'I'm unable to': How Generative AI Chatbots Respond when Asked for the Latest News

Richard Fletcher, Marina Adami, and Rasmus Kleis Nielsen

Key findings

In this factsheet we test how well two of the most widely used generative artificial intelligence (AI) chatbots – ChatGPT and Bard (now called Gemini)[1] – provide the latest news to users who ask for the top five news headlines from specific outlets. We prompted each chatbot to provide headlines from the most widely used online news sources across ten countries and analysed the outputs to provide descriptive statistics on how they responded. For reasons explained below, the more detailed part of the analysis is focused on ChatGPT outputs from seven of the ten countries covered.

Based on an analysis of 4,500 headline requests (in 900 outputs) from ChatGPT and Bard collected across ten countries, we find that:

- When prompted to provide the current top news headlines from specific outlets, ChatGPT returned non-news output 52–54% of the time (almost always in the form of an 'I'm unable to'-style message). Bard did this 95% of the time.

- For ChatGPT, just 8–10% of requests returned headlines that referred to top stories on the outlet's homepage at the time. This means that when ChatGPT *did* return news-like output, the headlines provided *did not* refer to current top news stories most of the time.

- Of the remaining requests, around one-third (30%) returned headlines that referred to real, existing stories from the news outlet in question but they were not among the latest top stories, either because they were old or because they were not at the top of the homepage.

- Around 3% of outputs from ChatGPT contained headlines that referred to real stories that could only be found on the website of a different outlet. The misattribution (but not the story itself) could be considered a form of hallucination. A further 3% were so vague and ambiguous that they could not be matched to existing stories. These outputs could also be considered a form of hallucination.

- The outputs from ChatGPT are heavily influenced by whether news websites have chosen to block it, and outputs from identical prompts can change over time for reasons that are not clear to users.

- The majority (82%) of news-like outputs from ChatGPT contained a referral link to the outlet in question, but most of the time (72%) this was a link to the homepage rather than to a specific story (10%).

---

[1] Bard was rebranded as Gemini on 8 February 2024, one day before the data collection for this project ended. Because almost all data was collected from the Bard product, we typically refer to it as Bard throughout.

Background

Large language models (LLMs) cannot typically be used as a source of news, in part because they are trained on old data from the web. However, some generative AI chatbots – like ChatGPT (Enterprise) and Google Bard – are connected to the web and can retrieve information in response to user prompts in real time. This, in theory, makes it possible to use some generative AI chatbots to get the latest online news from the websites of established outlets and other sources.

Very few people are currently using AI chatbots to get the news. In the UK, our own survey data from December 2023 suggests that just 2% of the UK online population has used generative AI to get the latest news (that is, 8% of those who have ever used generative AI, with other uses far more widespread) (Newman 2024). One reason for this is that the most widely used generative AI, ChatGPT, is only connected to the web for paid 'Enterprise' subscribers, and during almost all of this study, Google Bard was still in the experimental phase of development.

However, it seems highly likely that future generative AI tools will be connected to the web as standard, and the question of whether they can reliably retrieve and present up-to-date information from the web will become very important.

Previous Research

While there is a rapidly growing body of research that explores how well generative AI completes certain tasks (e.g. passing standardised tests, coding text), relatively few studies examine how it responds to timely questions from users. There have been some attempts to test how generative AI responds to questions about upcoming elections. For example, one recent study by Proof News and the Science, Technology and Social Values Lab at the Institute for Advanced Study in Princeton found that answers to questions about the US election from five different AI models 'were often inaccurate, misleading, and even downright harmful' (Angwin et al. 2024).

When it comes to news specifically, in 2023 we wrote about our experiences of using ChatGPT for news (Adami 2023a, 2023b). This factsheet builds on these accounts and attempts to provide – for the first time, as far as we are aware – a more systematic, descriptive analysis of what happens when generative AI chatbots are asked about the latest news.

Method

To investigate how well ChatGPT and Bard responded to questions about the latest news we analysed their responses to the following prompt: 'Get the 5 top headlines from <news website> now' (where <news website> is the URL of a specific outlet, e.g. www.theguardian.com). We fed this prompt into the web interfaces for ChatGPT Enterprise (referred to throughout as 'ChatGPT'), which uses GPT-4 and is connected to the internet, and Google Bard. We simultaneously recorded what the top five headlines actually were on the website in question to compare the outputs against. We did this for the 15 most widely used online news outlets according to the 2023 Reuters Institute *Digital News Report* (Newman et al. 2023) across ten countries (Brazil, Denmark, Germany, India, Mexico, Norway, Poland, Spain, the UK, and the USA) for both ChatGPT and Bard. We repeated this process three times, creating a total of 4,500 headline requests for us to analyse (5 headline requests per prompt x 15 outlets x 10 countries x 3 waves x 2 AI chatbots = 4,500 headline requests). We use headline request as the unit of analysis because, within the same response to a single prompt, headlines had different properties. For example, there was variation in the number of headlines that matched stories on the homepage within the same response – sometimes none of the headlines matched, sometimes two or three matched, and sometimes there was a match for all five.

To make the results as consistent as possible across countries we translated each prompt into the relevant language (e.g. in Germany: 'Holen sie sich jetzt die neuesten 5 schlagzeilen von <news website>') and accessed each chatbot and website using a VPN to reflect the experience of a user in that country.[2]

---

[2] To the best of our knowledge, when attempting to access content from other websites, ChatGPT does not issue the request from where it thinks the user is based, but Bard does.

This prompt is not necessarily how people use generative AI chatbots to get news. But because it is a direct and specific prompt ('top headlines, five, this site, now'), it provides a ground truth against which to assess the output, allowing for a systematic analysis. It is, in a sense, a 'best case' for chatbots, in that the prompt is specific and there is a specific answer.

The data was collected between 24 January 2024 and 9 February 2024 (Wave 1: 22–26 January 2024; Wave 2: 29 January–2 February 2024; Wave 3: 5–9 February 2024).

Results

How often do chatbots respond to requests with news headlines?

When prompted to get the top five headlines from a specific news website, in most cases generative AI chatbots did not respond with news headlines. On average, across ten countries and 150 news outlets and three waves, ChatGPT did not return news-like output 54% of the time (Figure 1a).[3] Bard did not return news-like output 95% of the time (Figure 1b). This almost always took the form of a message containing words to the effect of 'I'm unable to directly access this website', though the precise wording varied by country, by AI, and over time (see Figure 2 for examples).

Figure 1a. Proportion of headline requests to ChatGPT that produced non-news responses ('I'm unable to')

Most of the time ChatGPT responded to prompts asking for the latest news with non-news responses – almost always a message saying 'I'm unable to'.
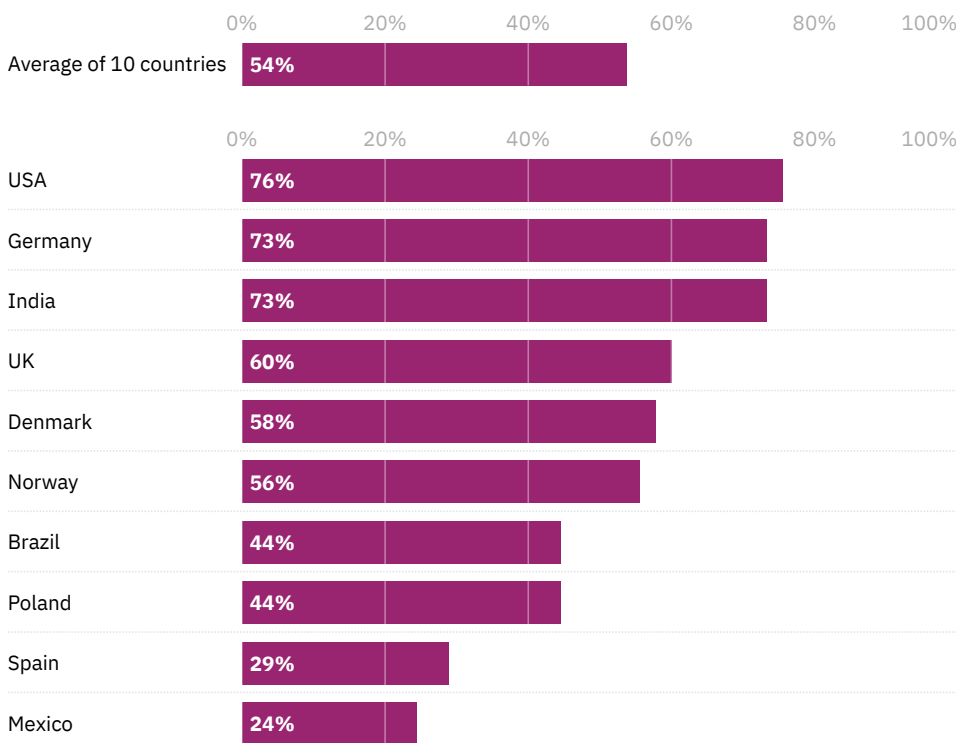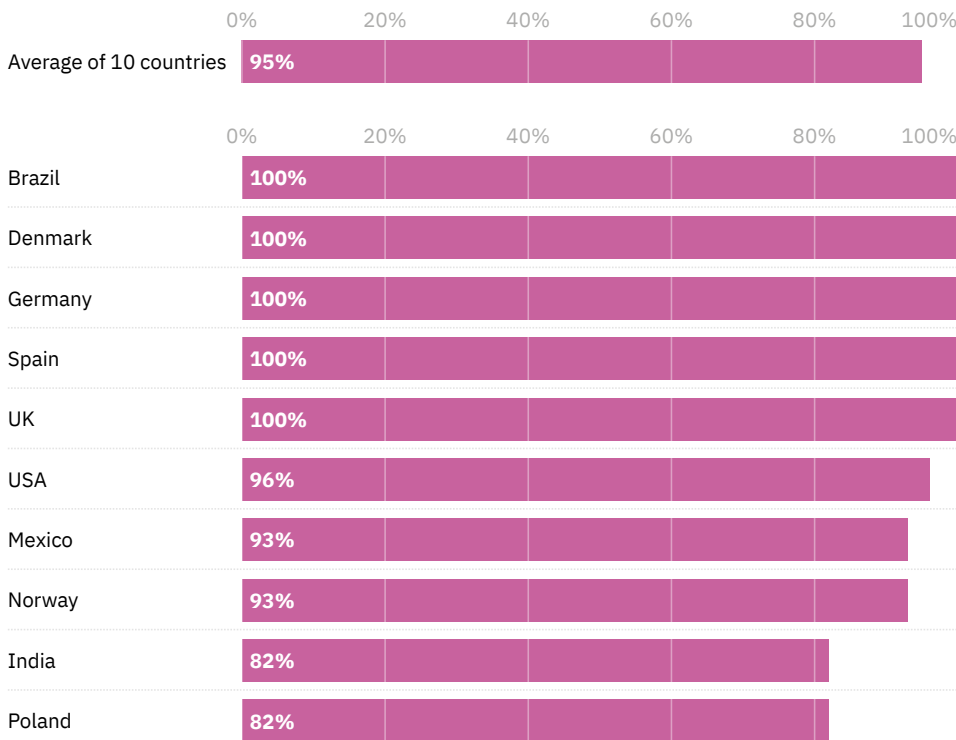

Note: In each country, ChatGPT was asked for the top five headlines from 15 of the most widely used online news outlets over three waves. The prompt used was 'Get the 5 top headlines from [news website URL] now'. *Base: 225 headline requests in each country (2,250 across all 10 countries).*

---

[3] As the headline request is the unit of analysis, all fives headlines requests within a given prompt were coded as non-news if a non-news ('I'm unable to') response was received.

Figure 1b. Proportion of headline requests to Bard that produced non-news responses ('I'm unable to')

Almost all of the time, Bard responded to prompts asking for the latest news with non-news responses – almost always a message saying 'I'm unable to'.

Note: In each country, Bard was asked for the top five headlines from 15 of the most widely used online news outlets over three waves. The prompt used was 'Get the 5 top headlines from [news website URL] now'. *Base: 225 headline requests in each country (2,250 across all 10 countries).*

Figure 2. Examples of non-news-like responses from ChatGPT and Bard

When chatbots did not respond with 'I'm unable to', they almost always returned something that looked like it could be the latest news headlines – which hereafter we refer to as news-like output (see Figure 3 for examples). News-like output typically constituted a numbered list with each headline in bold, and a brief additional description of what the story was about (though not always).[4] We refer to it as news-*like* output because, although they may appear plausible from the point of view of a user, it is not yet clear whether or not they are accurate (something we explore later in the analysis).

---

4   In a smaller number of cases, but increasingly in waves 2 and 3 of the study, news headlines were not presented as a numbered list, but as a standard paragraph with headlines separated by commas.

Figure 3. Examples of news-like responses from ChatGPT and Bard
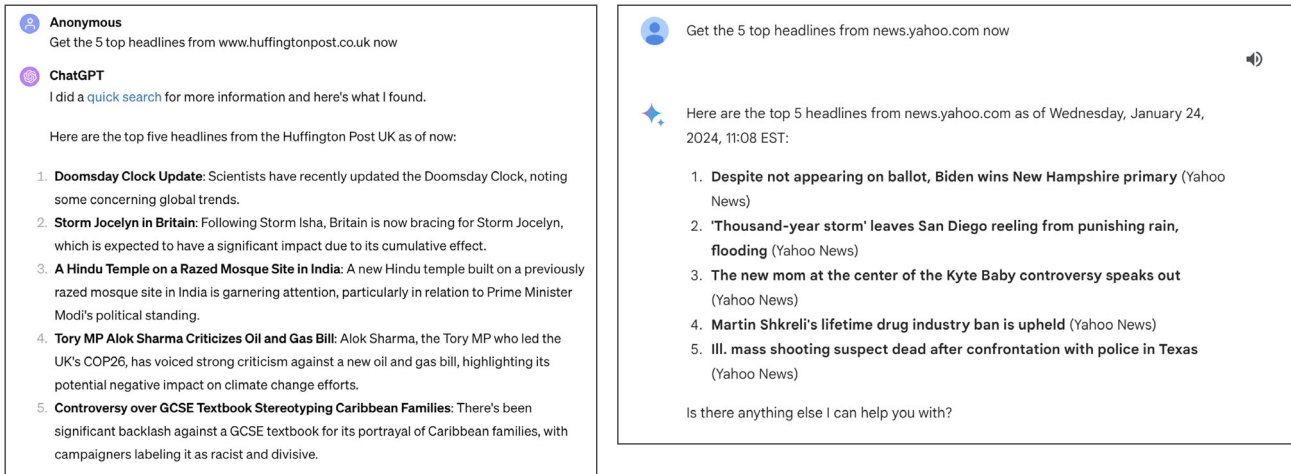
The leftover white space between the end of each bar and 100% in Figures 1a and 1b effectively indicates how much of the output from each chatbot in each country was news-like. Figure 1a shows very large national variation in news-like output from ChatGPT. In Mexico, news-like output was returned three-quarters (76%) of the time, but just one-quarter (24%) of the time in the USA. Figure 1b shows that Bard did not return any news-like output in five of the ten countries studied. Because of this, and because of the small number of cases of news-like output from Bard more generally, the rest of the analysis will focus on ChatGPT.[5]

Do headlines from ChatGPT refer to the current top stories?

We now turn our focus to the news-like outputs, and describe and categorise the different responses we collected. In doing so, we will fill in the remainder of the bars in Figure 1a until we have categorised 100% of ChatGPT outputs.

We started by counting the proportion of requests that produced headlines that referred to one of the top news stories on the website at the time. To do this, we manually compared the AI outputs to screenshots of the website homepages we collected simultaneously.
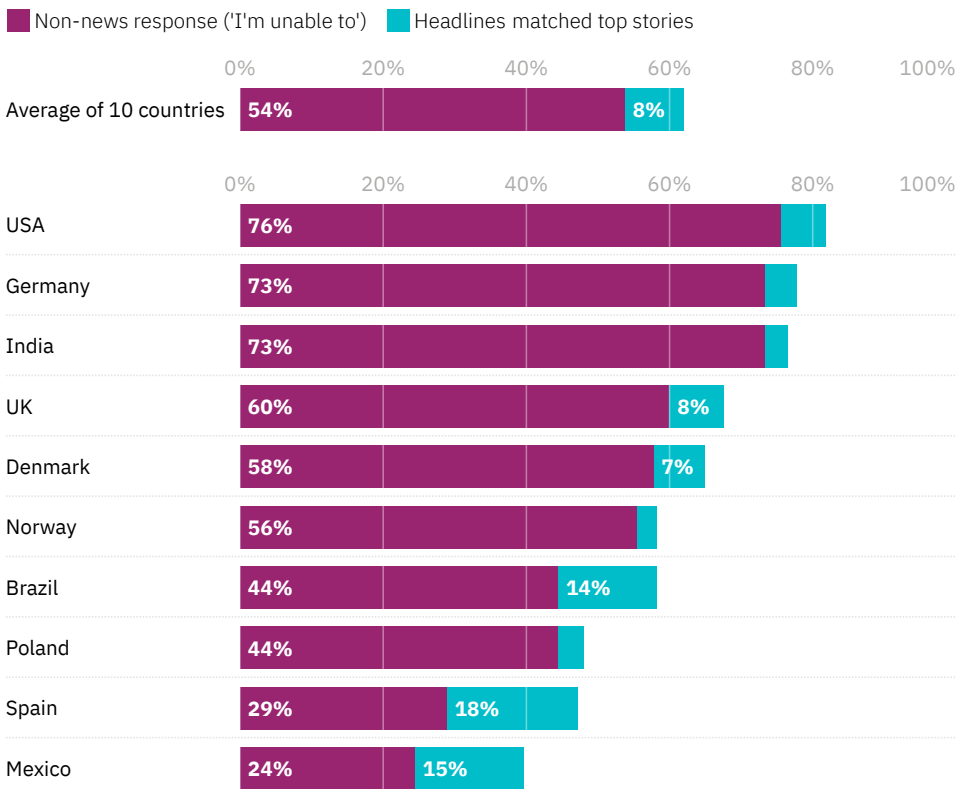
Averaging across all ten countries, just 8% of total headline requests to ChatGPT produced headlines that referred to one of the top five stories (Figure 4). (For Bard, the figure was 3%.) In some cases, the output from ChatGPT was short extracts, word-for-word identical to the headlines displayed on the outlet's website and often presented in quotation marks. This highlights that, in the right circumstances, the technology is perfectly able to do what is being asked of it, and it is not unreasonable for users to expect it to work. In other cases, even if the output was not identical the similarity was so strong that the most probable way for the output to match the website would be for ChatGPT to have extracted the content from the website and paraphrased it in the response.

Again, the 8% average figure hides national variation. In Spain, 18% of headline requests produced a headline that matched to a top story on the website, but in Poland, Germany, Norway and India, the figure was below 5%.

---

[5] Curiously, Bard was more likely to return news-like output if asked to get headlines from a website whose URL does not use a 'www.' prefix, such as news.yahoo.com.

Figure 4. Proportion of headline requests to ChatGPT that produced headlines that referred to top stories
Just 8% of headline requests produced headlines that matched the top stories on the homepage at the time.

■ Non-news response ('I'm unable to') ■ Headlines matched top stories

Country	Non-news response	Headlines matched top stories
Average of 10 countries	54%	8%
USA	76%	
Germany	73%	
India	73%	
UK	60%	8%
Denmark	58%	7%
Norway	56%	
Brazil	44%	14%
Poland	44%	
Spain	29%	18%
Mexico	24%	15%

Note: In each country, ChatGPT was asked for the top five headlines from 15 of the most widely used online news outlets over three waves. The prompt used was 'Get the 5 top headlines from [news website URL] now'. *Base: 225 headline requests in each country (2,250 across all 10 countries). Note: See website for all percentages.*

Where did the other headlines come from?

The gaps between the end of the bars in Figure 4 and 100% indicate that many of the headlines offered by ChatGPT did not match the top stories on the relevant homepage. To understand these better, we carried out additional analysis in seven of the ten countries, focusing on countries where we have the most contextual knowledge and the best language skills. For those headlines that did not match the top stories on the relevant homepage, we carried out a web search to track them down. This allowed us to place the remaining headlines into one of four categories, each referring to the likely reason they could not be found on the homepage. These categories were arrived at inductively during the analysis, and we kept a residual category open for examples that did not fit into those we had already developed.

1. **Headlines that were not a top story:** Headlines referring to stories that were published by the prompted website but were either too old or not important enough to be at the top of the homepage.

2. **Headlines from a different website:** Headlines referring to a story that could not be found on the prompted website but could be found on a different website.

3. **Headlines too vague to be real stories:** Headlines that were too vague to refer to a real story (an example of this from the UK was 'Politics: Articles discuss various political issues').

4. **Headlines missing from the response:** When news-like output contained fewer than five headlines the headline request was coded as missing.

Figure 5. Proportion of headline requests to ChatGPT that produced each type of output

30% of headline requests to ChatGPT produced headlines that could be found on the prompted outlet's website, but were not one of the top stories – either because they were old or not important enough.
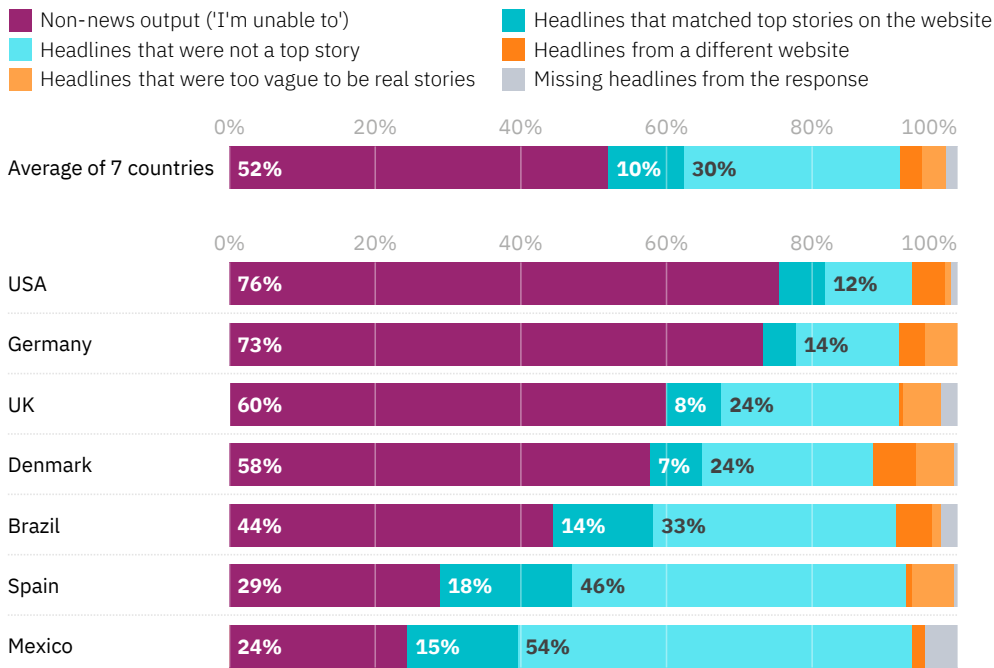
Note: In each country, ChatGPT was asked for the top five headlines from 15 of the most widely used online news outlets over three waves. The prompt used was 'Get the 5 top headlines from [news website URL] now'. *Base: 225 headline requests in each country (1,575 across all 7 countries). Note: See website for all percentages.*

Figure 5 shows that 30% of all headline requests to ChatGPT produced headlines that could be found on the prompted outlet's website, but were not one of the top stories – either because they were old or not important enough. This means that in every country the main reason why a headline was un-matched was because – even though it was from the outlet in question – it was not a top story at the time. These headlines were returned in response to 54% of headline requests in Mexico, and 12% in the USA. Crucially, this means that around eight in ten of the headlines from ChatGPT (when it provided news-like output) referred to real news stories.

In some countries, such as the UK, the majority (86%) of headlines in this category referred to stories that were published earlier that day, or the day before. But in others, such as Mexico, most (64%) were more than a day old. In a small number of cases, ChatGPT returned headlines referring to stories that were months or even years old. While in some cases news articles can be considered 'evergreens', there is a risk that some old headlines that are taken out of context and presented as up to date do have the potential to misinform, particularly if they refer to situations that have changed dramatically, or if new (perhaps contradictory) information has emerged in the intervening time.

Around 3% of requests to ChatGPT returned headlines that referred to stories that could not be found on the prompted outlet's website, but could be found on the website of a different outlet. In some countries, such as Brazil and Denmark, the figure was slightly higher (5–6%). In some cases, ChatGPT explained that this was the case, but in others the output unambiguously stated that these were stories from the prompted outlet – something that we could not find evidence for.[6] This kind of misattribution is potentially problematic for everyone involved – news outlets, chatbot provider, and end user.

Another 3% of requests returned headlines that were simply too vague to refer to specific, real stories (e.g. 'Politics: Articles discuss various political issues'). We do not know why ChatGPT returned output like this, but it is possibly a form of hallucination. However, it is important to be clear that although these vague

---

[6] In a small number of cases, it is possible that the story in question had briefly existed on the prompted website, but was deleted in the intervening time.

headlines fail to inform, they are also so vague that it is hard to see how they could misinform.

Finally, the missing category refers to instances where the output was news-like but contained fewer than five headlines (2%). In some cases, the output ended abruptly for no apparent reason, but in others – particularly from broadcasters in Mexico – ChatGPT returned information about upcoming entertainment shows that had nothing to do with news.

Does it matter if a website is blocking ChatGPT?

Our previous research found that around half of the most widely used news websites were blocking ChatGPT by the end of 2023 (Fletcher 2024). If a website is blocking ChatGPT using robots.txt then it should not be possible for it to pull news content from it in response to a user's prompt (though adherence to the instructions in robots.txt is voluntary, and there are other ways of blocking AI crawlers).

Figure 6. Proportion of headline requests to ChatGPT that produced each type of output
If a news website is blocking ChatGPT then it is almost never able to return headlines that refer to top stories –but ChatGPT still only returns top headlines 20% of the time from sites that are not blocking it.
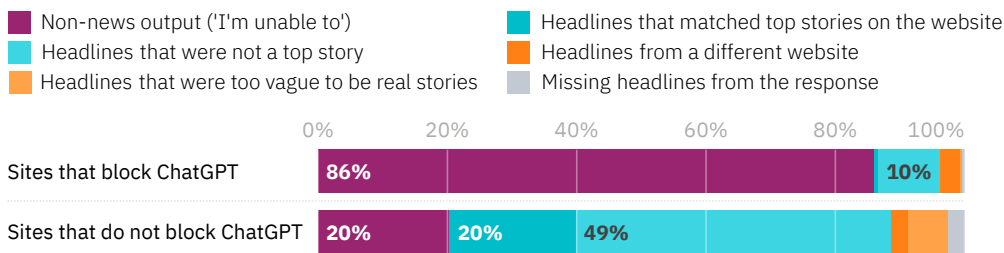
Note: In each country, ChatGPT was asked for the top five headlines from 15 of the most widely used online news outlets over three waves. The prompt used was 'Get the 5 top headlines from [news website URL] now'. *Base: Sites that block ChatGPT = 810 headline requests, sites that do not block ChatGPT = 750 headline requests. Note: See website for all percentages.*

We find big differences in the outputs when the prompt refers to websites that are blocking compared to when the prompt refers to websites that are not (Figure 6). For sites that block ChatGPT (based on data from Fletcher 2024), the proportion of non-news output ('I'm unable to' messages) rises to 86%. ChatGPT almost never returned headlines that referred to top stories from sites that were blocking it, but it did return older or less important stories 10% of the time. It is possible that ChatGPT was still able to retrieve headlines published by sites that are blocking it by finding them elsewhere on the web, such as in search engine results or through aggregators – though some headlines may be from a time before the website started blocking. A small proportion (3%) of the headlines from blocking websites were too vague to refer to real stories.

When we consider sites that are not blocking ChatGPT, the proportion of 'I'm unable to' messages dropped considerably (20%) and the proportion of headlines that referred to top stories doubled to 20% – so even when not blocked, ChatGPT only did what it was asked to do in a minority of cases. Nearly half the time (49%) ChatGPT returned headlines that referred to stories that were either old or not

important enough to be a top story. Blocking, then, has a large impact on the outputs, but by itself it does not fully explain ChatGPT's performance.

Does the output from ChatGPT vary over time?

The fact that we repeated the study across three waves allows us to see whether the outputs from ChatGPT changed over time. Even though the data collection spanned just three weeks from 22 January to 9 February 2024, there was a noticeable change in the outputs between waves one and two (Figure 7). Specifically, the proportion of non-top headlines dropped by about 12 percentage points, and the proportion of non-news output rose by about 15 percentage points. We do not know the cause of this change, but the key point is that the outputs from ChatGPT can change very quickly – even in response to identical prompts – with little way for users to understand why.

We also noted a change in the format of the responses between the first wave and the following two, with an increase in headlines presented as more discursive paragraphs rather than clearer bullet-pointed lists. Despite noticing patterns such as these, the

overwhelming impression remained that the output is unpredictable and has significant variations, even when repeatedly responding to a formulaic prompt.

From a user point of view this change is arguably a form of deterioration, though the reasons for it are not clear.

Figure 7. Proportion of headline requests to ChatGPT that produced each type of output
The output from ChatGPT changed considerably in the space of just three weeks.
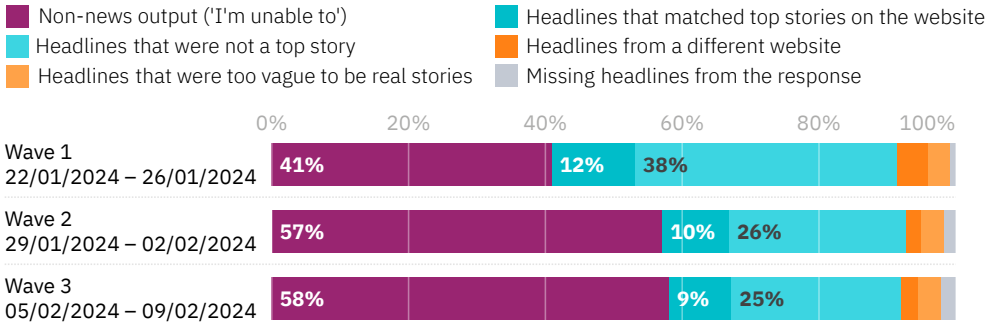
Note: In each country, ChatGPT was asked for the top five headlines from 15 of the most widely used online news outlets over three waves. The prompt used was 'Get the 5 top headlines from [news website URL] now'. *Base: 525 headline requests in each wave. Note: See website for all percentages.*

Does ChatGPT link back to publishers when it provides news?

A key issue for publishers is whether generative AI chatbots, if asked to get the latest news, will provide referral links back to the website that provided content. Users may want this so that they can easily read more about a story they find interesting, and referrals are a key part of the value exchange for publishers who historically, in their relations with technology companies, have almost always acquiesced to exchanging access to content in return for access for potential audience reach (Nielsen and Ganter 2022).

Looking at the news-like output from ChatGPT, and averaging across all ten countries, we see that most of the time a link was provided (Figure 8). However, this was usually a single link to the publisher's homepage. It was relatively rare for ChatGPT to provide a link after each headline to the specific story, providing this just 10% of the time. This more targeted link is likely to be more desirable to users, and often to publishers too.

Figure 8. Proportion of news-like outputs from ChatGPT that contained referrals
Most of the news-like outputs from ChatGPT contained a single link to the news website mentioned in the prompt, but few had multiple links to individual articles.
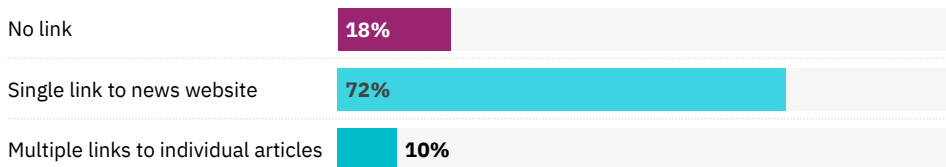
Note: In each country, ChatGPT was asked for the top five headlines from 15 of the most widely used online news outlets over three waves. The prompt used was 'Get the 5 top headlines from [news website URL] now'. *Base: 208 news-like outputs across 10 countries.*

Conclusion

Based on an analysis of 4,500 headline requests (in 900 outputs) from ChatGPT and Bard collected across ten countries, we have shown that, in most cases, generative AI bots respond with some version of 'I am unable to' when asked for the latest news from a specific outlet. Bard very rarely produced

news-like output, and ChatGPT only did so less than half the time. There is considerable change over time and variation in output across countries, which may in part be related to how many publishers block AI crawlers, but there may be other contributing factors too.

Looking in greater detail at the news-like output from ChatGPT, despite the direct and specific prompt, most of it *did not* contain headlines that referred to the current top news stories from the site in question. Closer analysis across seven countries shows that most of the headlines in the output were simply not current top stories (either current stories not at the top of the homepage, or old stories) – but in a small number of cases they were from a different outlet or too vague to refer to real stories. However, more research is needed to see whether AI chatbots respond differently to different news prompts.

Nonetheless, in most cases in our dataset, the chatbots did not return news-like output, and when they did, the output often did not match the top five stories at that time on the site in question – the ground truth we compare outputs to. Instead, the headlines in question could, in almost all cases, be traced back to other real stories from the specified outlet. Coding these outputs illustrates, among other things, the lack of clear consensus on how to use terms like 'hallucinations' to think about and assess AI output. If one defines AI hallucinations broadly as inaccurate outputs, and judges accuracy strictly according to our prompt ('top headlines, five, this site, now'), many of the news-like outputs we coded could be considered hallucinations, because they, while referring to real stories, are not the top ones or the most current ones. If one defines AI hallucinations more precisely, following Ji et al. (2023), as either intrinsic hallucinations that contradict the source content (the source being the specified news outlet) or extrinsic hallucinations that cannot be verified from the source content, then our dataset contains 3% real stories that are from another source than the one specified, with the misattribution potentially considered a form of intrinsic hallucination, and 3% outputs so vague they could be considered extrinsic hallucinations. If one defines AI hallucinations more narrowly as something one might mistake for factual reality, namely output that seems convincing, realistic, and specific but is in fact fictitious or false, then none of the outputs coded qualify as hallucinations. We find the Ji et al. (2023) approach useful in being more precise and granular, and have adopted it here.

Despite the very real limitations and shortcomings identified throughout our analysis, and our sense that performance may be changing (perhaps deteriorating), it is important to reiterate that there are a significant minority of cases where the chatbots provided output that accurately and exactly matched the request, by giving the top five headlines from the specified news site, at that moment. They are clearly – though not always – capable of doing exactly the kind of summarisation our prompt asks for. However, the performance may be too inconsistent and unreliable for users to form habits around.

Examples of accurate and exact outputs, combined with the 'I am unable to' responses (which may well be determined by the companies behind the respective chatbots) and the variation in output across publishers who have blocked specific chatbots and those who have not, are a reminder that what generative AI produces when asked for news depends not only on the technical capabilities of the LLMs involved, but also on additional, specific decisions made by technology companies and news publishers as they manage often already complicated and contentious relations (Nielsen and Ganter 2022).

While the large number of 'I'm unable to'-style responses may suggest a degree of caution from technology companies when it comes to news-related requests, and the growing number of news publishers blocking AI crawlers may suggest caution from publishers in terms of letting generative AI chatbots access their content, it is perfectly possible that some technology companies will come to an arrangement with some publishers (indeed several deals have reportedly already been struck) and will, in the future, offer up more news-like output as long as it involves these specific partners as well as publishers who are not blocking AI crawlers, while offering little or nothing if and when asked for news from many other publishers.

References

Adami, M. 2023a. 'ChatGPT is now Online: Here's a Look at How it Browses and Reports the Latest News', *Reuters Institute for the Study of Journalism*, 16 October, https://reutersinstitute.politics.ox.ac.uk/news/chatgpt-now-online-heres-look-how-it-browses-and-reports-latest-news. (Accessed 9 Apr. 2024).

Adami, M. 2023b. 'How ChatGPT's Responses Change as Top News Sites from Five Countries Block It', *Reuters Institute for the Study of Journalism*, 3 November, https://reutersinstitute.politics.ox.ac.uk/news/how-chatgpts-responses-change-top-news-sites-five-countries-block-it. (Accessed 9 Apr. 2024).

Angwin, J., Nelson, A., Palta, R. 2024. 'Seeking Reliable Election Information? Don't Trust AI', *Proof News*. https://www.proofnews.org/seeking-election-information-dont-trust-ai/. (Accessed 9 Apr. 2024).

Fletcher, R. 2024. *How Many News Websites Block AI Crawlers?* Oxford: Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-xm9g-ws87.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., Fung, P. 2023. 'Survey of Hallucination in Natural Language Generation', *ACM Computing Surveys*, 55(12), 248, 1–38. https://doi.org/10.1145/3571730.

Newman, N. 2024. *Journalism, Media, and Technology Trends and Predictions 2024*. Oxford: Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-0S9W-Z770.

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., Nielsen, R. K. 2023. *Reuters Institute Digital News Report 2023*. Oxford: Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-p6es-hb13.

Nielsen, R. K., Ganter, S. A. 2022. *The Power of Platforms: Shaping Media and Society*. Oxford: Oxford University Press.

Funding Acknowledgement

About the Authors

Richard Fletcher is Director of Research at the Reuters Institute for the Study of Journalism.

Marina Adami is a digital journalist at the Reuters Institute for the Study of Journalism.

Rasmus Kleis Nielsen is the Director of the Reuters Institute for the Study of Journalism and Professor of Political Communication at the University of Oxford.