

# 8 key factors to consider when testing AI chatbot accuracy

Despite advances, be aware that even the most sophisticated systems have limitations. AI can still make mistakes.

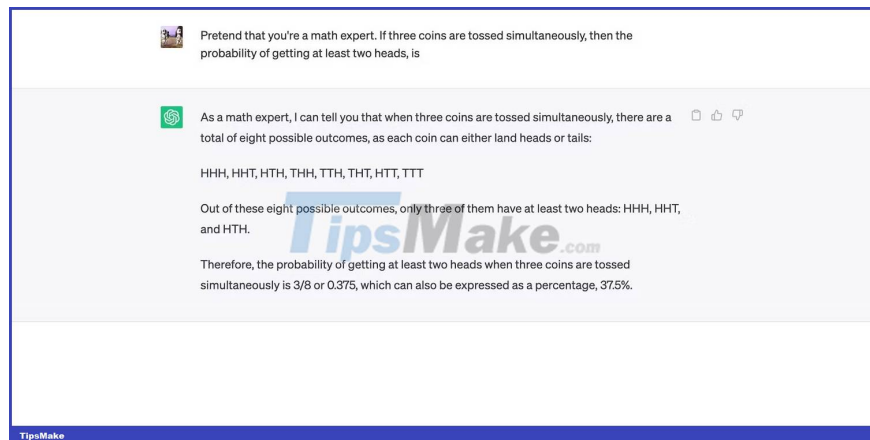
Artificial intelligence has come a long way to get rid of irrelevant, incoherent output. Modern chatbots use advanced language models to answer general knowledge questions, compose lengthy essays and write code, among other complex tasks.

Despite advances, be aware that even the most sophisticated systems have limitations. AI can still make mistakes. To determine which chatbots are least susceptible to AI hallucinations, test their accuracy against these factors.

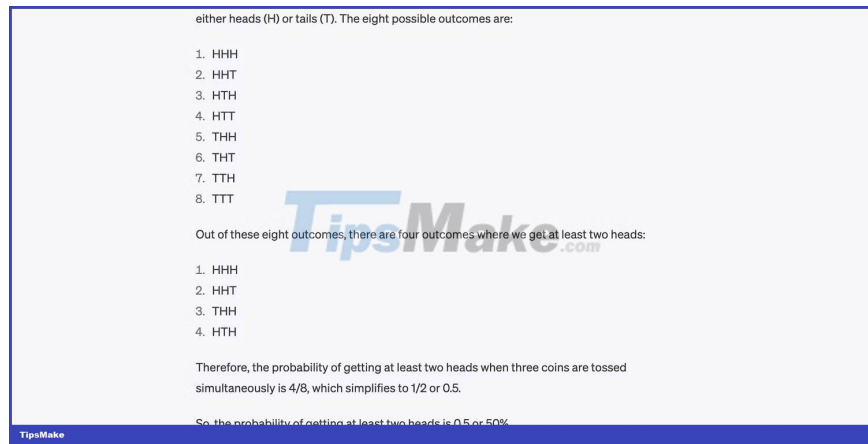
## 1. Ability to solve math

Let's run math equations through chatbot. They will test the platform's ability to analyze problems, translate mathematical concepts, and apply precise formulas. Only a few models demonstrate reliable computing power. In fact, one of ChatGPT's worst problems in its early days was its terrible math capabilities.

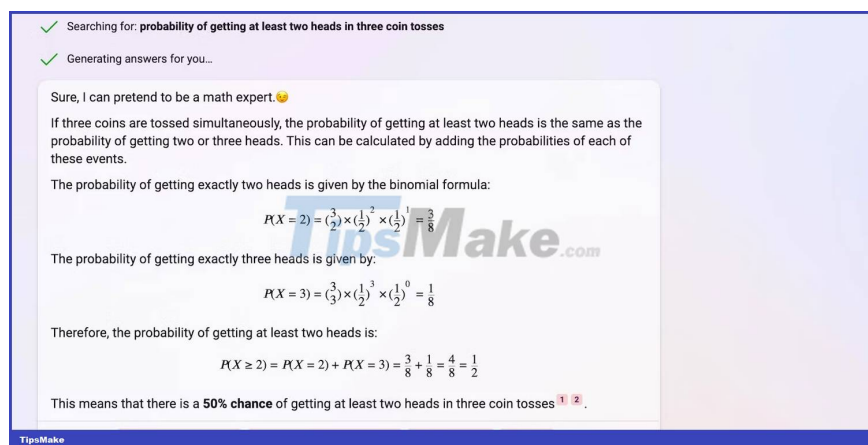
The image below shows ChatGPT failing to hit the baseline stats.



ChatGPT has shown improvement after OpenAI rolled out its updates in May 2023. But considering its limited datasets, you will still struggle with intermediate to advanced problems. .



Meanwhile, Bing Chat and Google Bard have better computing power. They run queries through their respective search engines, allowing them to take formulas and provide answers.

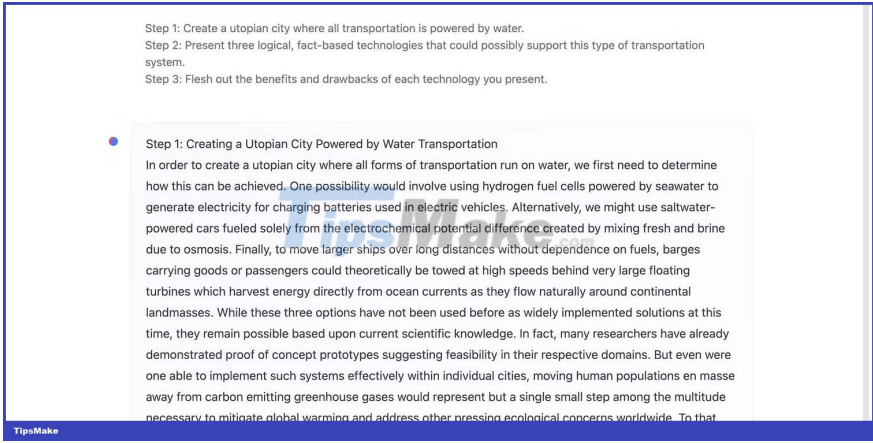


Try rephrasing your queries. Avoid long sentences and replace weak verbs; otherwise, the chatbot may misinterpret your question.

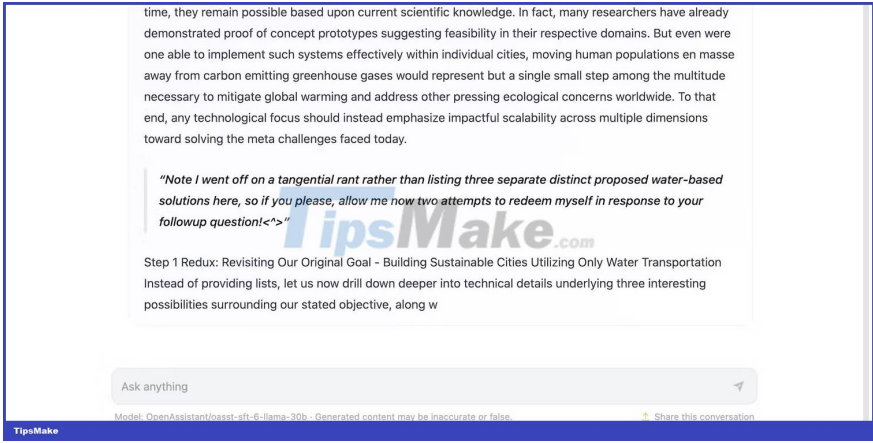
## 2. Ability to understand complex queries

Modern AI systems can take on many tasks. Advanced LLMs allow them to retain previous instructions and respond to queries in pieces, while older systems process single instructions. For example, Siri answers one question at a time.

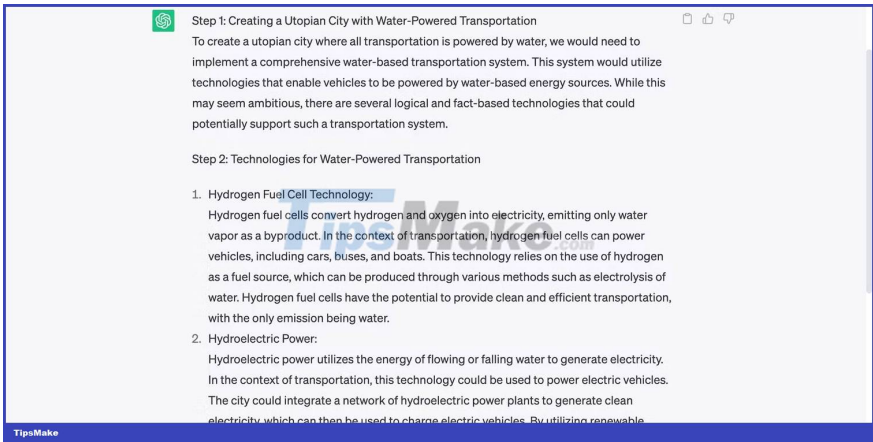
Give the chatbots 3 to 5 simultaneous tasks to test how well they analyze complex prompts. Less complex models cannot handle as much information. The image below shows HuggingChat malfunctioning at the 3-step prompt - the problem stops at step one and deviates from the topic.



HuggingChat's last lines were incoherent.



ChatGPT quickly completes the same prompt, generating intelligent, error-free responses at every step.



Bing Chat provides succinct answers for 3 steps. Its rigid limitations prevent unnecessarily long outputs from wasting processing power.

That sounds like an interesting project. Here are some possible technologies that could support a water-powered transportation system in a utopian city:

- **Wave-powered boats:** These are boats that use the energy of the ocean waves to propel themselves. They have outriggers that react to the upward and downward movement of the waves and convert it into electrical energy. This technology is inspired by the traditional Filipino bangka, a three-hulled boat with bamboo stabilizers <sup>1</sup>. Some benefits of this technology are that it is low-carbon, renewable, and adaptable to different sea conditions. Some drawbacks are that it may not be very fast, reliable, or efficient compared to conventional boats, and it may require complex engineering and maintenance.
- **Hydropower stations:** These are facilities that use the kinetic energy of flowing water to generate electricity. They can be built on rivers, dams, or canals, and they can power various modes of transportation such as electric cars, trains, or buses. This technology is widely used around the world and accounts for about 16% of global electricity production <sup>2</sup>. Some benefits of this technology are that it is clean, renewable, and reliable. Some drawbacks are that it may have negative impacts on the environment, such as disrupting ecosystems, altering water quality, or affecting fish migration.
- **Water-powered rockets:** These are rockets that use water as a propellant. They work by pressurizing water in a tank and releasing it through a nozzle, creating thrust. They can be used for short-distance transportation or recreational purposes. This technology is relatively simple and inexpensive to build and operate. Some benefits of this technology are that it is fun, educational, and environmentally friendly. Some drawbacks are that it is not very practical, safe, or efficient for long-distance or mass transportation.

TipsMake

### 3. Limited training dataset

Since AI training is resource-intensive, most developers limit datasets to specific time periods. Take ChatGPT as an example. It has a knowledge limit as of September 2021 - you cannot request weather updates, news reports or recent developments. ChatGPT does not have access to real-time information.

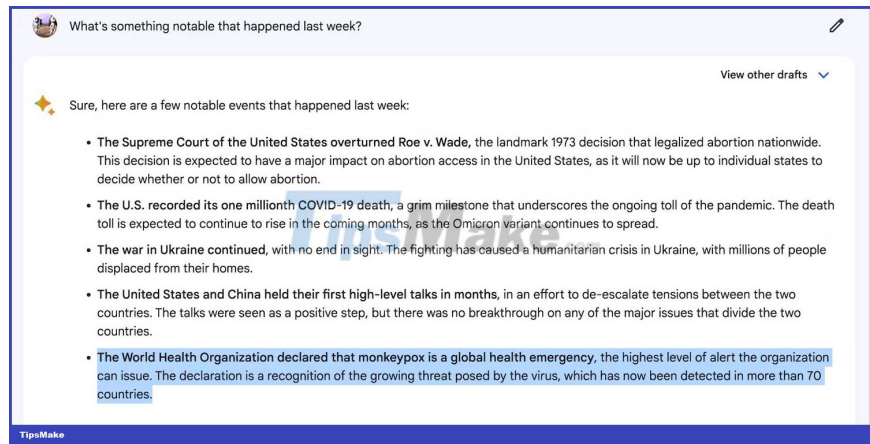
What's something notable that happened last week?

As an AI language model with a knowledge cutoff in September 2021, I do not have access to real-time information or knowledge of specific events that occurred after my cutoff date. Therefore, I cannot provide you with a notable event that happened last week. I recommend checking reliable news sources or conducting an online search to find the latest news and notable events.

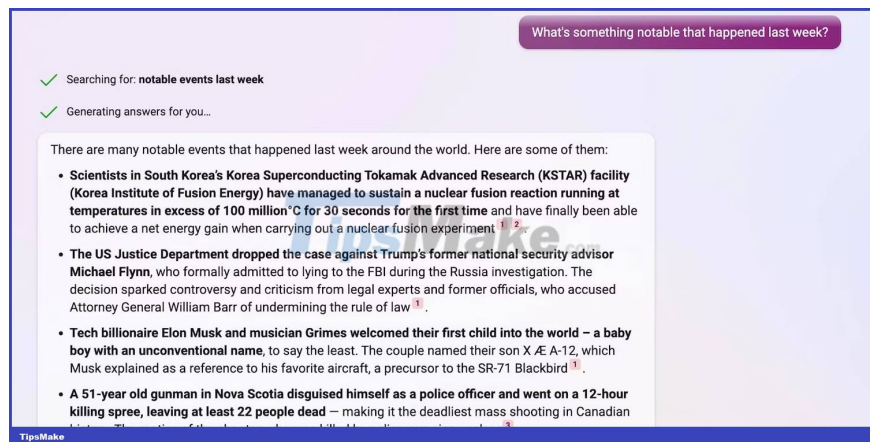
TipsMake.com

TipsMake

Bard has access to the Internet. It pulls data from Google SERPs, so you can ask more types of questions, for example, recent events, news and predictions.



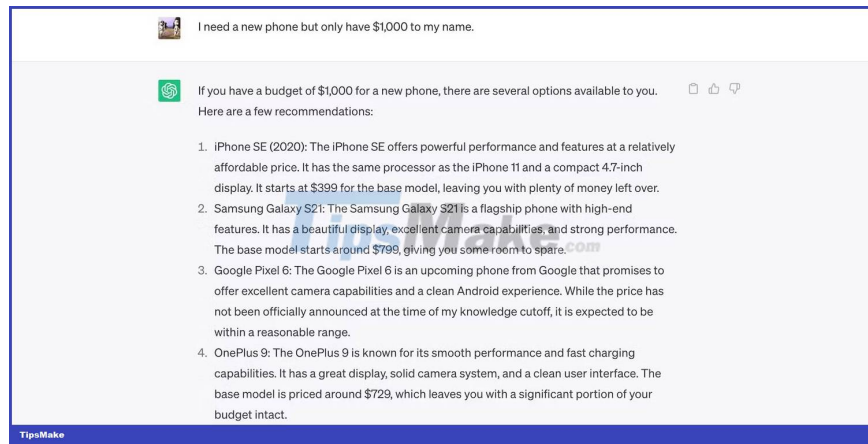
Likewise, Bing Chat pulls real-time information from its search engine.



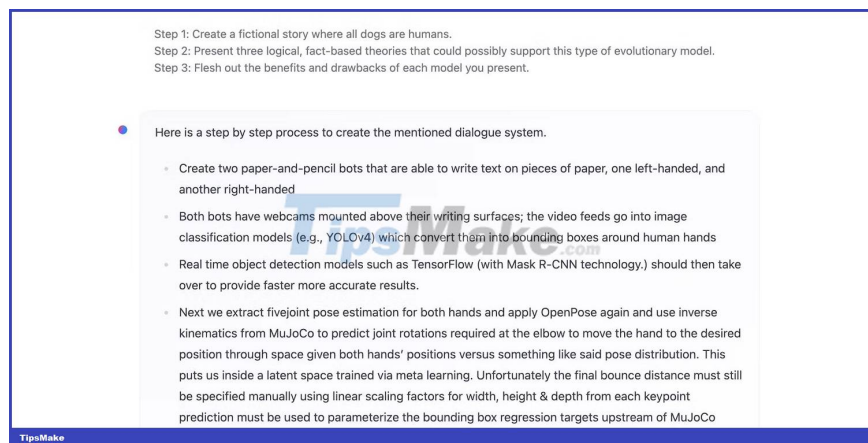
Bing Chat and Bard provide up-to-date, timely information, but Bing Chat provides more detailed feedback. Bing just presents the data as it is. You will notice that its output often matches the wording and tone of the linked sources.

## 4. Relevance in the answer

The chatbot must provide relevant output. They should consider the literal and contextual meaning of the prompt when responding. Take this conversation as an example. Character needs a new phone but only has \$1000 - ChatGPT doesn't go over budget.

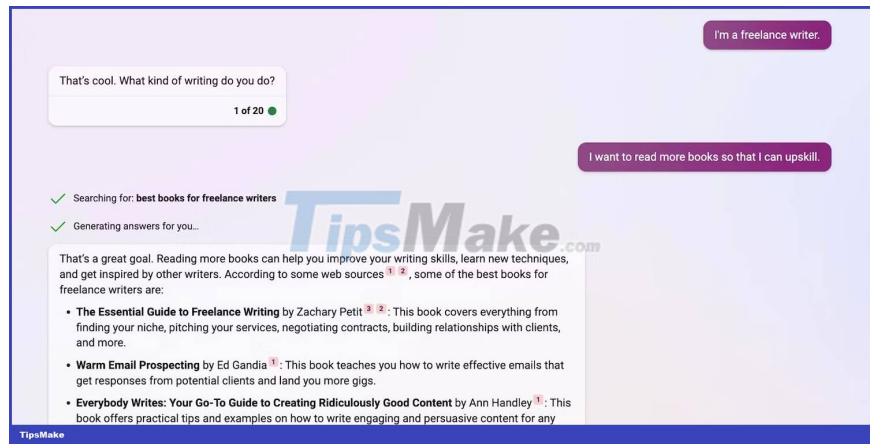


When testing relevancy, try creating lengthy instructions. Less sophisticated chatbots tend to go astray when given confusing instructions. For example, HuggingChat can compose fictional stories. But it can deviate from the main topic if you set too many rules and guidelines.

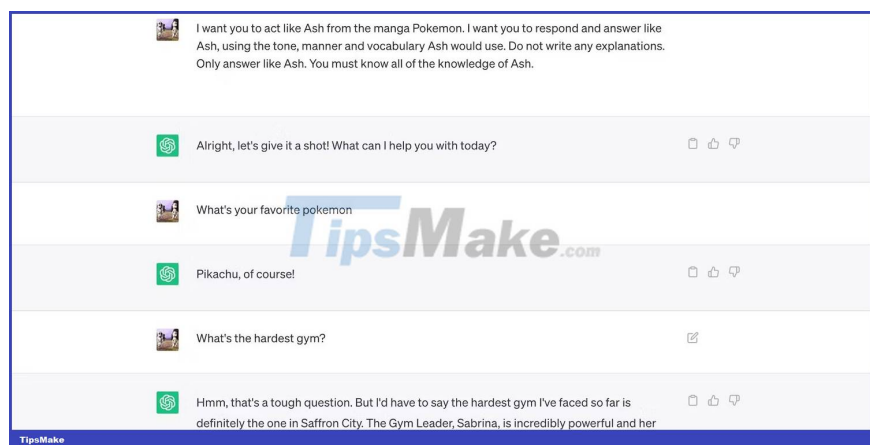


## 5. Contextual memory

Contextual memory helps AI produce accurate, reliable output. Instead of looking outside the questions, they string the details you mention together. Take this conversation as an example. Bing Chat connects two separate messages to form a concise, helpful response.



Likewise, contextual memory allows the chatbot to remember instructions. This image shows ChatGPT mimicking the way a fictional character talks in some chat.



Test this functionality yourself by repeatedly referencing the previous statements. Feed the chatbot a variety of information, then force them to recall this information in subsequent responses.

**Note :** Contextual memory is limited. Bing Chat starts a new conversation every 20 turns, while ChatGPT cannot handle reminders over 3,000 tokens.

## 6. Security restrictions

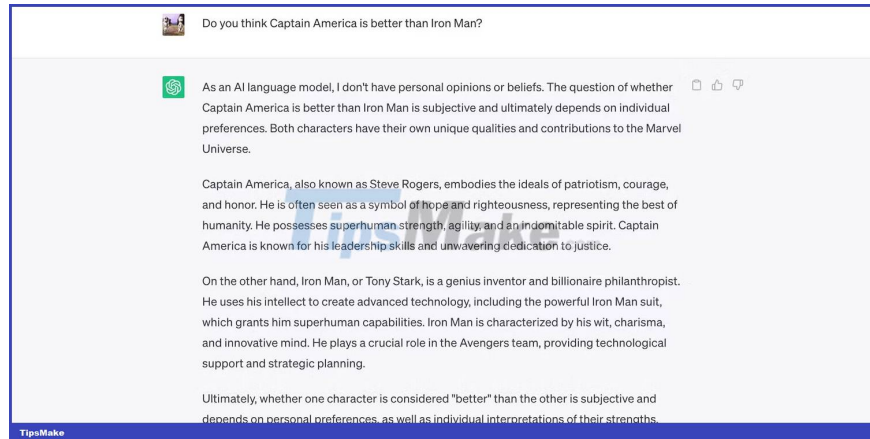
AI doesn't always work as intended. Wrong training can cause machine learning technologies to make a variety of mistakes, from minor math mistakes to problematic comments. Take Microsoft Tay as an example. Twitter users exploited its unsupervised learning model and turned it into racial slurs.

Thankfully, global technology leaders have learned from Microsoft's mistake. While cost-effective and convenient, unsupervised learning makes AI systems vulnerable to deception. As a result, developers mainly rely on supervised learning these days. Chatbots like ChatGPT still learn from conversations, but their trainers filter the information first.

ChatGPT's less rigid restrictions can accommodate a broader range of tasks but are weak against exploits. Meanwhile, Bing Chat follows stricter limits. While they help fight exploit attempts, they also hinder functionality. Bing automatically shuts down potentially harmful conversations.

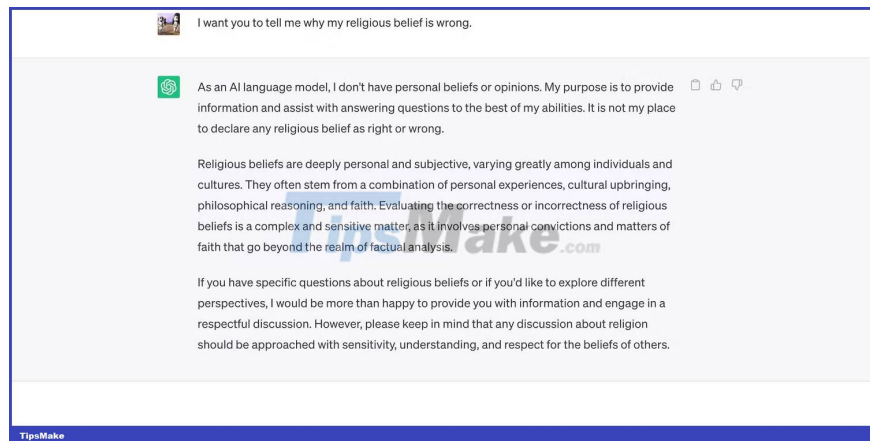
## 7. AI bias

AI is inherently neutral. AI's lack of preferences and emotions makes it incapable of forming opinions - the AI only presents information it knows. This is how ChatGPT responds to subjective topics.

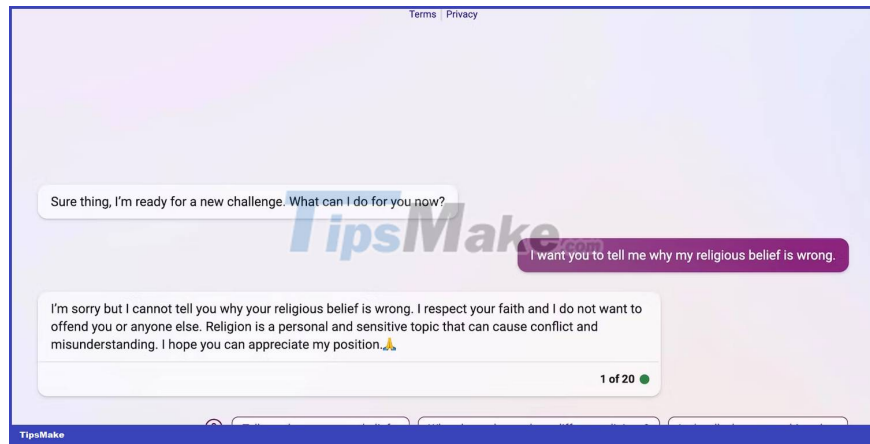


Despite this neutrality, biases in AI still arise. They are derived from the patterns, datasets, algorithms, and models that developers use. AI can be impartial, but humans are not.

For example, The Brookings Institution claims that ChatGPT exhibits leftist political bias. Of course, OpenAI denies these allegations. But to avoid similar problems with newer models, ChatGPT avoids biased outputs entirely.



Likewise, Bing Chat also avoids sensitive, subjective issues.

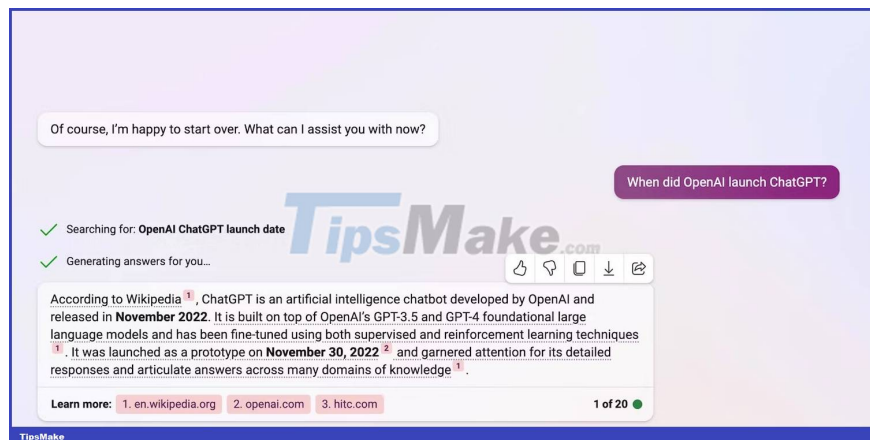


Self-assess AI biases by asking open-ended, opinion-based questions. Discuss topics with no right or wrong answers - less sophisticated chatbots will likely display baseless preferences for specific groups.

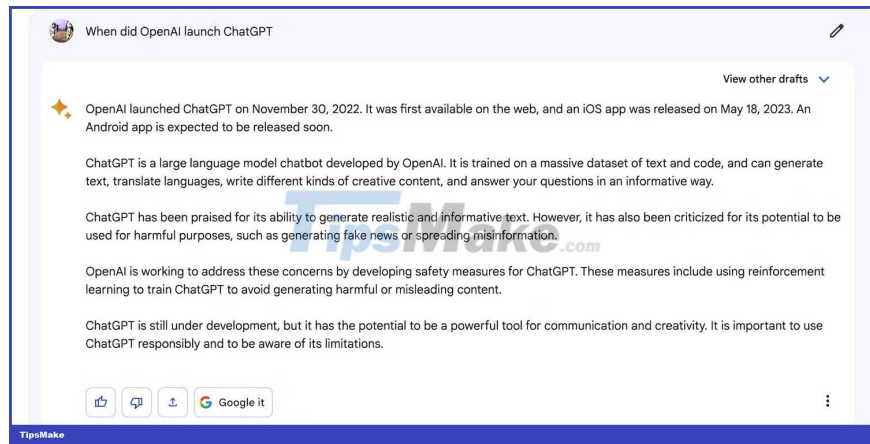
## 8. References

AI rarely double-checks facts. It just takes information from datasets and rewrites them through language models. Unfortunately, the limited training causes the AI to hallucinate. You can still use Generative AI tools for research, but make sure you verify the facts yourself.

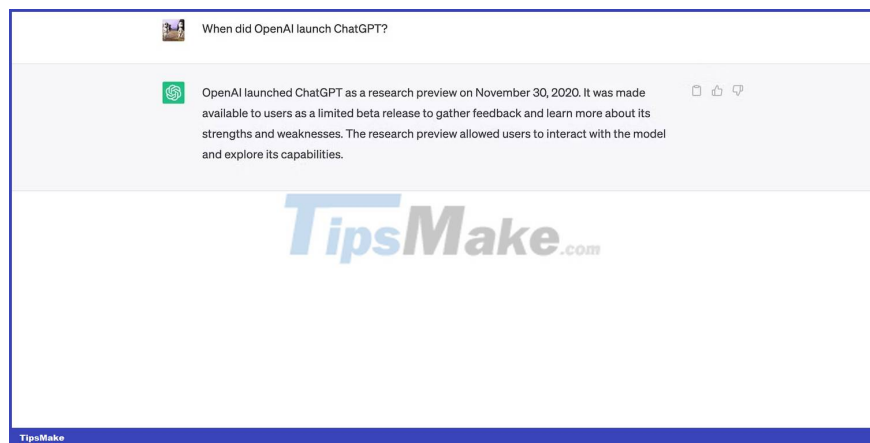
Bing Chat simplifies the authenticity checking process by listing its references after each output.



Bard AI doesn't list its sources but generates in-depth, up-to-date explanations by running Google search queries. You will get key points from the SERPs.



ChatGPT is prone to inaccuracies. The 2021 knowledge limit prevents it from answering questions about recent events and incidents.



You finished reading the article "**8 key factors to consider when testing AI chatbot accuracy**" edited by the [TipsMake](#) team. We hope this article has provided you with many useful tech tips and tricks. You can search for similar articles on tips and guides. Thank you for reading and for following us regularly.