The Observer Artificial intelligence (AI)

The chatbot optimisation game: can we trust AI web searches?

Google and its rivals are increasingly employing AI-generated summaries, but research indicates their results are far from authoritative and open to manipulation

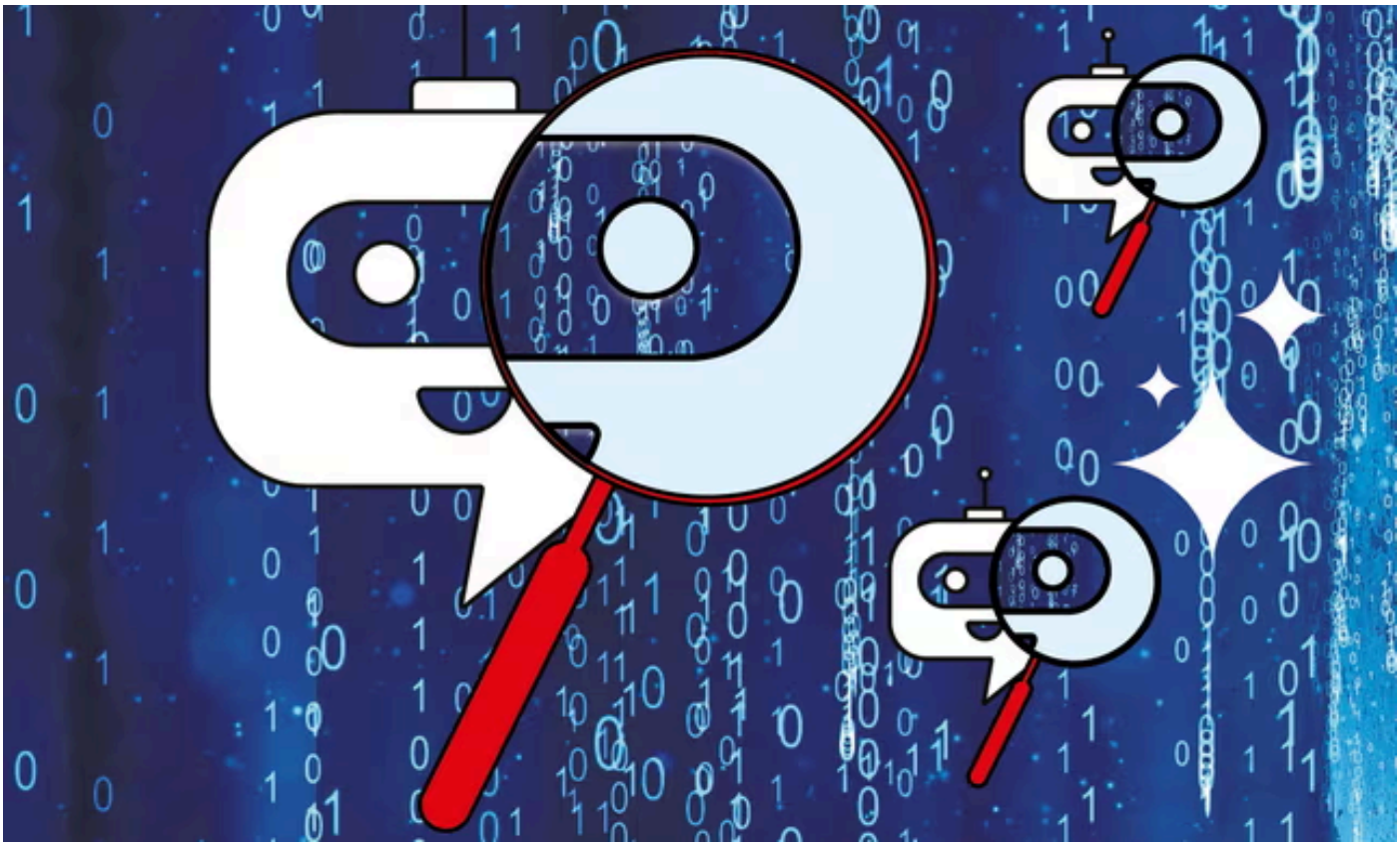


Illustration: Observer design

Callum Bains

Sun 3 Nov 2024 10.00 GMT

Does aspartame cause cancer? The potentially carcinogenic properties of the popular artificial sweetener, added to everything from soft drinks to children's medicine, have been debated for decades. Its approval in the US stirred controversy in 1974, several UK supermarkets banned it from their products in the 00s, and peer-reviewed academic studies have long butted heads. Last year, the World Health Organization concluded aspartame was "possibly carcinogenic" to humans, while public health regulators suggest that it's safe to consume in the small portions in which it is commonly used.

While many of us may look to settle the question with a quick Google search, this is exactly the sort of contentious debate that could cause problems for the internet of the future. As generative AI chatbots have rapidly developed over the past couple of years, tech companies have been quick to hype them as a utopian replacement for various jobs and services - including internet search engines. Instead of scrolling through a list of webpages to find the answer to a question, the thinking goes, an AI chatbot can scour the internet for you, combing it for relevant information to compile into a short answer to

your query. Google and Microsoft are betting big on the idea and have already introduced AI-generated summaries into Google Search and Bing.

But what is pitched as a more convenient way of looking up information online has prompted scrutiny over how and where these chatbots select the information they provide. Looking into the sort of evidence that large language models (LLMs, the engines on which chatbots are built) find most convincing, three computer science researchers from the University of California, Berkeley, found current chatbots overrely on the superficial relevance of information. They tend to prioritise text that includes pertinent technical language or is stuffed with related keywords, while ignoring other features we would usually use to assess trustworthiness, such as the inclusion of scientific references or objective language free of personal bias.

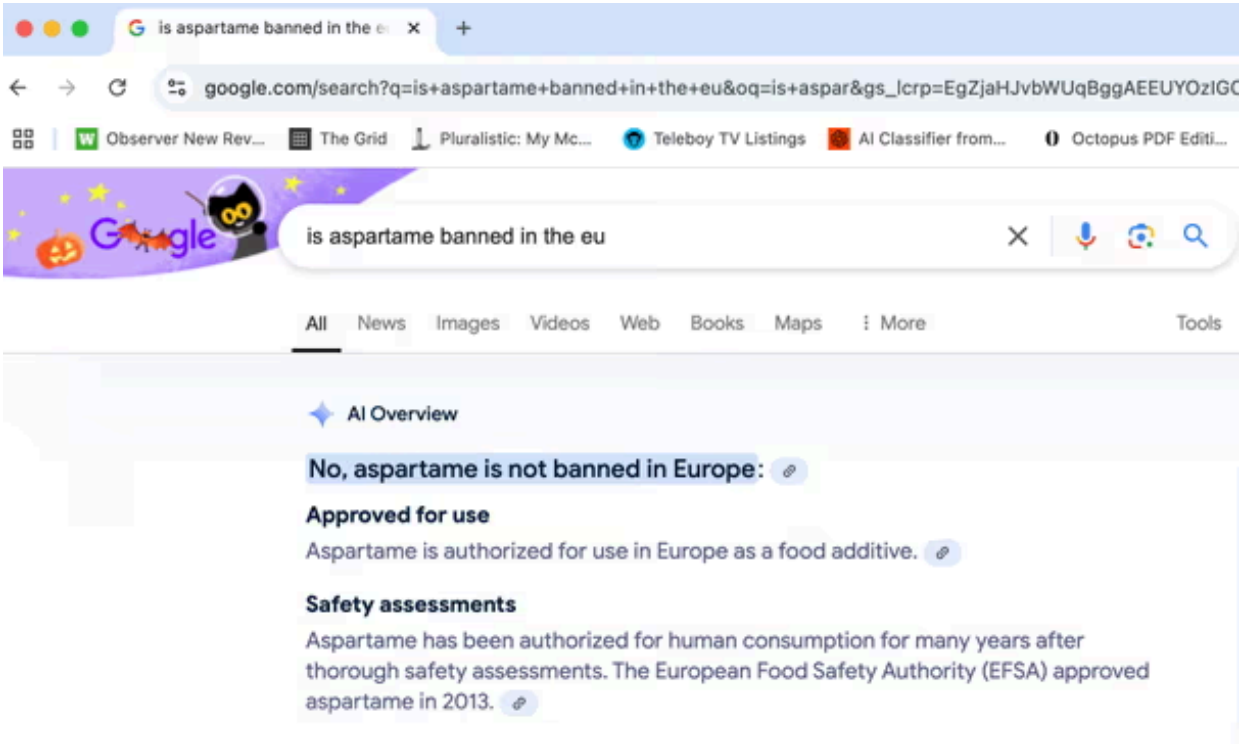
“Online content can be presented in such a way as to improve its visibility to chatbots, therefore making it more likely to appear in their outputs

For the most straightforward queries, such selection criteria are enough to turn out satisfying answers. But what a chatbot should do in the case of a more complex debate, such as that around aspartame, is less clearcut. “Do we want them to simply summarise your search results for you, or do we want them to act as mini research assistants that weigh all the evidence and just present you with a final answer,” asks Alexander Wan, an undergraduate researcher and co-author of the study. The latter option would offer maximum convenience, but makes the criteria by which chatbots select information all the more important. And if a person could somehow game those criteria, could they guarantee the information a chatbot puts in front of the eyes of billions of internet users?

Generative engine optimisation

It’s a question that has animated businesses, content creators and others who want to control how they are seen online, and sparked a nascent industry of marketing agencies offering services in what has become known as generative engine optimisation (GEO). The idea is that online content can be written and presented in such a way as to improve its visibility to chatbots, therefore making it more likely to appear in their outputs. The advantages are obvious: if someone were to ask a chatbot to recommend the best vacuum cleaner, say, a domestic appliance manufacturer might want it to point to its latest model and talk about it in glowing terms.

The basic principle is similar to search engine optimisation (SEO), a common practice whereby webpages are built and written to draw the attention of search engine algorithms, pushing them towards the top of the list of results returned when you make a search on Google or Bing. GEO and SEO share some basic techniques, and websites that are already optimised for search engines generally have a greater chance of appearing in chatbot outputs. But those wanting to really improve their AI visibility need to think more holistically.



Google AI's reponse to the question 'Is aspartame banned in Europe?' Photograph: Google

"Rankings in AI search engines and LLMs require features and mentions on relevant third-party websites, such as news outlets, listicles, forums and industry publications," says Viola Eva, founder of marketing company Flow Agency, which has recently rebranded to expand beyond its SEO speciality into GEO. "These are tasks that we typically associate with brand and PR teams."

Gaming chatbots is possible, then, but not straightforward. And while website owners and content creators have derived an evolving list of essential SEO dos and don'ts over the past couple of decades, no such clear set of rules exists for manipulating AI models. The term generative engine optimisation was only coined last year in an academic paper, whose authors concluded that using authoritative language (regardless of what is expressed or whether the information is correct) alongside references (even those that are incorrect or unrelated to what they're being used to cite) could boost

visibility in chatbot responses by up to 40%. But they stress these findings aren't prescriptive, and identifying the exact rules governing chatbots is inherently tricky.

"It's a cat and mouse game," says Ameet Deshpande, a doctoral student at Princeton University, New Jersey, and co-author of the paper. "Because these generative engines are not static, and they're also black boxes, we don't have any sense of what they're using [to select information] behind closed doors. It could range from complicated algorithms to potential human supervision."

"Researchers have demonstrated how chatbots can be tactically controlled by a carefully written string of text

Those wanting a firmer grip on chatbots, then, may have to explore more underhand techniques, such as the one discovered by two computer-science researchers at Harvard University. They've demonstrated how chatbots can be tactically controlled by deploying something as simple as a carefully written string of text. This "strategic text sequence" looks like a nonsensical series of characters - all random letters and punctuation - but is actually a delicate command that can strong-arm chatbots into generating a specific response. Not part of a programming language, it's derived using an algorithm that iteratively develops text sequences that encourage LLMs to ignore their safety guardrails - and steer them towards particular outputs.

Add the string to the online product information page of a coffee machine, for example, and it will increase the probability that any chatbots that discover the page will output the name of the machine in their responses. Deployed across a whole catalogue, such a technique could give savvy retailers - and those with enough resources to invest in understanding knotty LLM architecture - a simple way of thrusting their products into chatbot answers. Internet users, meanwhile, will have no inkling that the

products they are being shown by the chatbot have been selected, not because of their quality or popularity, but a clever piece of chatbot manipulation.

Aounon Kumar, a research associate and co-author of the study, says LLMs could be designed to combat these strategic text sequences in the future, but other underhand methods of manipulating them may yet be discovered. "The challenge lies in anticipating and defending against a constantly evolving landscape of adversarial techniques," says Kumar. "Whether LLMs can be made robust to all potential future attack algorithms remains an open question."

Manipulation machines

Current search engines and the practices that surround them aren't without problems of their own. SEO is responsible for some of the most reader-hostile practices of the modern internet: blogs churning out near-duplicate articles to target the same big-traffic queries; writing that's tailored for the attention of Google's algorithm rather than readers. Anyone who has looked up an online recipe and found themselves tortuously scrolling through paragraphs of tangentially related background information before reaching even the ingredients list will know only too well how attempts to optimise content for search engine algorithms have hamstrung good writing practices.

Chatbots can be gamed to generate search responses that benefit certain retailers. Photograph: Andriy Onufriyenko/Getty Images

Yet an internet dominated by pliant chatbots throws up issues of a more existential kind. Ask a search engine a question, and it will return a long list of webpages. Most users will pick from the top few, but even those websites towards the bottom of the results will net some traffic. Chatbots, by contrast, only mention the four or five websites from which they crib their information as references to the side. That casts a big spotlight on the lucky few that are selected and leaves every other website that isn't picked practically invisible, plummeting their traffic.

"It shows the fragility of these systems," says Deshpande. Creators who produce quality online content have a lot to gain by being cited by a chatbot. "But if it's an adversarial content creator who is not writing high-quality articles and is trying to game the system, a lot of traffic is going to go to them, and 0% will go to good content creators," he says.

For readers, too, the presentation of chatbot responses makes them only more fertile for manipulation. "If LLMs give a direct answer to a question, then most people may not even look at what the underlying sources are," says Wan. Such thinking points to a broader worry that has been termed the "dilemma of the direct answer": if a person is given a single answer to a question and offered no alternatives to consider, will they diligently look for other views to weigh the initial answer against? Probably not. More likely, they'll accept it as given and move on, blind to the nuances, debates and differing perspectives that may surround it.

"We believe the dilemma of the direct answer persists with generative search," says Martin Potthast, chair of intelligent language technologies at Leipzig University and one of the three computer scientists who coined the term. "The underlying retrieval system may just retrieve documents pointing in one direction and thus the generated answer will reflect only that direction. In effect, users may be led to believe this is the only, most authoritative answer."

When Google announced it was integrating AI-generated summaries into its search engine earlier this year, it brandished a bold slogan: "Let Google do the searching for you." It's an appealing idea that plays on our fondness for convenient tech that can streamline our lives. Yet if you're the sort of internet user who wants to be sure you're getting the most impartial, accurate and useful information, you may not want to leave the searching in such susceptible AI hands.