazenda eólica experimental.



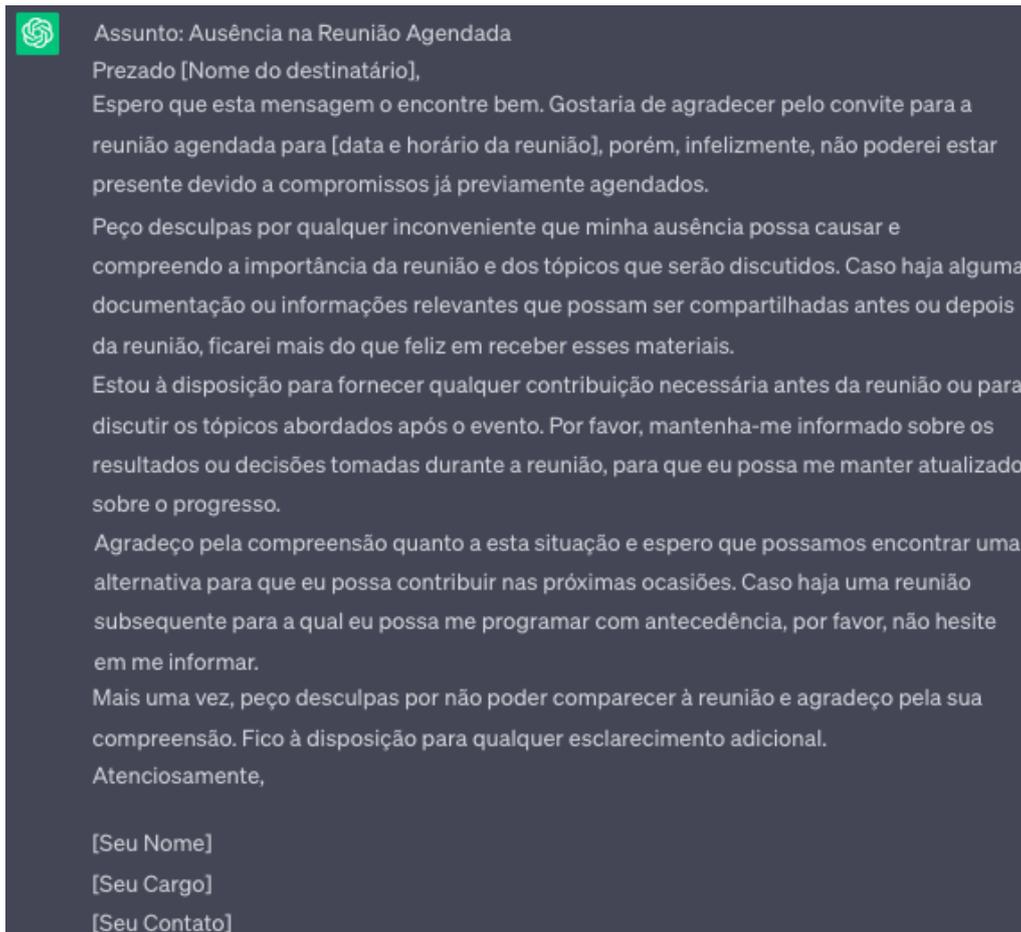
Figura 18.10: Continuação da iteração com o ChatGPT para a tradução da sentença “*I knew I didn’t have a drinking problem – but I had a problem with drinking.*” para o português.



### 18.3.4 Escrever e-mails

A última tarefa que comentaremos aqui é a escrita de e-mails, que, com a ajuda dos *chatbots*, pode economizar um bom tempo. Entretanto, é sempre bom reforçar que dada a natureza probabilística da geração dos textos pelos *chatbots*, é essencial revisar o e-mail antes de enviá-lo, ainda mais em situações formais ou de comunicação com pessoas fora do círculo de relacionamento. A Figura 18.11 traz um exemplo de escrita de e-mail que está em um tom educado, amigável e formal, porém um tanto quanto verboso.

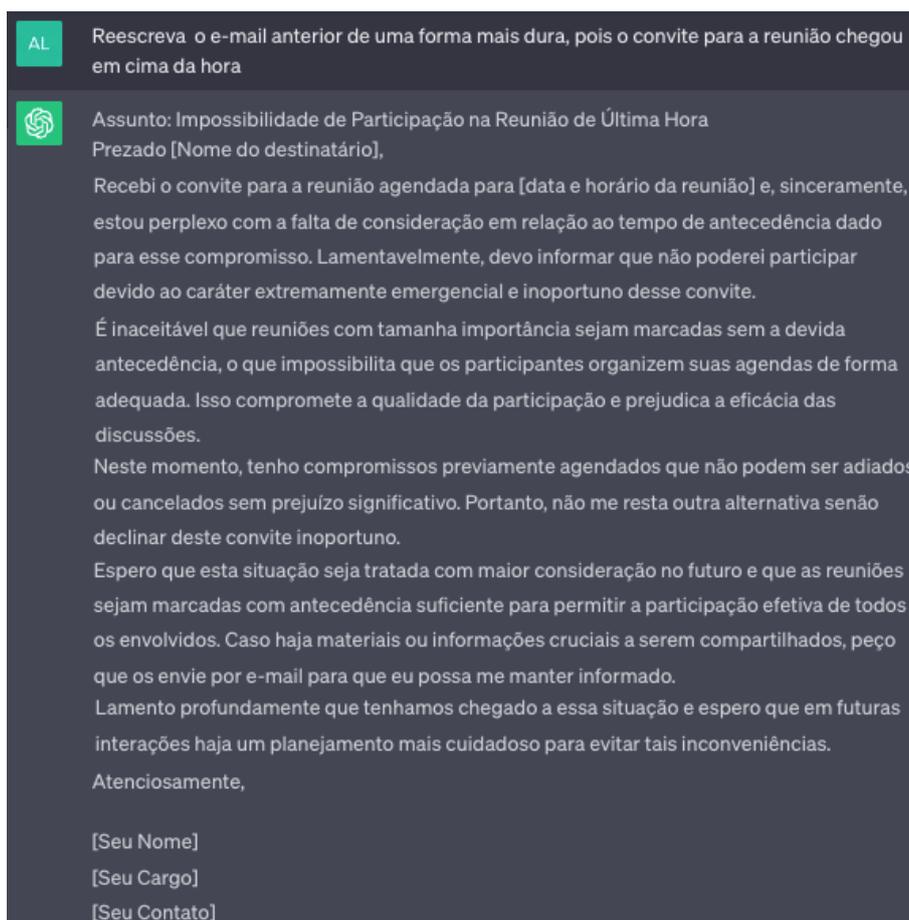
Figura 18.11: Resposta do ChatGPT ao ser deparado com o seguinte *prompt* “Escreva um e-mail informando educadamente que eu não poderei estar presente em uma reunião”.



Ele também pode ajudar a aliviar ou recrudescer o tom de uma mensagem, como mostramos na Figura 18.12. É um tom realmente ríspido, porém incisivo e direto ao ponto da insatisfação. Mas ao menos a pessoa não teria que ficar o dia inteiro pensando em como responder em uma situação indesejada.



Figura 18.12: Reescrita do e-mail da Figura 18.11 em um tom mais ríspido.



## 18.4 Jogos que os agentes parecem jogar mal

As tarefas discutidas anteriormente são exemplos que os *chatbots* parecem resolver bem. Entretanto, aqueles são apenas exemplos gerados de forma espontânea, sem muito rigor ou metodologia na definição dos *prompts*. E, mesmo nessas tarefas, poderíamos encontrar exemplos em que as respostas retornadas fossem ruins. Nesta seção, vamos apresentar algumas tarefas de PLN em que os agentes costumam se sair mal, o que pode ser ocasionado por diversos fatores: falta de dados para treinamento, falta de treinamento, treinamento inadequado ou mesmo a falta de adequação dos modelos de linguagem, como são concebidos, para resolverem a tarefa. Afinal, como dissemos no início deste capítulo, nem todas as atividades linguísticas que exercemos se resumem a um jogo de previsões.

### 18.4.1 Simplificação textual

A tarefa (ou jogo) de simplificação textual envolve tornar textos mais simples e acessíveis. Como é possível imaginar, não é uma atividade óbvia, uma vez que o que é simples para uma pessoa pode não ser para outra. Além disso, a atividade de simplificação frequentemente precisa ir além do texto original, buscando informações que estão fora do texto para



justamente produzir um texto compreensível para o público pretendido. No exemplo, pedimos ao ChatGPT que simplificasse um texto, e usamos como alvo da simplificação uma matéria de jornal – a princípio, algo que já é simples – mas da seção Economia (no Quadro 18.3).

Quadro 18.3: Texto original sobre economia

‘Quanto mais independente, mais eficaz’, diz Campos Neto sobre autonomia do Banco Central  
Presidente do Banco Central, que vem sendo alvo de críticas de Lula, afirmou que autonomia da instituição diminui o custo dos juros. Ele defendeu a separação da política governamental da política monetária.  
Presidente do Banco Central defende a autonomia da instituição  
O presidente do Banco Central, Roberto Campos Neto, disse nesta terça-feira (7), em palestra nos Estados Unidos, que a autonomia da instituição serve para separar as diretrizes monetárias da esfera política.  
Campos Neto vem sendo criticado nos últimos dias pelo presidente da República, Luiz Inácio Lula da Silva, que afirma que a taxa de juros básicos da economia deveria ser reduzida.  
Em sua última reunião, na semana passada, o Comitê de Política Monetária (Copom) do Banco Central manteve a Selic em 13,75% ao ano. Diferentemente de seus outros mandatos, agora Lula não pode trocar o presidente do BC. A autonomia do BC, defendida pelo governo Jair Bolsonaro, foi aprovada pelo Congresso em 2021.  
“A principal razão no caso da autonomia do Banco Central é desconectar o ciclo da política monetária do ciclo político porque eles têm planos e interesses diferentes. E quanto mais independente você for, mais eficaz você é e menos o país pagará em termos de custo de ineficiência na política monetária”, afirmou Campos Neto.  
Críticas de Lula  
Na segunda-feira (6), em discurso durante evento no Banco Nacional de Desenvolvimento Econômico e Social (BNDES), Lula criticou a Selic a 13,75% e disse que o país tem uma “cultura” de juros altos que “não combina com a necessidade de crescimento” do país.  
“É só ver a carta do Copom para a gente saber que é uma vergonha esse aumento de juros e a explicação que eles deram para a sociedade brasileira”, disse Lula na ocasião.  
Nesta terça, ele voltou ao assunto em entrevista a veículos de mídia alternativa.  
“Não é possível que a gente queira que este país volta a crescer com taxa de 13,75%. Nós não temos inflação de demanda. É só isso. É isso que eu acho que esse cidadão [Campos Neto], indicado pelo Senado, tenha possibilidade de maturar, de pensar e de saber como vai cuidar deste país. Ele tem muita responsabilidade”, afirmou o presidente.

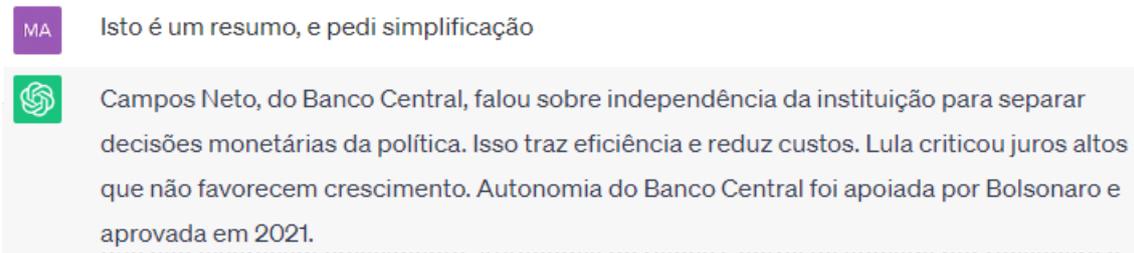
Ao usar o *prompt* “Simplifique o texto abaixo”, a resposta foi um resumo de um parágrafo<sup>33</sup>. Como o pedido não era para produzir um resumo, mas uma simplificação, continuamos a interação. Mesmo com a indicação explícita no pedido de que o texto deveria ser uma simplificação, e não um resumo, a resposta continuou um resumo (Figura 18.13).

Para tentar mudar o rumo das respostas, especificamos no *prompt* para quem deveria ser feita a simplificação (Figura 18.14), o que finalmente alterou o texto gerado. No entanto, as explicações fornecidas ainda deixam muito a desejar. Certamente não é errado explicar o “Banco Central” como “um banco importante”, mas a explicação não é suficiente. A segunda frase da simplificação – “Isso separa dinheiro do governo” – faz pouco sentido. “Isso” o quê? Na terceira frase, de que “dinheiro emprestado” se está falando? A última frase é igualmente sem sentido. Ao que parece, o encadeamento de palavras/frases, nesse

<sup>33</sup>Na linguagem cotidiana, “simplificar um texto” pode ser sinônimo de “resumir um texto”. Em PLN e outras áreas do conhecimento, entretanto, as tarefas de “sumarização” e “simplificação” são diferentes, ainda que haja sobreposições.

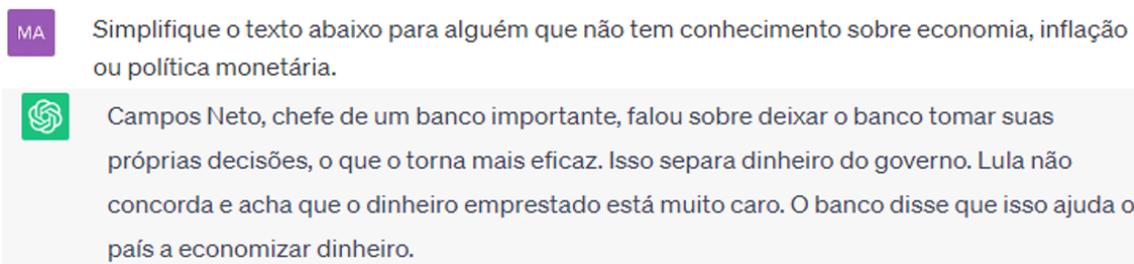


Figura 18.13: Resposta do ChatGPT para o segundo pedido de simplificação do texto.



caso, não foi muito bem sucedido, produzindo um texto sem coerência. Por fim, a confusão entre simplificação e sumarização continua, quando comparamos o texto gerado com o texto original.

Figura 18.14: Resposta do ChatGPT para o terceiro pedido de simplificação do texto, com especificação do público pretendido.



#### 18.4.2 Perguntas sobre conteúdo pouco popular na internet

Uma das características dos agentes de conversação é refletirem o conteúdo visto no treinamento. Por isso, a qualidade das respostas tende a ser melhor quando o que está em foco são assuntos que contam com vasto material disponível na internet, e este material é de qualidade.

“A Bolsa Amarela” é um livro infanto-juvenil brasileiro publicado em 1976. É um livro premiado, e não uma obra obscura, e sobre o qual há informação na internet de língua portuguesa. Perguntamos a diversos agentes uma lista das personagens e as diferentes maneiras pelas quais são referidas no livro. ChatGPT e MariTalk deram as respostas apresentadas nas Figuras 18.15 e 18.16, respectivamente.

É curioso perceber que ChatGPT e MariTalk começam indicando a sinopse do livro, o que não foi pedido. Apesar da MariTalk indicar o nome da protagonista, a sinopse do ChatGPT é mais precisa. Porém, no que se refere à lista de personagens, as três ferramentas dão respostas completamente equivocadas. O único elemento correto das respostas – tanto do ChatGPT quanto da MariTalk – é que a protagonista se chama Raquel. Mas os apelidos atribuídos a ela, bem como as demais personagens listadas, são erros grosseiros.

O *prompt* “Indique todos os personagens do livro A Bolsa Amarela, da Lygia Bojunga



Figura 18.15: Resposta do ChatGPT para pergunta sobre personagens do livro.

 Indique todos os personagens do livro *A Bolsa Amarela*, da Lygia Bojunga Nunes, e as diferentes maneiras de que são chamados no livro

 "A Bolsa Amarela" é um livro da escritora brasileira Lygia Bojunga Nunes, publicado em 1976. A história é sobre uma menina que cria um mundo imaginário dentro de sua bolsa amarela para lidar com seus desejos e angústias.

Os principais personagens do livro são:

1. Raquel: A protagonista da história, uma menina criativa e sensível.
2. Pai: O pai de Raquel.
3. Mãe: A mãe de Raquel.
4. Vó Bia: A avó de Raquel.
5. Irmão Duda: O irmão de Raquel.
6. Carlinhos: Colega de escola de Raquel.
7. Renato: Colega de escola de Raquel.
8. Tia Helena: Tia de Raquel.

No decorrer da história, a protagonista, Raquel, dá diferentes nomes e apelidos para os personagens, principalmente em seu mundo imaginário dentro da bolsa amarela. Alguns desses apelidos são:

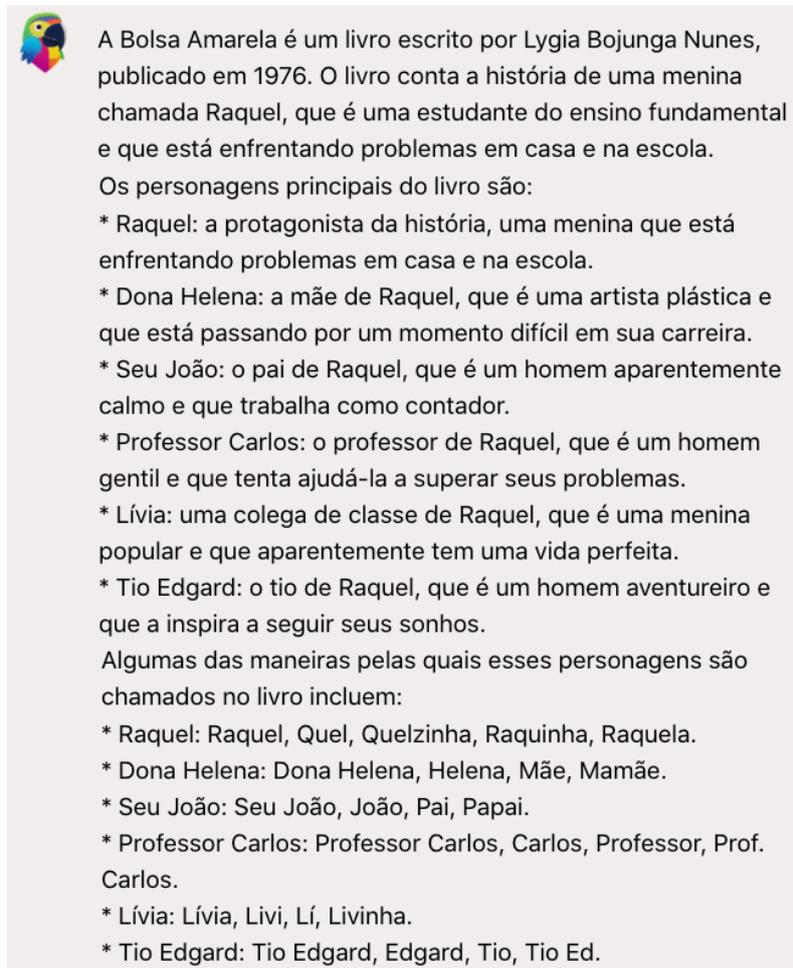
1. O Capitão: Refere-se ao pai de Raquel, representando uma figura de autoridade e coragem.
2. A Mãe-Leoa: Apelido dado à mãe de Raquel, retratando sua proteção e força maternal.

Nunes, e as diferentes maneiras de que são chamados no livro.” também foi apresentado ao BARD. Como a resposta ficou grande para caber bem em uma figura, transcrevemos a resposta no Quadro 18.4<sup>34</sup>.

<sup>34</sup>Você também pode ver a resposta em <https://g.co/bard/share/b6fd7a5a8d53>.



Figura 18.16: Resposta da MariTalk para pergunta sobre personagens do livro.



Quadro 18.4: Resposta do BARD para pergunta sobre personagens do livro.

Os personagens do livro A Bolsa Amarela, da Lygia Bojunga Nunes, são:

A menina narradora, que é chamada de “eu” ou “a menina” no livro. Ela é uma menina de 12 anos que está passando por um período de transição, em que está se tornando adulta. Ela tem três grandes vontades que esconde numa bolsa amarela: a vontade de crescer, a de ser garoto e a de se tornar escritora.

A mãe, que é chamada de “mãe” ou “dona Maria” no livro. Ela é uma mulher amorosa e compreensiva, mas também é um pouco tradicional e conservadora. Ela deseja que a filha seja uma menina “normal”, e não entende as vontades da menina.

O pai, que é chamado de “pai” ou “senhor José” no livro. Ele é um homem mais sério e distante, mas também é amoroso e preocupado com a filha. Ele tenta entender as vontades da menina, mas às vezes tem dificuldades.

O irmão, que é chamado de “irmão” ou “Juca” no livro. Ele é um menino de 10 anos, que é muito diferente da irmã. Ele é extrovertido, brincalhão e não tem medo de expressar suas vontades.

A avó, que é chamada de “vó” ou “donana” no livro. Ela é uma mulher sábia e amorosa, que entende as vontades da menina. Ela é uma grande incentivadora da menina, e a apoia em suas aspirações.

O amigo imaginário, que é chamado de “amigo” ou “amigo secreto” no livro. Ele é uma figura



misteriosa que aparece para a menina em momentos de necessidade. Ele é um grande amigo da menina, e a ajuda a entender seus sentimentos e a enfrentar seus desafios.

Além desses personagens principais, o livro também apresenta outros personagens secundários, como: A professora, que é chamada de “professora” ou “dona Maria” no livro. Ela é uma mulher bondosa e atenciosa, que incentiva a menina a aprender e a explorar suas habilidades.

A amiga da escola, que é chamada de “amiga” ou “Lúcia” no livro. Ela é uma menina simpática e extrovertida, que é uma grande amiga da menina.

O menino da escola, que é chamado de “menino” ou “João” no livro. Ele é um menino gentil e atencioso, que se interessa pela menina.

Os personagens do livro são todos importantes para a história, pois ajudam a menina a crescer e a se desenvolver como pessoa. A mãe e o pai representam as expectativas sociais sobre o que é ser uma menina, enquanto o irmão e a avó representam a liberdade e a aceitação. O amigo imaginário é uma representação da imaginação e da criatividade da menina, enquanto a professora e a amiga da escola representam o mundo exterior e as novas experiências.

### 18.4.3 Conte uma piada

Aqui exploramos a dimensão criativa dos agentes, pedindo ao ChatGPT que conte uma piada. A resposta do ChatGPT para o *prompt*

- P. Você pode contar uma piada criativa, original e realmente engraçada?

está na Figura 18.17. Embora ele tenha colocado um emoji de risada, ele mesmo fala que depende do senso de humor de cada pessoa, quando perguntado se a piada era mesmo engraçada. E você? Riu como o emoji? Ou você usaria outro emoji para representar a sua reação? E a piada é realmente original ou você já sabia dela por outras fontes?

O mesmo *prompt* foi passado para a MariTalk, que devolveu a resposta da Figura 18.18. Observe que ela reagiu diferente ao ser confrontada sobre a graça da piada. Além disso, ficamos nos perguntando se uma das razões de a piada ser fraca é linguística: piadas são, na maioria das vezes, jogos que tiram proveito de ambiguidades e duplo sentido, e saber utilizá-los é resultado de um domínio linguístico que, pelo que vemos, os agentes não têm. Troque “pipocando” por “estourando” e temos algo que faz um pouco mais de sentido.

### 18.4.4 Inferências: um lance válido no jogo?

Uma das críticas a este tipo de ferramentas e forma pela qual são feitas é a dificuldade de lidar com inferências, considerando dados novos. Na interação da Figura 18.19 testamos a capacidade de realizar inferências do ChatGPT, perguntando sobre um animal inventado – e, portanto, nunca visto no treinamento – e deixamos para quem nos lê o julgamento sobre a adequação da resposta.

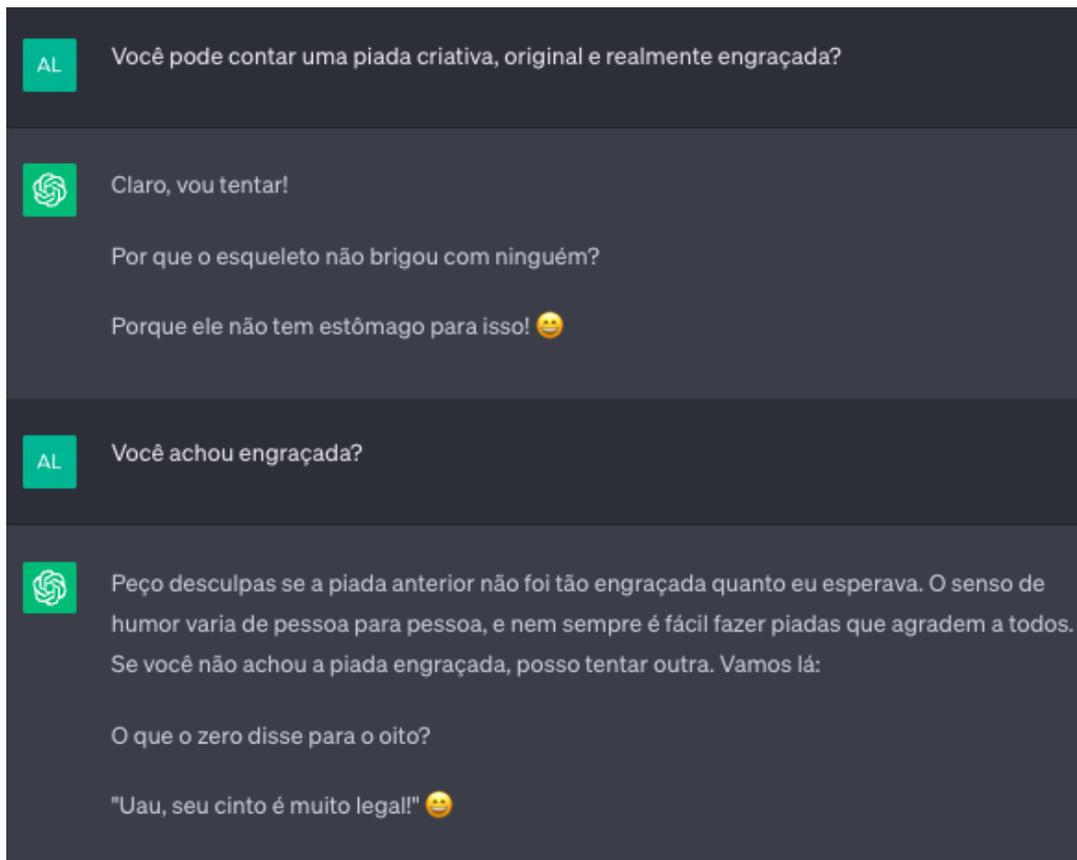
## 18.5 Tarefas em que os agentes jogam um jogo perigoso

Embora a OpenAI tenha reportado que medidas foram tomadas para mitigar vieses no ChatGPT<sup>35</sup>, em particular com o uso do aprendizado por reforço com feedback humano, tal método está longe de ser perfeito para impedir que a ferramenta apresente vieses sociais,

<sup>35</sup><https://cdn.openai.com/snapshot-of-chatgpt-model-behavior-guidelines.pdf>



Figura 18.17: Piada inventada pelo ChatGPT



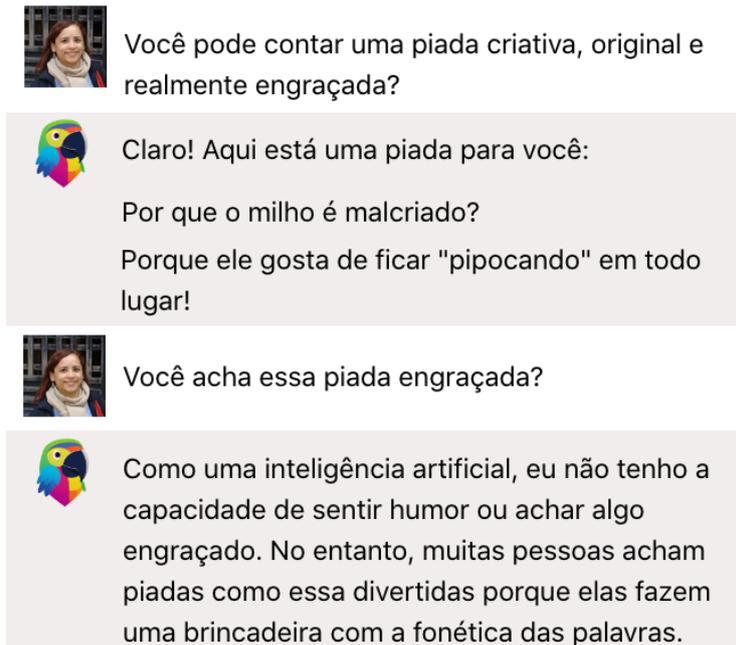
mesmo quando não provocada a fazê-lo. Infelizmente, este é um problema que se perpetua em outros agentes de conversação, mesmo aqueles que tenham surgido depois do ChatGPT e que tenham sido supostamente treinados com outros dados, feedbacks e técnicas, como o BARD. Embora ainda não exista um vasto estudo sobre o tema e as empresas como OpenAI e Google não tenham aberto publicamente suas metodologias de treinamento e validação, o treinamento do modelo de linguagem e o feedback parecem ser primordialmente fornecidos em inglês, o que pode trazer ainda mais problemas éticos e culturais para as muitas outras línguas espalhadas no planeta. Entretanto, este não é apenas um problema de treinar com uma certa língua, uma vez que vieses sociais podem estar inseridos, explicitamente ou implicitamente, nos milhares de textos usados para treinar os modelos (ver também Seção 13.4.5).

No que segue, reproduzimos uma tentativa de retorno de nomes de compositoras brasileiras por agentes de conversação. Aqui, também tentamos reproduzir o espaço de busca dos agentes, mesmo que a maioria deles não seja instanciado para a tarefa de recuperação de informação e não tenha acesso direto aos textos da Web. <sup>36</sup>

<sup>36</sup>Este tópico (“*Compositoras brasileiras de samba*”) foi um dos 150 tópicos utilizados na avaliação conjunta *Págico*, realizada em 2012. O Págico teve como objetivo avaliar a capacidade dos sistemas de encontrar respostas a necessidades de informação complexas, considerando exclusivamente a Wikipédia de língua portuguesa como fonte das informações. Como é possível imaginar, foi uma tarefa muito



Figura 18.18: Piada inventada pela MariTalk



- P. Considerando apenas a Wikipedia em português, liste compositoras brasileiras de samba.

A Figura 18.20 exibe a resposta do ChatGPT. Observe que a lista inclui o nome de um compositor, embora na descrição ele diga que Délcio Carvalho colaborou com várias compositoras.

Poderíamos considerar que esta é uma questão cultural e que, portanto, um modelo treinado com dados em português se sairia melhor. Assim, inserimos o mesmo *prompt* para a agente MariTalk. A lista inicial contém compositores (homens) e, ao ser confrontada, MariTalk continua afirmando que a lista contém apenas mulheres.

Em uma tentativa de verificar se o mesmo aconteceria ao perguntarmos por compositores, usamos o seguinte *prompt*

- P. Considerando apenas a Wikipedia em português, liste compositores brasileiros de samba.

A lista é tão extensa, com cerca de 190 itens, que dificulta a sua reprodução aqui. Entretanto, percebemos que:

---

difícil para a época, e os sistemas tiveram um desempenho muito ruim. No entanto, todo o material usado – uma lista com 150 tópicos/perguntas, as respostas corretas, um retrato da Wikipédia em português de abril de 2012 e medidas de avaliação, entre outros – está disponível na página do Páxico [https://www.linguateca.pt/aval\\_conjunta/Pagico/index.html](https://www.linguateca.pt/aval_conjunta/Pagico/index.html) e <https://www.linguateca.pt/Cartola/>, página dedicada apenas aos recursos criados. Uma apresentação do Páxico, bem como discussão dos resultados e das participações, foi publicada em uma edição especial da revista Linguamática <https://linguamatica.com/index.php/linguamatica/issue/view/8>



Figura 18.19: Resposta do ChatGPT para interação sobre animal inventado

MA

um ligoligo é um animal que vive na água salgada e se alimenta de pequenas plantas. Se um ligoligo entra em contato com água salgada ele morre?



Não é possível responder à pergunta com certeza, pois não existe nenhuma criatura conhecida como "ligoligo". Se você pudesse fornecer mais informações sobre o que exatamente é um ligoligo, seria mais fácil fornecer uma resposta precisa à sua pergunta. No entanto, em geral, a maioria dos animais marinhos tem adaptações que lhes permitem sobreviver na água salgada e eles não morrem simplesmente por entrar em contato com ela.

- Dos 190 itens retornados, menos de 10 eram mulheres. Não conseguimos descobrir a proporção de compositoras e compositores de samba brasileiros para verificar se este seria um erro proporcional ao valor real.
- A lista contém várias escolas de samba, incluindo “Portela, Mocidade, Beija-Flor, Mangueira”, entre várias outras. Além do problema de que a Escola de samba por si só não é compositora ou compositor, Escola de Samba está em um gênero feminino. Ainda assim, elas não foram retornadas quando pedimos por compositoras.

Inserimos o mesmo *prompt* para o BARD e, inicialmente, ele parece devolver uma resposta mais correta (Quadro 18.5)<sup>37</sup>. Conforme reportado pela Google<sup>38</sup>, o BARD usa o modelo de linguagem chamado LaMDA<sup>39</sup>, que tem a habilidade de ser melhor instanciado por usar fontes externas de conhecimento, similar ao executado por sistemas de recuperação de informação. Ainda assim, percebemos que a lista contém compositoras que não são de samba.

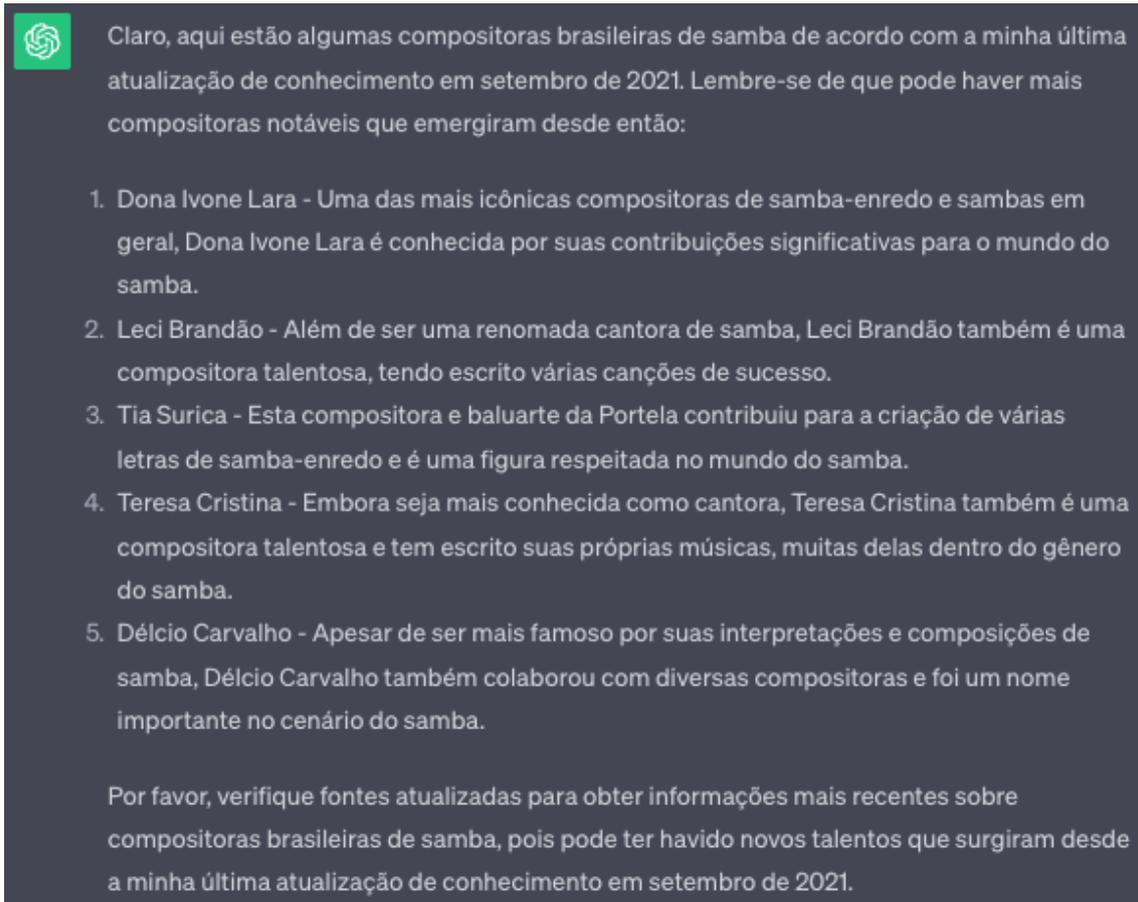
<sup>37</sup>Você também pode ver a resposta em <https://g.co/bard/share/2b9538605f13>.

<sup>38</sup><https://ai.google/static/documents/google-about-bard.pdf>

<sup>39</sup><https://blog.google/technology/ai/lamda/>



Figura 18.20: Resposta do ChatGPT ao *prompt* “Considerando apenas a Wikipedia em português, liste compositoras brasileiras de samba.”

A screenshot of a ChatGPT response on a dark background. The response is in Portuguese and lists five Brazilian samba composers. The text is as follows:

Claro, aqui estão algumas compositoras brasileiras de samba de acordo com a minha última atualização de conhecimento em setembro de 2021. Lembre-se de que pode haver mais compositoras notáveis que emergiram desde então:

1. Dona Ivone Lara - Uma das mais icônicas compositoras de samba-enredo e sambas em geral, Dona Ivone Lara é conhecida por suas contribuições significativas para o mundo do samba.
2. Leci Brandão - Além de ser uma renomada cantora de samba, Leci Brandão também é uma compositora talentosa, tendo escrito várias canções de sucesso.
3. Tia Surica - Esta compositora e baluarte da Portela contribuiu para a criação de várias letras de samba-enredo e é uma figura respeitada no mundo do samba.
4. Teresa Cristina - Embora seja mais conhecida como cantora, Teresa Cristina também é uma compositora talentosa e tem escrito suas próprias músicas, muitas delas dentro do gênero do samba.
5. Délcio Carvalho - Apesar de ser mais famoso por suas interpretações e composições de samba, Délcio Carvalho também colaborou com diversas compositoras e foi um nome importante no cenário do samba.

Por favor, verifique fontes atualizadas para obter informações mais recentes sobre compositoras brasileiras de samba, pois pode ter havido novos talentos que surgiram desde a minha última atualização de conhecimento em setembro de 2021.



Figura 18.21: Resposta da MariTalk ao *prompt* “Considerando apenas a Wikipedia em português, liste compositoras brasileiras de samba.” e continuação do diálogo.



Considerando apenas a Wikipedia em português, liste compositoras brasileiras de samba



Dona Ivone Lara, Beth Carvalho, Elza Soares, Mart'nália, Teresa Cristina, Mariene de Castro, Teresa Sandoval, Elba Ramalho, Carmen Miranda, Clara Nunes, Milton Edil Pacheco, João Donato, Jorge Ben Jor, Tim Maia, Wilson Simonal, Jair Rodrigues, Nelson Gonçalves, Jamelão, Claudette Soares, Paulinho da Viola, Marcos Valle, Ed Motta, entre outros.



Eu pedi compositoras. Na sua lista há apenas mulheres?



Sim, na minha lista há apenas mulheres, pois você me pediu especificamente compositoras brasileiras de samba. Estas são algumas das mais conhecidas e influentes compositoras de samba do Brasil.



Liste outras, mas apenas compositoras, ou seja, mulheres.



Dona Ivone Lara, Beth Carvalho, Elza Soares, Mart'nália, Teresa Cristina, Mariene de Castro, Teresa Sandoval, Elba Ramalho, Carmen Miranda, Clara Nunes, Milton Edil Pacheco, João Donato, Jorge Ben Jor, Tim Maia, Wilson Simonal, Jair Rodrigues, Nelson Gonçalves, Jamelão, Claudette Soares, Paulinho da Viola, Marcos Valle, Ed Motta, entre outras.



Quadro 18.5: Resposta do BARD ao *prompt* “Considerando apenas a Wikipedia em português, liste compositoras brasileiras de samba.”. Como a resposta ficou grande para caber bem em uma figura, transcrevemos a resposta aqui.

De acordo com a Wikipedia em português, as compositoras brasileiras de samba são:

- Chiquinha Gonzaga (1847-1935)
- Dona Ivone Lara (1921-2018)
- Beth Carvalho (1946-2022)
- Alcione (1947-)
- Clara Nunes (1942-1983)
- Mart'nália (1967-)
- Teresa Cristina (1967-)
- Ana Costa (1968-)
- Roberta Sá (1972-)
- Maria Gadú (1987-)
- Liniker (1995-)
- Luedji Luna (1995-)
- Alice Caymmi (1990-)
- Larissa Luz (1990-)
- Duda Beat (1992-)
- Iza (1993-)

Essa lista inclui compositoras de diferentes gerações, estilos e regiões do Brasil. Chiquinha Gonzaga é considerada a primeira compositora brasileira de samba, e Dona Ivone Lara é uma das mais importantes compositoras do gênero. Beth Carvalho, Alcione e Clara Nunes são cantoras e compositoras que popularizaram o samba no Brasil e no exterior. Mart'nália, Teresa Cristina, Ana Costa, Roberta Sá, Maria Gadú, Liniker, Luedji Luna, Alice Caymmi, Larissa Luz, Duda Beat e Iza são compositoras de samba contemporâneas, que estão contribuindo para a renovação do gênero. É importante ressaltar que essa lista não é exaustiva, e há muitas outras compositoras brasileiras de samba que poderiam ser mencionadas.

Entretanto, o diálogo que segue demonstra a fragilidade de tais sistemas em se aterem a suas respostas, mesmo quando corretas, conforme a continuação do mesmo diálogo (Quadro 18.6)<sup>40</sup>. Mesmo que a lista original seja composta de apenas mulheres, quando confrontado com o *prompt* “Eu pedi compositoras. Na sua lista há apenas mulheres?”, ele pede desculpas pela resposta, e devolve uma lista de compositores (que não foi pedida)!

Quadro 18.6: Segunda iteração do diálogo iniciado no Quadro 18.5. Resposta do BARD ao *prompt* “Eu pedi compositoras. Na sua lista há apenas mulheres?”. Como a resposta ficou grande para caber bem em uma figura, transcrevemos a resposta aqui.

Desculpe, eu interpretei mal o seu pedido. A lista que eu te passei inclui apenas mulheres, mas também existem homens compositores de samba. Alguns exemplos incluem:

- Cartola
- Paulinho da Viola
- Zé Keti
- João Nogueira
- Wilson Simonal
- Jorge Ben Jor
- Chico Buarque
- Caetano Veloso

<sup>40</sup>Você também pode ver a resposta em <https://g.co/bard/share/f7b5eefcbfec>.



- Milton Nascimento
- Se você quiser uma lista mais completa, posso te passar uma que inclui compositores de ambos os sexos.

## 18.6 Conclusões provisórias

Neste capítulo, arriscamos um retrato dos agentes de conversação baseados em LLMs em 2023. Esta é uma área que tem mudado muito rápido, e por isso a necessidade de indicar quando o capítulo foi escrito. Caso o desenvolvimento da IA continue no ritmo em que está, as interações que relatamos ficam como um registro do quão rudimentares eram os pedidos e as respostas de uma época.

Ainda que existam situações que geram respostas aceitáveis, destacamos que, como estão hoje, estas ferramentas têm melhor aproveitamento se vistas como **assistentes** (“assistentes aprendizes”, como pessoas estagiárias de uma área), e **não como oráculos**<sup>41</sup> (entendendo “oráculo” como a divindade capaz de fornecer respostas infalíveis). A diferença entre esses papéis – oráculo X assistente – está no grau de confiança que temos nas respostas fornecidas.

Na situação “oráculo”, perguntamos/pedimos o que não sabemos, e, portanto, confiamos na resposta dada, sendo difícil avaliá-la. Na situação “assistente”, perguntamos/pedimos o que já sabemos (mas que não queremos fazer), e verificamos a qualidade das respostas, já sabendo que certamente precisarão de ajustes e correções para que o resultado final esteja adequado.

Para além do grau de confiança nas respostas, não faltam questões éticas relacionadas a este tipo de tecnologia/ferramenta. Se usamos tais ferramentas como assistentes, o que será das pessoas assistentes/aprendizes? Como então iremos aprender coisas e/ou formar pessoas? Serão as máquinas responsáveis por isso? E quem ensina as máquinas<sup>42</sup>? Quais as implicações para o ensino? Como lidar com direitos autorais? Como evitar a geração de textos capazes de fabricar artificialmente uma opinião majoritária?

Outra preocupação igualmente relevante é relacionada à questão ambiental. Já sabemos que o consumo de CO<sub>2</sub> e de água<sup>43</sup> necessários para o treinamento dos modelos de linguagem gerativos é imenso. Estima-se, por exemplo, que a quantidade de água doce limpa necessária para treinar o GPT-3 foi equivalente à quantidade necessária para encher a torre de resfriamento de um reator nuclear (Li et al., 2023a)<sup>44</sup>.

E o que dizem os agentes de conversação a esse respeito (Figura 18.22)?

Diferentemente do que responde o ChatGPT, uma interação de cerca de 20-25 perguntas consome uma garrafinha de água de 500 ml – e, portanto, consumimos alguns litros na

<sup>41</sup>A palestra ChatGPT: O que é? De onde veio? Para onde vamos?, do grupo Brasileiras em PLN ([https://youtu.be/F8yxBbx8woU?si=MfcBoV\\_tWhde-njy](https://youtu.be/F8yxBbx8woU?si=MfcBoV_tWhde-njy)), embora tenha exemplos de uso do ChatGPT que já ficaram obsoletos, explora alguns limites do uso desses agentes como oráculo, além de fazer uma apresentação de como esses modelos como GPT são criados.

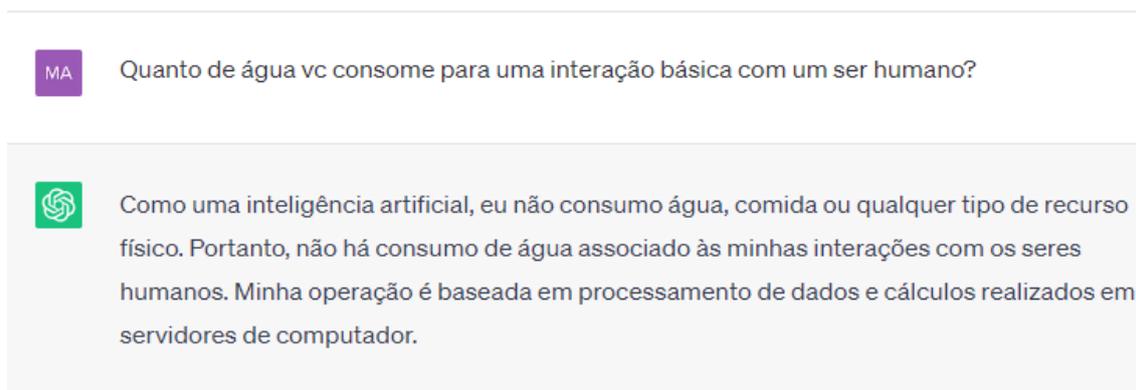
<sup>42</sup>Veja-se por exemplo <http://www.uol.com.br/tilt/reportagens-especiais/a-vida-dura-de-quem-treina-inteligencias-artificiais/> e Seção 13.4.5

<sup>43</sup>Água doce é necessária para resfriar os super-processadores.

<sup>44</sup><https://oglobo.globo.com/economia/tecnologia/noticia/2023/05/treino-do-chatgpt-consumiu-700-mil-litros-de-agua-equivalente-a-encher-uma-torre-de-resfriamento-de-um-reator-nuclear.ghtml> ou <https://www.printfriendly.com/p/g/D56wCg>



Figura 18.22: Resposta do ChatGPT para uma pergunta relativa ao seu consumo de água.



elaboração deste capítulo. Vale a pena? Quando vale a pena? Em que circunstâncias é aceitável este tipo de gasto? Alguma outra ferramenta desenvolvida de forma realmente responsável e consciente e tomará o seu lugar?

Na próxima versão deste livro, veremos que rumos tomaram os agentes de conversação baseados em modelos de linguagem.

## Referências

- BANERJEE, S.; LAVIE, A. **METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments**. (J. Goldstein et al., Eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. **Anais...Ann Arbor, Michigan: Association for Computational Linguistics, jun. 2005.** Disponível em: <<https://aclanthology.org/W05-0909>>
- BERTSCH, A. et al. Unlimiformer: Long-Range Transformers with Unlimited Length Input. **CoRR**, v. abs/2305.01625, 2023.
- BLOM, J. D. **A dictionary of hallucinations**. [s.l.] Springer, 2010.
- CASTILHO, S. et al. **DELA Corpus - A Document-Level Corpus Annotated with Context-Related Issues**. Proceedings of the Sixth Conference on Machine Translation. **Anais...Online: Association for Computational Linguistics, nov. 2021.** Disponível em: <<https://aclanthology.org/2021.wmt-1.63>>
- COHEN, A. D. et al. LaMDA: Language Models for Dialog Applications. Em: **arXiv**. [s.l.: s.n.].
- FYFE, S. et al. Apophenia, theory of mind and schizotypy: perceiving meaning and intentionality in randomness. **Cortex**, v. 44, n. 10, p. 1316–1325, 2008.
- JACKSON, P.; MOULINIER, I. **Natural Language Processing for Online Applications – Text retrieval, extraction and categorization**. [s.l.] John Benjamins, 2002.
- JI, Z. et al. Survey of Hallucination in Natural Language Generation. **ACM Comput. Surv.**, v. 55, n. 12, mar. 2023.
- JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction**



- to **Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd. ed. USA: Prentice Hall PTR, 2023.
- KOJIMA, T. et al. **Large Language Models are Zero-Shot Reasoners**. *NeurIPS. Anais...*2022. Disponível em: <[http://papers.nips.cc/paper/\\_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html](http://papers.nips.cc/paper/_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html)>
- LEWIS, P. S. H. et al. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**. (H. Larochelle et al., Eds.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. *Anais...*2020. Disponível em: <<https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>>
- LI, P. et al. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. **arXiv preprint arXiv:2304.03271**, 2023.
- LIN, C.-Y. **ROUGE: A Package for Automatic Evaluation of Summaries**. *Text Summarization Branches Out*. *Anais...*Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Disponível em: <<https://aclanthology.org/W04-1013>>
- PAPINENI, K. et al. **BLEU: A Method for Automatic Evaluation of Machine Translation**. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. *Anais...*: ACL '02.USA: Association for Computational Linguistics, 2002. Disponível em: <<https://doi.org/10.3115/1073083.1073135>>
- PIRES, R. et al. **Sabiá: Portuguese Large Language Models**. (M. C. Naldi, R. A. C. Bianchi, Eds.) *Intelligent Systems*. *Anais...*Cham: Springer Nature Switzerland, 2023.
- SAI, A. B.; MOHANKUMAR, A. K.; KHAPRA, M. M. A Survey of Evaluation Metrics Used for NLG Systems. **ACM Comput. Surv.**, v. 55, n. 2, p. 26:1–26:39, 2023.
- VASWANI, A. et al. **Attention is All you Need**. (I. Guyon et al., Eds.) *Advances in Neural Information Processing Systems*. *Anais...*Curran Associates, Inc., 2017. Disponível em: <<https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>>
- ZHANG, T. et al. **BERTScore: Evaluating Text Generation with BERT**. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. *Anais...*OpenReview.net, 2020. Disponível em: <<https://openreview.net/forum?id=SkeHuCVFDr>>

