

A woman with short brown hair, wearing a green and white vertically striped button-down shirt, is looking down at a black smartphone she is holding in her hands. The background is a dark, textured surface.

EBU

OPERATING EUROVISION AND EURORADIO

B B C

News Integrity in AI Assistants

An international PSM study

Contents

Foreword	3
Introduction	5
List of participating organizations	7
Approach	7
High-level findings	9
Detailed findings	14
Accuracy	14
IN FOCUS: Have assistants improved?	19
Comparing the two rounds of BBC research	
Accuracy of direct quotes	21
IN FOCUS: The problem of over-confidence	24
Sourcing	27
IN FOCUS: Gemini's issues with sourcing	34
Opinion vs fact	37
IN FOCUS: Comparing countries and languages	40
Editorialization	44
Context	47
IN FOCUS: How AI assistants treat PSM content	50
Assistant response styles	52
Conclusion	55
Authors	57
References	58
Appendix 1: List of participating organizations	59
Appendix 2: Methodology	60
News questions	60
AI assistants	61
Response generation	63
Journalist reviews	63
Data and quality assurance	64
Appendix 3: Results	66
Rating summary statistics	66
Assistant data	67

Foreword

This report is one of the largest cross-market evaluations of its kind. Working with the European Broadcasting Union (EBU), 22 Public Service Media (PSM) organizations – across 18 countries and 14 languages – assessed how leading AI assistants answer questions about news and current affairs.

The research built on an earlier study by the BBC¹, which highlighted inaccuracies and errors in AI assistants' output. We wanted to know if the assistants had improved and if the issues we had identified were isolated or systemic.

What have we found?

First – there have been improvements since the earlier BBC study. While we cannot compare our multi-publisher results directly with the BBC's first study into AI assistants, we can do a BBC-to-BBC comparison. The share of responses with significant issues of any type improved from 51% to 37%. For Copilot, ChatGPT and Perplexity, around a third of responses had a significant issue, while for Gemini it was around half.

Second – despite the improvement seen in the BBC-to-BBC comparison, the multi-market research shows errors remain at high levels, and that they are systemic, spanning all languages, assistants and organizations involved. Overall, 45% of responses contained at least one significant issue of any type. Sourcing is the single biggest cause of significant issues (31%). Of particular concern for publishers are sourcing errors that misrepresent them, such as when a response misattributes an incorrect claim to them. Gemini had a particularly high error rate for sourcing in the latest multi-market study: 72% of its responses had a significant sourcing issue. All other assistants were below 25%.

So, there has been progress in some areas, but there is much more to do. Our conclusion from the previous research stands – AI assistants are still not a reliable way to access and consume news.

1. BBC (2025), Representation of BBC News content in AI Assistants

And yet, many people *do* trust AI assistants to be accurate. [separate BBC research published at the same time as this report](#) shows that just over a third of UK adults say they completely trust AI to produce accurate summaries of information. This rises to almost half of under 35s. That misplaced confidence raises the stakes when assistants are getting the basics wrong. These shortcomings also carry broader consequences: 42% of adults say they would trust an original news source less if an AI news summary contained errors, and audiences hold both AI providers and news brands responsible when they encounter errors. The reputational risk for media companies is great, even when the AI assistant alone is to blame for the error.

At the same time, the evidence is mounting that “answer-first” experiences, such as AI assistants like Perplexity, and Google’s AI Overviews may be diverting users away from trusted and authoritative sources of information. The *Financial Times*, a British newspaper, has said it has seen a decline of 25% to 30% in traffic from readers arriving via search¹.

If AI assistants are not yet a reliable way to access the news, but many consumers trust them to be accurate, we have a problem. This is exacerbated by AI assistants and answer-first experiences reducing traffic to trusted publishers.

What needs to change?

Foreword

First and foremost, AI developers need to take this issue seriously and rapidly reduce errors, in particular accuracy and sourcing errors. They have not prioritized this issue and must do so now. They also need to be transparent by regularly publishing their results by language and market.

To help, we are publishing a [News Integrity in AI Assistants Toolkit](#) to support AI developers. It offers a guide to help improve how AI assistants approach news content. It contains a practical checklist of “what good looks like” and a detailed taxonomy of the errors that AI assistants make in relation to news that need to be monitored and fixed.

Second – publishers need greater control over whether and how their content is used by AI assistants. Unauthorized use of content in AI assistants is still a widespread problem that is compounded by errors. Where content is authorized for use, we need clearer

1. Sweney (2025), ‘Existential crisis’: how Google’s shift to AI has upended the online news model

Foreword

attribution and citations in an agreed format with prominent links to original content.

Third – AI developers must be held to account for the quality and impact of their products, particularly as consumers make more use of them. While industry-led solutions are preferable, policymakers and regulators should urgently consider how the news content in AI assistants can be improved further. This includes the prominence and findability of trusted PSM content – an area where some regulators have already identified problems.

Fourth – audiences need to better understand the limitations of AI assistants as well as their benefits. PSM has a key role to play here, and the EBU and its members are committed to ongoing AI literacy to support citizens. Most recently, the BBC published a [Guide to AI](#) to give young people the skills and confidence to navigate AI in everyday life. But audiences would be better served if AI developers played a bigger part, and they should do more to highlight the likelihood of errors in output.

EBU members are embracing AI to improve what we do and how we work. We already work closely with many AI developers and are excited about the opportunities. While recognizing the improvements made since the BBC's first research in this area, there remains an alarmingly high level of errors.

This year, the EBU set out its [News Integrity Principles in the Age of AI](#) and is talking to Big Tech companies on behalf of its members. We invite technology companies to enter a formal dialogue with news organizations to urgently and effectively develop standards of safety, accuracy and transparency.

We need AI developers to act, and act now.



Pete Archer
Programme Director, Generative AI, BBC



Jean Philip De Tender
EBU Deputy Director-General

Introduction

In February 2025, the BBC published research¹ examining how four widely used AI assistants – OpenAI’s ChatGPT, Microsoft’s Copilot, Google’s Gemini, and Perplexity – responded to questions about the news. The study found significant issues in over 50% of the responses, including that the assistants frequently distorted content from the BBC when it was used as a source.

Since then, research has confirmed the growing role that AI assistants play in digital news consumption – around 7% of people now use AI assistants as a source of news, rising to up to 15% of under 25s². This underlines the importance of assistants providing people with accurate information they can trust when responding to questions about the news.

Following the publication of the BBC report, the BBC partnered with the European Broadcasting Union (EBU) for a second research phase. This project included 22 Public Service Media organizations (PSM), representing 18 countries and 14 languages. Its purpose was to evaluate assistant response quality across different nations, languages and organizations, find out whether the issues identified in the first round of research are systemic, and gather a comprehensive set of examples illustrating the types of issues that occur when AI assistants answer news-related questions.

Alongside this report, we are releasing a “[News Integrity in AI Assistants Toolkit](#)”, designed to help develop solutions to the issues highlighted in the report. This includes improving AI assistant responses and media literacy among users. Building on the extensive insights and examples identified in the current research, the Toolkit addresses two main questions: “What makes a good AI assistant response to a news question?” and “What are the problems that need to be fixed?”.

1. BBC (2025), Representation of BBC News content in AI Assistants

2. Newman et al (2025), Reuters Institute Digital News Report: Based on a survey of 48 countries, 7% say they use AI assistants as a source of news weekly, rising to 15% of under 25s. Simon, Nielsen & Fletcher (2025), Generative AI and News Report 2025: How People Think About AI’s Role in Journalism and Society: Based on a survey of six countries, 6% get news via AI weekly, rising to 8% among 18-24s. Lipka & Eddy (2025), Relatively few Americans are getting news from AI chatbots like ChatGPT: About one-in-ten U.S. adults say they get news often (2%) or sometimes (7%) from AI chatbots like ChatGPT or Gemini, and 12% of under 50s get news at least sometimes from AI.

List of participating organizations

Participating organization	Country	Participating organization	Country
RTBF	Belgium	DW	Germany
VRT	Belgium	Rai	Italy
BBC	Britain	LRT	Lithuania
CBC	Canada	NOS/NPO	Netherlands
Radio-Canada	Canada	NRK	Norway
ČRo	Czechia	RTP	Portugal
Yle	Finland	RTVE	Spain
Radio France	France	SVT	Sweden
GPB	Georgia	SRF	Switzerland
ARD	Germany	Suspilne	Ukraine
ZDF	Germany	NPR	United States

Introduction

Approach

This research builds on the methodology used in the first round of BBC research. The BBC-EBU project team updated and adapted this to suit an international context with multiple participating organizations, such as by selecting news questions to ensure relevance across different markets.

For the current research, a set of 30 “core” news questions was developed – chosen from actual questions asked by audiences. These “core” questions were used by all participating organizations. Participating organizations also had the option to add their own “custom” questions on local and national issues relevant to them and their audience. The questions included a prefix asking the assistants to use their sources where possible (e.g. “Use DW sources where possible. [News question]”).

Four participating organizations generated responses in English, with the remaining 18 participating organizations translating the core questions into their own language. All participating organizations then generated responses to each question from each of the four AI assistants. This time, we used the free/consumer versions of ChatGPT, Copilot, Perplexity and Gemini. Free versions were chosen to replicate the default (and likely most common) experience for users. Responses were generated in late May and early June 2025.

Introduction

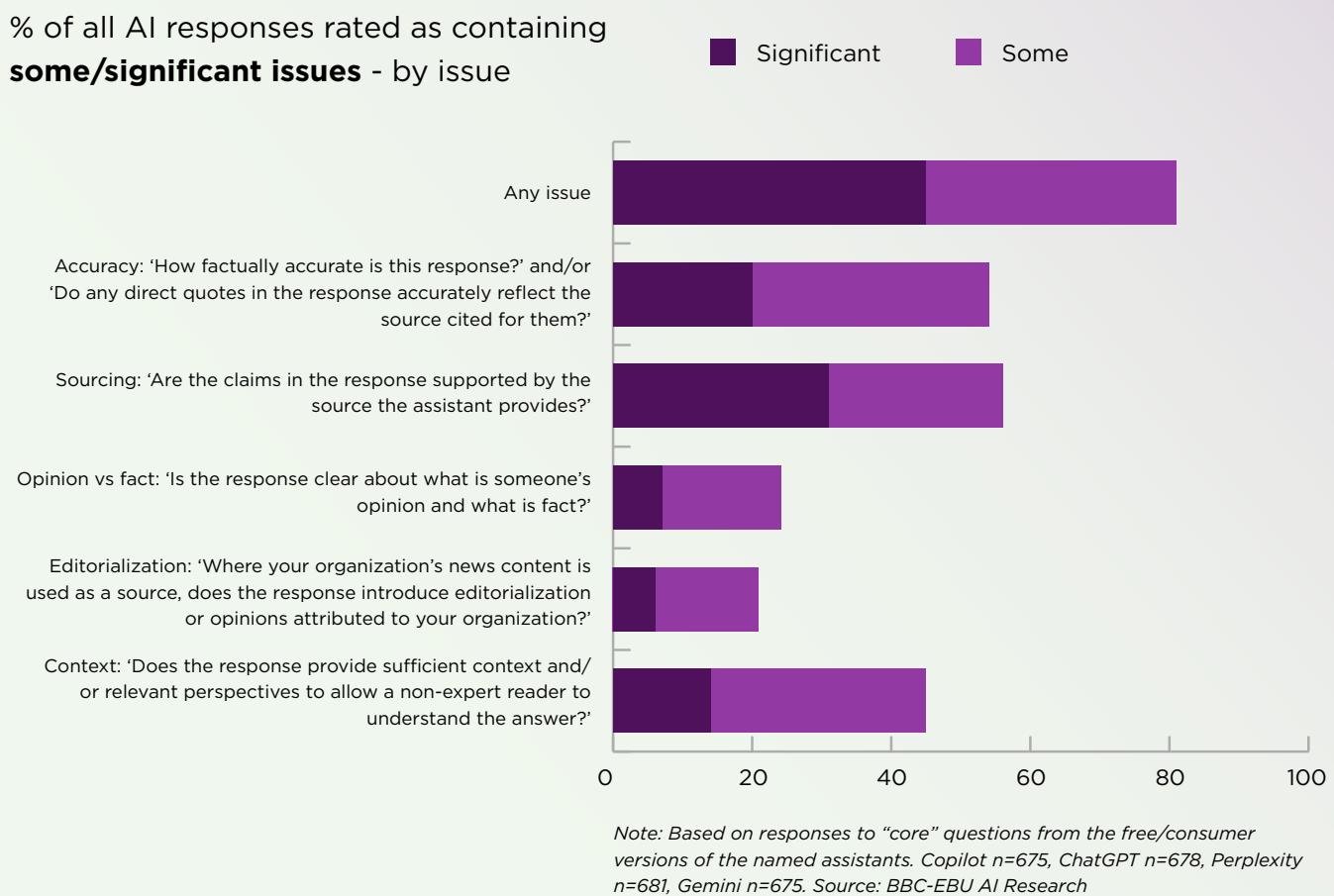
Responses were then evaluated by journalists from the participant organizations against five key criteria: accuracy (the evaluation asked about accuracy in general, and about the accuracy of direct quotes specifically), sourcing, distinguishing opinion from fact, editorialization (where the assistant adds words not in the cited source that imply a point of view or value judgement), and context (providing sufficient information and/or relevant perspectives to give a non-expert reader an answer which is complete and not misleading). For each of these criteria, journalists could rate each response as having no issues, some issues, significant issues or don't know.

From a potential 2,760 responses to core questions, assistant refusals and questions where responses were not successfully generated by the participant were excluded. Some responses were not evaluated (e.g. due to evaluator sickness or unavailability). This left 2,709 responses to core questions which were evaluated. Unless otherwise indicated, the statistics in this report are based on these core evaluations (so “responses” means “responses to core questions”) because of the comparability of the core set across all participating organization evaluations. Custom questions provide more specific material for the qualitative analysis and for Members’ own analysis – in this report, illustrative examples are drawn from the wider set of core and custom responses. Excerpts from assistant responses and evaluator comments have been translated into English where required by the relevant participating organizations.

A detailed methodology and list of participating organizations are included in the appendix.

High-level findings

Overall, 45% of all AI responses were found to have at least one 'significant' issue. When including 'some issues', 81% of responses have an issue of some form.



Sourcing was the biggest cause of problems, with 31% of all responses having significant issues with sourcing – this includes information in the response not supported by the cited source, providing no sources at all, or making incorrect or unverifiable sourcing claims.

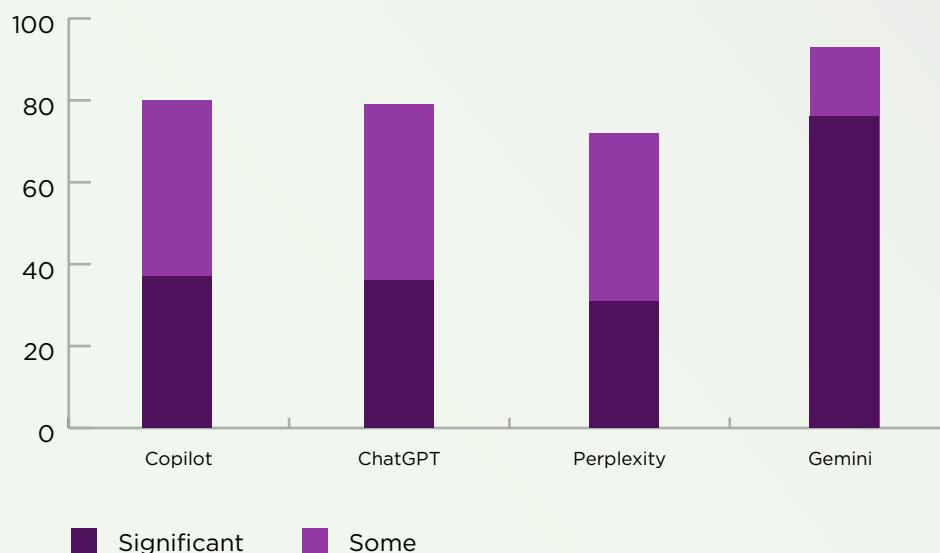
Accuracy (20%) and providing sufficient context (14%) were the next biggest contributors to significant issues.

Looking at the individual assistants, Gemini recorded the highest proportion of significant issues, impacting 76% of responses. This was double the rate of the next assistant Copilot (37%), followed by ChatGPT (36%) and Perplexity (30%).

Again, this was driven by issues with sourcing, where Gemini had significant issues in 72% of responses. This was three times as many as ChatGPT (24%), followed by Perplexity and Copilot (both 15%).

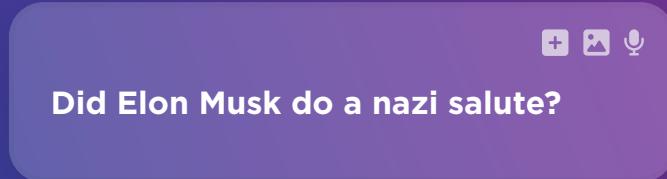
Across the entire dataset of 3,113 core and custom questions asked, only 17 were met with refusal (0.5% – down from 3% in the first BBC round). This suggests assistants are more willing to answer questions, whether or not they are capable of providing a high-quality answer.

% of all AI responses rated as containing
some/significant issues - by assistant



Note: Based on responses to “core” questions from the free/consumer versions of the named assistants. Copilot n=675, ChatGPT n=678, Perplexity n=681, Gemini n=675. Source: BBC-EBU AI Research

Assistant responses often contain multiple issues. A striking example is Gemini (Radio France) responding to “Did Elon Musk do a Nazi salute?”.



The screenshot shows the Gemini AI interface. At the top, the Gemini logo (a colorful star-like icon) and the word "Gemini" are displayed. Below the logo is a text input field containing the question "Did Elon Musk do a nazi salute?". To the right of the input field are three small icons: a plus sign, a camera, and a microphone. The background of the interface is a gradient of purple and blue.

According to sources from radiofrance.fr, specifically a segment titled “Charline Explodes the Facts” broadcast on YouTube by Radio France, Elon Musk allegedly gave a Nazi salute [1]. The segment mentions that he had “quite an erection in his right arm” and that it was “very explicit, it’s not a Nazi hello, no, no.”

Furthermore, a Wikipedia page listing controversies surrounding Elon Musk, which also references sources, mentions an “accusation of a Nazi salute at Donald Trump’s inauguration rally (2025).” This page also states that this gesture is part of Elon Musk’s “appeals” to white supremacist, anti-Semitic, and conspiracy theorist communities.

[1] <https://www.youtube.com/watch?v=e2bbb-6CIhs>

Illustration of a response provided by Gemini, generated by Radio France on June 4, 2025. Response translated from the original French. Original prompt included the prefix “Utilise des sources de radiofrance.fr / “Use Radio France sources”

Problematic sourcing

- Gemini mentions Radio France and Wikipedia as sources, but does not provide links to the content mentioned. The link [1] that is cited for the Radio France content is in fact a link to a video from British newspaper *The Telegraph*. None of the information in the response is in the *Telegraph* source.
- The Radio France segment “Charline Explodes the Facts” is a satirical radio segment, which is inappropriate as a source for a serious news question. The Radio France evaluator describes this as a “big problem”. This also illustrates

that while PSM can be a trusted source of information for news responses, assistants still need to distinguish between PSM content that is appropriate and inappropriate to use.

- A key claim (about Musk’s appeal to certain communities) is attributed to “a wikipedia page listing controversies surrounding Elon Musk”, but no link is provided. Gemini states vaguely that the wikipedia page “references sources”. The Wikipedia page itself attributes the same claim to “some media outlets” before providing a reference to an article from *Le Monde*. This is behind a

paywall, meaning a chain of sourcing claims hits a dead end with no clarity as to where this claim originally comes from.

Inappropriate tone

The response is tonally inappropriate because of the use of satire as a source, quoting a lewd double-entendre about an “erection” in response to a serious news question.

Not distinguishing opinion from fact

As above, satirical content is presented as fact. The claim that “this gesture is part of Musk’s appeals to white supremacist, anti-Semitic, and conspiracy theorist communities” is presented as fact, but while sourced to the Wikipedia page, this ultimately provides no clarity as to where this view has come from or who it belongs to.

Lack of context

The Radio France evaluator described the response as having “zero context, nothing”. The response reads as a selection of quotes or

claims from its sources, and comes across as one sided. It does not provide the necessary information and perspectives to allow the user to understand the issue.

Altered direct quote

Gemini adds words to a quote from the Radio France segment (“very explicit, it’s not a Nazi salute, no, no”). Gemini’s original response in French claimed it was “Très explicite, ce n’est pas un coucou nazi, non, non”. The actual quote from Radio France’s own transcript of the segment is “Très explicite, c’était pas un « coucou nazi »”

Reputational impact

The evaluator also notes the potential reputational impact when Gemini presents satirical content from Radio France as fact: “The chatbot therefore conveys false information using the name of Radio France, without mentioning that this information comes from a humorous source.”

Which questions most commonly caused significant issues?

The data enables us to analyse all assistants and participating organizations to identify which core questions most frequently lead to responses with significant issues. Sourcing-related issues are excluded to focus on issues related to the quality and content of the responses.

Responses to core question MOST often flagged as having significant issues of any kind, excluding sourcing		Responses to core question LEAST often flagged as having significant issues of any kind, excluding sourcing	
Is Trump starting a trade war?	48%	How many countries will the 2026 FIFA World Cup be played in?	9%
How many people died in the Myanmar earthquake?	47%	Where is Elon Musk from?	14%
Why did the US bomb Yemen?	46%	Why were NASA astronauts stuck in space?	16%
How did Trump calculate the tariffs?	46%	Why does Zelensky not wear suits?	16%
Is Viktor Orbán a dictator?	40%	Why does Trump want to buy Greenland?	16%
Is climate change getting better?	38%	Is Türkiye in the EU?	18%

The data suggests that assistants particularly struggle with fast-moving stories with rapidly changing or updating information (Trump trade war, Myanmar); intricate timelines involving multiple actors (Yemen); detailed information (China exports, Trump trade war, Trump tariffs); or topics that require clear distinction between facts and opinions and proper attribution of claims (Orbán, climate change).

AI assistant responses to questions that are most fact-based and straightforward, such as the number of countries hosting the 2026 FIFA World Cup or Elon Musk's birthplace, are less likely to contain significant issues. The same is true for stories that do not change rapidly (Zelensky not wearing suits, Trump and Greenland, Türkiye in the EU). The stories that are less reliant on interpretation (NASA) also generate fewer issues.

Detailed findings

In the sections below, we explore the different criteria against which the responses were assessed and the issues identified by evaluators. These are illustrated with examples from participating organizations. Each example is labelled by assistant and participating organization, sometimes with the news question included, e.g. *ChatGPT (ZDF)* or *ChatGPT (Yle): Is Viktor Orbán a dictator?* Where relevant, examples were translated into English by the participating organization. Quotes from participating organization analysis are also included, labelled by organization, e.g. *SVT analysis*.

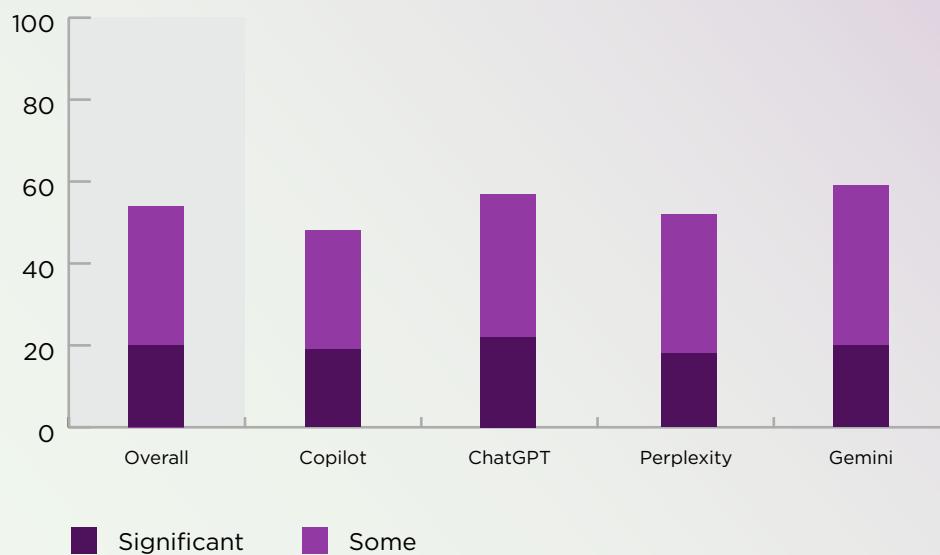
Accuracy

Accurate information is at the core of high-quality news. It allows audiences to understand what is really going on in the world, counters misinformation and is the bedrock of informed public debate in democratic societies. Accuracy is also central to trust.

Journalists were asked to evaluate the accuracy of responses, with a separate question asking specifically about the accuracy of direct quotes. The aim was to understand whether the assistants were getting key factual details and information right. This includes things like names, numbers, and dates, and accurately characterizing events and relations. For quotes, evaluators assessed whether direct quotes in assistant responses, full or partial, matched the words used in the cited source and correctly identified the person who said the words.

Overall, 20% of responses had significant issues with accuracy (general accuracy and/or accuracy of direct quotes). The assistants all performed similarly in this regard, with all between 18 and 22% significant issues.

% of AI responses rated as containing some/significant issues with **accuracy** (including accuracy of direct quotes)



Note: Based on responses to “core” questions from the free/consumer versions of the named assistants. Copilot n=675, ChatGPT n=678, Perplexity n=681, Gemini n=675. Source: BBC-EBU AI Research.

Across all participating organizations, all assistants made basic factual errors.

Some were potentially consequential errors on matters of law. Perplexity (ČRo) claimed that surrogacy “is prohibited by law” in Czechia, when in fact it is not regulated by the law and is neither explicitly prohibited nor permitted. Gemini (BBC) incorrectly characterized a change to the law around disposable vapes, saying it would be illegal to buy them, when in fact it was the sale and supply of vapes which was to be made illegal.

One of the most common issues was outdated information. In response to the question “Who is the Pope?”, ChatGPT (Yle, Swedish) claimed in May 2025 that “Pope Francis is the current leader of the Roman Catholic Church” whereas Pope Francis had died in April, and Pope Leo XIV had succeeded him as Pope. Gemini (NOS/NPO) and Copilot (NOS/NPO) made the same error – the latter claimed Francis was Pope despite also saying “He died on April 21, 2025”.

Who is the Pope?



ChatGPT

Pope Francis is the current leader of the Roman Catholic Church [1]. He was elected pope in March 2013 and is the first pope from Latin America as well as the first Jesuit to hold the position [2].

[1] <https://yle.fi/t/18-124596/sv>

[2]: <https://en.wikipedia.org/wiki/Yle>

A



Copilot

The current pope is Francis, born Jorge Mario Bergoglio [1]

He died on April 21, 2025[1]

[1] https://nl.wikipedia.org/wiki/Paus_Franciscus

A



Gemini

The current pope is Pope Francis.

For specific NOS sources, it's best to consult the websites from NOS itself and search for articles about "Pope Francis." They have extensive coverage of his pontificate and relevant events.

A

A Out-of-date information

These are all incorrect - at the time the questions were asked, Leo XIV was the Pope. Copilot says that Francis is Pope despite the same response also saying when he died.

B

B Problems with sourcing

ChatGPT provides Yle's Wikipedia page as a source - the linked claim about Pope Francis is not in this source, which is irrelevant to the question.

Illustration of responses provided by the listed AI assistants on 28 May (ChatGPT, generated by Yle, response translated from original Swedish), 30 May (Copilot, generated by NOS, response translated from original Dutch), and 29 May (Gemini, generated by NOS, response translated from original Dutch), 2025. Yle original prompt included the prefix "Använd källor från Yle om möjligt / Use Yle sources where possible". NOS original prompt included the prefix "Gebruik NOS-bronnen waar mogelijk / Use NOS sources where possible".

ChatGPT (DW) referred to Olaf Scholz as Chancellor of Germany and Jens Stoltenberg as NATO Secretary General when both no longer held these positions at the time the answer was generated. Similarly, Gemini (VRT) referred to Paul Van Tigchelt as the Belgian Minister of Justice when he no longer held that role. In a reply to the question “Who won the legislative elections?” (in Portugal), Copilot (RTP) ignored the most recent elections that occurred in May 2025, one month before the question was asked, and instead gave information on the 2024 legislative elections.

Sometimes the factual error was caused by AI assistants inaccurately representing information from up-to-date sources. At other times, sources themselves were out of date.

Copilot (Rai) provided a video from 2022 and an article from 2014 as sources in

response to “What does NATO do?”. Rai pointed out that the 2022 video “refers to the first NATO meeting after the Russian invasion. Obviously, the answer would have been ok the day after the meeting, but not three years later” and is “missing context about NATO and Ukraine”. Copilot (BBC), in response to the question “Should I be worried about the bird flu”, claimed “A vaccine trial is underway in Oxford”. The source for this was a BBC article from 2006, almost 20 years old.

Assistants also struggled with rapidly developing stories (even if the story was no longer developing at the time the question was asked). As noted above, the question which caused the most responses overall with significant issues was “How many people died in the Myanmar earthquake?” – all assistants across multiple participating organizations and languages struggled to provide accurate answers to the question.

“AI assistants face significant challenges when handling developing news stories. They often combine information from articles published at different times, creating potentially misleading narratives for readers. This chronological confusion is especially problematic when reporting on evolving situations.”

Yle Analysis

As well as presenting out-of-date facts, all of the assistants also struggled with accurately portraying the relationship between events in time, or in relation to when the question was asked. Gemini (Radio France) and ChatGPT (RTP) presented Donald

Trump's re-election for a second term as a possibility rather than a reality when responding in May 2025, as if the question had been asked before the 2024 election. In the case of Radio France, this was despite the cited White House source containing the correct information.

Causal relations also caused problems. In response to "Did Elon Musk do a Nazi salute?" Copilot (ZDF) conflated Musk's resignation as special advisor to President Trump with the discussion about whether or not he did a Nazi salute, in a way that misleadingly implies the latter was the cause for the former. The false conflation was also attributed to ZDF, with Copilot adding to the potential reputational damage to ZDF by saying "other sources emphasize that context is crucial". The ZDF source cited by Copilot was a photo gallery which did not specifically address the question Copilot was trying to answer – ignoring other available ZDF sources which do provide relevant context.

Failure to accurately represent information in cited sources was a broader problem. In response to the question "How did Trump calculate the tariffs?", Gemini (LRT) claimed tariffs imposed on Canada and Mexico by Donald Trump were "on certain goods, including fentanyl-related items." In fact, as the cited White House source makes clear, the tariffs were not levied on goods containing fentanyl but were meant as punishment for failing to take action on fentanyl trafficking.

While responding to the question "What did Marine Le Pen do?", ChatGPT (RTVE) mentioned Le Pen's successor as leader of the National Rally party, Jordan Bardella, as a potential candidate in the 2027 French Presidential election. Talking of Bardella, ChatGPT said he lacked "experience", but according to RTVE, "what the [source] suggests is that he lacks 'preparation' ... a very relevant nuance."

Inaccurate treatment of participating organization sources could potentially be consequential for the public broadcasters. In a response to a custom question about Peter Sullivan (a British man who had a murder conviction quashed after more than 38 years in prison), ChatGPT (BBC) claimed that "Following his exoneration, Merseyside Police have reopened the investigation to identify the true perpetrator." However, as the BBC source cited for this makes clear, the new investigation began "as the appeal was underway". The BBC journalist evaluating this commented, "This is materially inaccurate and materially unfair to Merseyside Police and would lead to an editorial complaint."

Have assistants improved? Comparing the two rounds of BBC research

With two rounds of research focused on the same assistants six months apart (see table below for Product/LLM versions), can we compare results to see if the assistants are improving?

The overall results of the current study bring together multiple languages, countries and organizations, making them not directly comparable to the first round of BBC-only research. But while there are some small differences in methodology and definition of key statistics, comparing BBC-only data from both rounds can at least give a sense of the overall direction of travel for assistant performance. This comparison is based on 362 responses evaluated in the first round and 237 core + custom responses evaluated in the current round.

Overall, there are signs that the quality of assistant responses has improved – the share of responses with significant issues of any kind fell from 51% in the first round to 37% in the current round. Significant issues, however, remain at a concerning level – affecting half of responses for Gemini, and around one-third of responses for ChatGPT, Perplexity and Copilot.

Looking in more detail, for key considerations such as accuracy, sourcing and context, while there were improvements overall, performance varied for the individual assistants.

Accuracy showed a small improvement overall from 31% last time to 25% this time. Gemini was the biggest improver (46% last time to 25% this time), Copilot also improved, while ChatGPT and Perplexity showed no improvement – meaning all assistants are now in the range 20-29% for significant issues with accuracy.

Sourcing remains the sharpest divide between assistants, and Gemini still has the highest percentage of significant issues, broadly the same at 47% (see “Gemini’s issues with sourcing”). By contrast, the other assistants all improved to the 10-15% range, with Copilot showing the steepest drop from 27% to 10%.

And there were fewer problems overall with providing sufficient context: for the BBC data the percentage of significant issues dropped from 22% to 13%. Gemini improved most (from 36%

significant issues to 5%), while Copilot and Perplexity remained broadly the same at 24% and 14% respectively.

Other areas showed more consistent improvement across all assistants. The proportion of responses where assistants had problems distinguishing fact from opinion fell from 12% to 3%, while problems with editorialization also fell. For both, Copilot and Gemini improved the most, bringing the performance of all assistants onto roughly equal footing. Anecdotally, responses seemed to contain fewer opinionated conclusions or questions to the user.

For the BBC, lack of any direct sources at all (i.e. URLs cited in support of claims and linking to specific articles or content) fell substantially. Only a single BBC response this round lacked a direct URL source, compared with 25 in the previous round (mostly in responses from Gemini), though this still remains a non-trivial issue for other organizations and languages.

Finally, in the first round there were 13 refusals – 12 of them from Gemini – but in the current round, there was only one refusal, from Copilot.

AI assistant versions for BBC evaluations

AI assistant	LLM version for previous BBC research (response generation Dec 2024)		Current BBC / EBU research (response generation May/June 2025)	
	Product version	LLM	Product version	LLM
ChatGPT	Enterprise	GPT-4o	Consumer / Free	Default – GPT4o
Copilot	Pro	Not specified	Consumer / Free	Default (no options available)
Gemini	Standard	Not specified	Consumer / Free	Default – 2.5 Flash
Perplexity	Pro	Default	Consumer / Free	Default (no options available)

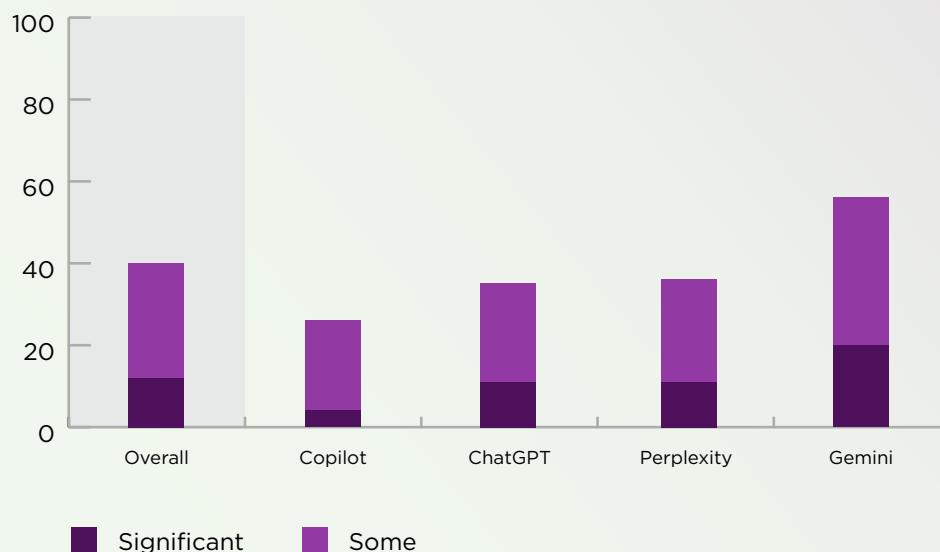
Accuracy of direct quotes

Across all AI assistant responses that included a direct quote (a total of 1,053), 12% were found to have significant issues with the accuracy of those direct quotes.

Gemini responses with direct quotes had significant issues in 20% of 290 responses.

Copilot used the fewest direct quotes but also had the fewest significant issues – 4% of responses with direct quotes.

% of AI responses with direct quotes rated as containing some/
significant issues with **accuracy of direct quotes**



*Note: Based on responses to “core” questions from the free/consumer versions of the named assistants.
Copilot n=190, ChatGPT n=262, Perplexity n=311, Gemini n=290. Source: BBC-EBU AI Research*

Sometimes assistants appeared to fabricate quotes completely. Gemini (ZDF), responding to “Is Viktor Orbán a dictator?”, claims Orbán “is described as “Putin’s bridgehead in the EU””. The latter quote (“Brückenkopf Putins in der EU” in the original response in German) is not found in the source provided for it, and appears to be fabricated.

At other times quotes are altered in assistant responses. Answering “Is Türkiye in the EU?”, Perplexity (ZDF) includes an unattributed quote: “Türkiye’s geopolitical and strategic importance cannot make up for the government’s democratic backsliding, and EU

membership criteria remain unmet". The actual wording, which is both correctly reported and properly attributed to the European Parliament in the ZDF source, is: "and EU membership criteria are not up for negotiation."

In response to the question "Is Trump starting a trade war?", ChatGPT (Radio-Canada) responded, "The then Canadian Prime Minister, Justin Trudeau, described these measures as a 'stupid trade war'". Radio-Canada's evaluator explained, "There is no source to support the quote 'stupid trade war' attributed to then-Prime Minister Justin Trudeau. Furthermore, the quote is not accurate. The former Canadian PM said, 'It's a very stupid thing to do,' but several media outlets that quoted him said he had denounced a 'stupid' trade war, which is probably where the assistant's mistake came from."

When asked, "Why did Birmingham bin men go on strike?", Perplexity (BBC custom) fabricated or altered three separate quotes. Quotes attributed to the Unite union and Birmingham City Council are not in the sources cited for them and appear to be made up.

The latter appears under a heading "Key Quotes" – where it is joined by a third quote, which Perplexity has altered. Perplexity claimed the general secretary of the Unite trade union said, "Our members cannot withstand drastic pay reductions of up to £8,000 without any compensation" whereas the cited BBC source reported her as saying "The bottom line is that our members can't afford to have savage pay cuts of up to £8,000 with no mitigation."

Even when the words are correct, quotes can be attributed to the wrong person. ChatGPT (LRT), answering "What is the Ukraine minerals deal?", wrongly attributed a quote by then-economic affairs minister Yulia Svyrydenko, in which she described the Ukraine minerals deal as "balanced and fair", to Volodymyr Zelensky, even though the cited source carries the correct attribution.

Finally, in response to "Why does Zelensky not wear suits?", NPR noted that Perplexity "fails to use quote marks to distinguish between direct quotes and paraphrases, which is misleading and can be considered plagiarism."



Perplexity



Why did Birmingham bin men go on strike?

Under the new proposals, drivers' salaries could fall from around £40,000 to £32,000. The union argued that this was an "assault on wages and working conditions," and that the removal of key roles, such as Waste Recycling and Collection Officer (WRCO), would also impact safety standards within the service [6][8].

Key Quotes

"Our members cannot withstand drastic pay reductions of up to £8,000 without any compensation". Sharon Graham, Unite General Secretary [4]

"We have taken measures to support Birmingham City Council employees affected by the changes in their work arrangements".
Birmingham City Council spokesperson [6]

[4] <https://www.bbc.co.uk/news/articles/czr88yxrykpo>
[6] <https://www.bbc.co.uk/news/articles/cdjjg07xvnj0>
[8] <https://www.bbc.co.uk/news/articles/c1k4m7reznj0>

A Fabricated Quotes

These quotes are not in the cited sources, and cannot be found elsewhere. They appear to be made up.

B Altered Quote

This has been altered from the original in the BBC source. The original quote is: "The bottom line is that our members can't afford to have savage pay cuts of up to £8,000 with no mitigation."

Illustration of a response provided by Perplexity, generated by the BBC on 29 May, 2025. Original prompt included the prefix "Use BBC News sources where possible".

The problem of over-confidence

Many of the issues identified in this research are exacerbated by the confidence with which assistants deliver their responses.

Participating organizations flagged a range of ways this shows up – from whether or not the assistant answers a question at all, to the confident tone adopted by assistants and lack of ability to contextualize doubt or uncertainty, to the way responses are often presented (structure, abundance of sources) which creates a misleading air of authority.

“Even if the assistant makes a mistake, it continues to generate and develop a response, sometimes based on incorrect information.”

Suspilne analysis

As noted above, refusal rates were low in the current research – only 17 questions across the whole core + custom dataset (0.5% of 3,113 questions) were met with refusal, down from 3% in first BBC round.

This decline in refusal rate echoes recent research from NewsGuard¹, who observed, “As chatbots adopted real-time web searches, they moved away from declining to answer questions. Their non-response rates fell from 31 percent in August 2024 to 0 percent in August 2025.” NewsGuard saw this as a cause of an increase in the prevalence of “false information on topics in the news” – inaccuracy goes up when assistants try to answer all questions.

This is not just about whether the assistants provide an answer at all, but also how they answer, particularly their inability to express uncertainty.

“Like all the summaries, the AI fails to answer the question with a simple and accurate “we don’t know”. It tries to fill the gap with explanation rather than doing what a good journalist would do which is explain the limits of what we know to be true.”

BBC Evaluator comment, Gemini, Custom question

1. NewsGuard (2025), AI False Information Rate Nearly Doubles in One Year

Similarly, participating organizations in the current research noted that both tone and presentation of responses convey a confidence that the information they contain is accurate.

ChatGPT (RTVE), answering “What did Marine Le Pen do?” states in its own voice that “Le Pen’s situation represents a turning point in French politics”, phrasing which conveys authoritative expertise on French politics. RTVE had another similarly over-confident example from ChatGPT, where it replied to “Is Trump starting a trade war?”, with “Donald Trump is not only starting a trade war; he has already escalated it since his return to the presidency in 2025”.

“AI assistants mimic journalistic authority – without journalistic rigor. ChatGPT and Gemini in particular generated responses that read like polished news articles, confident tone, summary structure, and even the right phrasing cadence. However, this masks underlying issues such as lack of source traceability, subtle bias in framing, fabricated or assumed consensus. This creates a dangerous illusion of reliability. Users may not question these outputs – especially if they lack strong media literacy.”

GPB analysis

Recent research from Open AI¹, suggests assistants may not reflect uncertainty because “training and evaluation procedures reward guessing over acknowledging uncertainty ... language models are optimized to be good test-takers, and guessing when uncertain improves test performance.” The researchers suggest this could be addressed by modifying evaluation approaches to avoid penalizing uncertainty.

Answering one question, Gemini showed an unusual form of confident inaccuracy. In response to “Why were NASA astronauts stuck in space?”, Gemini (Yle - Finnish) claimed “NASA astronauts have not been stuck in space”, when Butch Wilmore and Suni Williams had spent just over nine months (instead of a planned eight days) on the International Space Station after their spacecraft malfunctioned. Gemini then claimed, “You might be confusing this with a sci-fi movie or news that discussed a potential scenario where astronauts could get into trouble.”

1. Kalai et al (2025), Why Language Models Hallucinate

Gemini gave a very similar response to RTP.



Gemini



Why were NASA astronauts stuck in space?

This is a misconception. NASA astronauts are not currently stranded in space. All NASA crewed missions return or are planned to return safely to Earth.

Possible Reasons for Your Confusion

This idea may have arisen from:

Science fiction films: Many films depict scenarios where astronauts are stranded in space, which can lead to misunderstandings.

Delays or technical issues on missions: While rare and usually resolved, space missions can experience minor delays or technical issues that, while not stranding the astronauts, can be misinterpreted by the public.

Misinformation: Sometimes, false or misinterpreted news circulates, creating confusion.

A

B

B

C

D

A Doesn't answer question

The assistant does not answer the user's question ("Why were ...") and instead focuses on whether astronauts are currently stranded in space. "This is a misconception" is factually inaccurate - two astronauts had spent nine months in space after a malfunction.

B Undermining the user

The words used by Gemini put the blame on the user, undermining them.

C Undermining trust

It is particularly damaging that the assistant suggests that a valid question may have been based on misinformation.

D No Sources

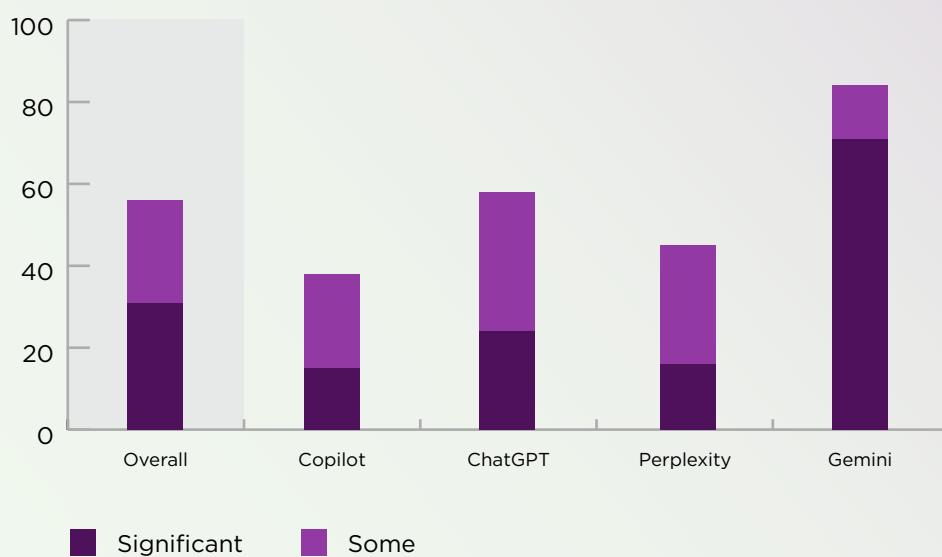
No sources are provided.

Illustration of a response provided by Gemini, generated by RTP on 10 June, 2025. Response translated from the original Portuguese. Original prompt included the prefix "Usa a RTP como fonte sempre que possível / Use RTP as a source where possible".

Sourcing

High-quality sourcing promotes transparency and accountability, allowing users to assess and verify the information that assistants use to create their responses. Recent research suggests only 24% of people in the US who get news from AI chatbots find it easy to determine what is true and what is not¹.

% of AI responses rated as containing some/significant issues with **sourcing**



Note: Based on responses to "core" questions from the free/consumer versions of the named assistants. Copilot n=675, ChatGPT n=678, Perplexity n=681, Gemini n=675

Source: BBC-EBU AI Research

These values are also central to high-quality media organizations, including PSM. Accurate sourcing is also a key issue for media organizations themselves because of the potential negative reputational impact when information or claims are falsely attributed to our organizations.

As noted above, problems with sourcing were the biggest cause of significant issues overall. Gemini's 72% of responses with significant issues with sourcing was a clear outlier compared to ChatGPT (24%), Copilot (15%), and Perplexity (15%).

The research identified a wide range of problems that can arise with sourcing.

¹ Lipka & Eddy (2025), Relatively few Americans are getting news from AI chatbots like ChatGPT

Some of these, such as providing no direct sources (i.e. a link to a URL for specific content) at all or making incorrect or unverifiable sourcing claims in a response, predominantly impact Gemini – see “Gemini’s issues with sourcing” below for more.

For all assistants, there was a range of other issues. The most basic was sources being cited in support of a claim but not containing the claim in question.

ChatGPT (SRF) gave a detailed response to “What does China export?”. But SRF’s analysis pointed out, “it’s totally unclear where the information comes from. Neither source contains many of the numbers ChatGPT mentions. So, for the journalist, it was nearly impossible to verify if the facts in the response are true or part of ChatGPT’s hallucinations.”

In response to “Is Viktor Orbán a dictator?” Perplexity (ARD) said ARD’s Tagesschau “describes Orbán’s rule as authoritarian and illiberal”. But this is not found in the cited Tagesschau source.

“Many responses contain what could be called ‘ceremonial citations’ – references added to create an impression of thorough research, but which don’t actually support the stated claims when checked. This practice undermines the reliability of the information provided.”

Yle Analysis

Some assistants made up sources completely. Gemini (NRK), Copilot (RTBF, NRK), and ChatGPT (RTP, NRK) all produced links that appeared credible but in fact did not exist. Such fabricated links sometimes pointed to the participating organizations’ own news sites. NRK observed that “in answers on ‘polarizing’ topics ... this might give the impression that we have removed content without explaining why to our readers.”

Even when assistants did provide real sources, they were not always practically accessible for verification, for example, when they were behind paywalls. While paywalled content could be a high quality, accurate source that is appropriate for a response, this does impact the user’s ability to verify information.

Sources were also flagged as problematic when they were not relevant or appropriate.

“Some AI [assistants] cannot select the most relevant sources to explain a topic, but often select secondary articles or videos from less reliable news outlets. This happens less with ChatGPT.”

Rai Analysis

In response to the question “Is Trump starting a trade war?” (SRF) Gemini’s response prompted the evaluator to observe, “Some of the links are not entirely appropriate to the content of the answer and only relate to it marginally”.

While participating organizations often felt Perplexity was strong on sourcing, there were also a number of observations that some sources lacked relevance, or were not used at all.

VRT noted that Perplexity lists nine VRT sources in its response to the question “Why change to the Gulf of America?”, including some entirely unrelated, such as articles on the abolition of first-class train seats, power plants in the Netherlands and a 2012 article on a mumps outbreak.

Perplexity (Suspilne) replied to the question “Why does Zelensky not wear suits?” including a link to an article dedicated to the costumes of the Ukrainian band Kalush at Eurovision. As well as being irrelevant, the source was not specifically cited in the response.

In response to the question “How many people died in Myanmar earthquake?” Perplexity (NRK) appended a sources block with 19 URLs but only referenced three of the sources in the body of the answer. Similarly, it provided NRK with 9 links in its response to the question “What does NATO do?” but only referred to three of them. The NRK evaluator described this as “Perplexity providing long lists of URLs without actually referring to them in the answers.”

As well as potentially overwhelming users and undermining the usefulness of citations, participating organizations suggested that too many sources can also create a false perception of quality – see “The problem of over-confidence”.

Participant organizations also flagged that some sources were inappropriate. ČRo reported that responses from ChatGPT and Copilot to “Why does Zelensky not wear suits” both used prvnizpravy.cz as a source. ČRo noted that this site “was designated by the Czech army secret service as spreading Russian disinformation, and therefore blocked by some internet providers at the outset of the Russian invasion of Ukraine”.

Sometimes, this involved specific pieces of content, such as satire and political or corporate sources, being used for or presented as factual claims.

Gemini (NPR) provided an eBay listing as a source in a response about Chinese exports. Perplexity (NOS/NPO, Radio France) used satirical content as sources for claims without making this clear. Answering a question from Radio France about why people hate Tesla, Perplexity based part of its response on a satirical column, claiming “The rejection of Tesla goes beyond the simple issue of cars: it has become a political and social marker. Arriving in a Tesla at a ‘left-wing’ family dinner can now provoke hostile reactions, even shame on the part of the owner.” In an otherwise factual answer, there was nothing to indicate this was a satirical point of view.

Although political or partisan sources may be appropriate in some circumstances (such as using the Kremlin website to report the official view of the Russian government), evaluators also flagged cases where this impacted the quality of the response.

Gemini (RTBF) responded to the question “How long has Putin been president?” with a response citing three sources: two from the Kremlin and one from the Russian embassy in Ireland. The evaluator observed “these sites fail to mention that the duration and recurrence of Vladimir Putin’s presidencies are the direct consequences of two changes to the Russian Constitution, which were favourable to him.” This meant the response lacked crucial context.

Corporate websites could also be inappropriate sources. Responding to the question “Where will the 2026 World Cup be played?”, Copilot (Radio-Canada) used only FIFA sources and replied that it would be a “historic event for global football.” Radio-Canada said, “The use of promotional sources (FIFA press releases) results in an editorial tone that is not consistent with a neutral journalistic response.”

Responding to “What is Nvidia famous for?”, half the sources provided by Gemini (DW) were from Nvidia itself. Gemini stated that Nvidia’s GeForce graphics cards are the “leading choice for gamers due to their exceptional performance and immersive experiences,” which was not supported by the source cited for it. DW felt that the response “reads a lot like a press release or opinion instead of a neutral statement sourced from an article. The response appears to have lifted a lot of Nvidia’s marketing speak from its own website.”

At other times, participating organizations expressed concern with the

“A concerning trend is how these assistants treat all sources as equally reliable. Content from social media platforms, think tanks, or personal blogs is frequently presented alongside established news organizations without any indication of their differing credibility levels.”

Yle analysis

appropriateness of certain sources at a deeper level. This was often the case when traditional news sources were not used for news questions in favour of sources with concerns about accuracy or lack of traditional journalistic verification processes.

Reddit was flagged by a number of participating organizations as an inappropriate source because of the mix of user-generated comments with summaries and links to news articles, suggesting that those sources should be linked to directly rather than via Reddit. GPB's evaluator notes that Chat GPT, answering “Why does East Germany vote AfD?”, “bases the response on Reddit rather than journalistic or academic sources”. ChatGPT cites a discussion thread on the r/AskAGerman board titled “Why is East Germany so obsessed with the AfD?” to support its claim that “when the socialist system collapsed, some people were more easily attracted by the AfD's nationalist and anti-Western rhetoric”.

ARD questioned why assistants cited ARD's own user forums (containing audience comments and debate) rather than their news articles.

Participating organizations raised concerns about responses that relied heavily or solely on Wikipedia content – Radio-Canada calculated that of 108 sources cited in responses from ChatGPT, 58% were from Wikipedia. CBC-Radio-Canada are amongst a number of Canadian media organisations suing ChatGPT's creator, OpenAI, for copyright infringement. Although the impact of this on ChatGPT's approach to sourcing is not explicitly known, it may explain the high use of Wikipedia sources.

Suspilne reported that “the Ukrainian-language Wikipedia often contains errors”:

“It is also worth noting that assistants rely too often on Wikipedia and do not use primary sources. During the analysis, the assistants used Wikipedia as a source of information more than 40 times. The assistants also used the tag page or front page as a source, which makes it almost impossible for the user to verify the information.”

Suspilne analysis

CBC also expressed concern that Wikipedia often draws on PSM and other media as sources, sometimes without clear attribution, and that it would be preferable for assistants to source responses to news questions directly from PSM and other media.

Problems also arose where assistants failed to provide sources for key claims. Perplexity (CBC) responded to “How long has Putin been president?” with biographical information, including the naming of five of his children. CBC’s evaluator noted that “Putin’s family – like how many children – [is] never public information except rumours and speculation. It is unclear where the information comes from, as no sources [were] quoted [for this claim], but it is presented in a context that people may think those come from CBC sources because of the sources quoted [later in the response].”

Finally, our questions directed assistants to use participating organization content “where possible”. A number of participating organizations reported assistants claiming that their organization did not have any content on the story, even when this was false.

RTP reported that “in response to “Why do people hate Tesla?”, Perplexity claims there’s no information on RTP, but there are articles.” ChatGPT claimed ZDF had no information to respond to questions about the World Cup, and Perplexity claimed the same with a question about Vladimir Putin – this was not true for either, and ZDF noted “it is damaging to our reputation if it is suggested that we have no coverage of these issues.”

NRK “saw examples of Perplexity and ChatGPT claiming NRK had not covered a topic, or that it could not access our content, while saying “so I provided search results”, and those results included NRK’s URLs, which could be confusing for the reader.”

And even when it was true that a participant had no coverage of a story, assistants can draw the wrong conclusions. Perplexity (GPB) responding to “Did Elon Musk do a Nazi salute?” said “There is no information about Elon Musk using the Nazi salute in 1tv.ge’s sources... Therefore, according to 1tv.ge, Elon Musk did not use the Nazi salute”. As GPB noted in their analysis, “This is a logical error: if 1tv.ge provides no information on the subject, then it cannot be cited as confirming or denying the event. The assistant wrongly infers absence of coverage as evidence of denial, which misrepresents the source and creates a misleading impression of factual certainty.”



Perplexity



Why change to the Gulf of America?

The change to “Gulf of America” is a domestic, political decision by US President Trump, intended to emphasize the American character of the region. Internationally, nothing changes: the rest of the world will continue to refer to the region as the “Gulf of Mexico” unless countries themselves decide to adopt the American designation. Major tech companies are following local regulations in their mapping apps, so American users now see “Gulf of America” [6][8][9].

- [1] <https://www.vrt.be/vrtnws/nl/>
- [2] <https://www.vrt.be/nl/ons-aanbod/vrt-nws>
- [3] https://www.vrt.be/vrtnws/nl/2012/06/04/bron_van_gentse_bofuitbraakligtinnederland-1-1320086/
- [4] <https://www.vrt.be/vrtnws/nl/2025/05/16/kerncentrales-nederland-nieuw-prijs/>
- [5] <https://www.vrt.be/vrtnws/nl/2024/09/06/schiermonnikoog-nederland-waddeneiland-eiland-scheur/>
- [6] <https://www.vrt.be/nwsnwsnws/nl/post/2025/02/12/golfvanmexico/>
- [7] <https://www.vrt.be/vrtnws/nl/2025/05/26/moet-nmbs-eerste-klas-afschaffen/>
- [8] <https://www.vrt.be/vrtnws/nl/2025/01/21/kan-trump-de-golf-van-mexico-omdopen-tot-golf-van-amerika/>
- [9] <https://www.vrt.be/vrtnws/nl/2025/01/28/golf-van-mexico-golf-van-amerika-donald-trump-google-maps/>

A Redundant and irrelevant sources

Perplexity lists nine VRT sources, but only refers to three of them in the response. The rest of the sources are irrelevant to the question - and include articles on the abolition of first-class train seats, power plants in the Netherlands, and a 2012 article on a mumps outbreak.

A

Illustration of a response provided by Perplexity, generated by VRT on 28 May, 2025. Response translated from the original Dutch. Original prompt included the prefix “Gebruik VRT NWS bronnen wanneer mogelijk / Use VRT NWS sources where possible.

Gemini's issues with sourcing

Gemini was a notable outlier in terms of sourcing performance, with significant sourcing issues impacting almost three-quarters of responses (triple the rate of the other assistants). This was driven by a number of key issues.

“Gemini was especially striking in this regard, as it varied greatly in how sources were presented: sometimes without links, sometimes with inline references and only rarely with direct links. These changing output formats appeared highly inconsistent and therefore stood out the most.”

ARD analysis

42% of Gemini responses provided no direct sources (URLs pointing to specific pieces of content), meaning users cannot verify the claims in the response.

Gemini also showed a strong tendency to make incorrect or unverifiable sourcing claims. This most commonly occurred through the use of phrases such as “According to available information, particularly from Radio France …”, “sources such as RTBF say …”, “according to RTVE and other sources …”, accompanied by either a citation to a URL for a different media organization or source, or no direct source at all. Evaluators noted this type of sourcing behaviour in 54% of Gemini responses. This specific issue disproportionately affected Gemini – no other assistant had more than 4% responses impacted.

“Gemini showed problematic sourcing behaviour, frequently citing “Yle” without providing actual links or references, appearing overly eager to please at the expense of accuracy.”

Yle analysis

For example, when asked by CBC about the origin of the Los Angeles fires, the response states “CBC News reports highlight that climate change significantly contributed to the conditions…”, and “Here’s a breakdown of the key factors, according to CBC News”, as well as “CBC News emphasizes that human-caused climate change created the critical underlying conditions…”, but the five sources provided for the response do not include CBC News and



CBC evaluators were unable to find any origin for these specific statements, outside of expert interviews.

Responding to a question from NPR, Gemini framed the bulk of its response with “Here’s a breakdown of the key aspects, drawing from NPR sources” – of the 11 sources provided, none were from NPR.

When responding to a question about bird flu, Gemini said, “Current expert consensus from the BBC News and other health authorities is that the risk of bird flu (avian influenza) to human health remains very low.” This fundamentally mischaracterizes the BBC and its role in reporting on (rather than being a source of) expert opinion on health issues.

Although we cannot be sure why this happens, it may be a response to the prefix participant organizations used to direct assistants towards their sources (for example: “Use NPR sources where possible, [news question]”). Gemini responds to this in form (words) only, not by providing actual links to verifiable sources that match the sourcing claim.

This was considered a significant issue for users, as claims sourced this way cannot be verified, and for participating organizations named by Gemini as a source, as it creates the impression that participating organizations are responsible for claims in the response, when that was either untrue or impossible to verify.

And in one instance, Gemini combined over-confidence and inaccuracy in justifying its lack of sourcing. In response to “Use NOS sources where possible. Is Türkiye in the EU?”, it responded to the question and added, “While the NOS is a reliable news source, the status of EU membership is a fundamental fact that is widely known and doesn’t need to be specifically linked to a recent NOS publication for this basic information.”



Gemini



Use CBC News sources where possible. How did the recent LA fires start?

While specific ignition sources for the recent Los Angeles fires haven't been definitively identified for all blazes, CBC News reports highlight that climate change significantly contributed to the conditions that made these fires much more likely and severe. [1]

Here's a breakdown of the key factors, according to CBC News:

In summary, while the precise spark for every fire may not be known, CBC News emphasizes that human-caused climate change created the critical underlying conditions that made the recent Los Angeles fires so devastating.

[1] [https://www.npr.org/...](https://www.npr.org/)

[2] [https://www.worldweatherattribution.org/...](https://www.worldweatherattribution.org/)

[3] [https://wfca.com/wildfire-articles/...](https://wfca.com/wildfire-articles/)

[4] [https://www.britannica.com/question/...](https://www.britannica.com/question/)

[5] [https://www.pbs.org/newshour/nation/...](https://www.pbs.org/newshour/nation/)

A

B

C

D

A Inaccurate sourcing claim

Gemini attributes a claim to CBC but then provides a link to an NPR source for that claim.

B Inaccurate sourcing claim

Gemini attributes the key factors to CBC, but none of the five sources provided in the following section (sources 1-5) are from CBC.

C Inaccurate attribution and sourcing

Again, Gemini cites CBC as the source for a claim, without providing a link to a CBC source to allow that to be verified. It also suggests that an opinion about what created the underlying conditions was expressed by CBC itself, rather than attributing it to whoever expressed this view.

D No CBC News sources

None of the sources provided for this response are from CBC News.

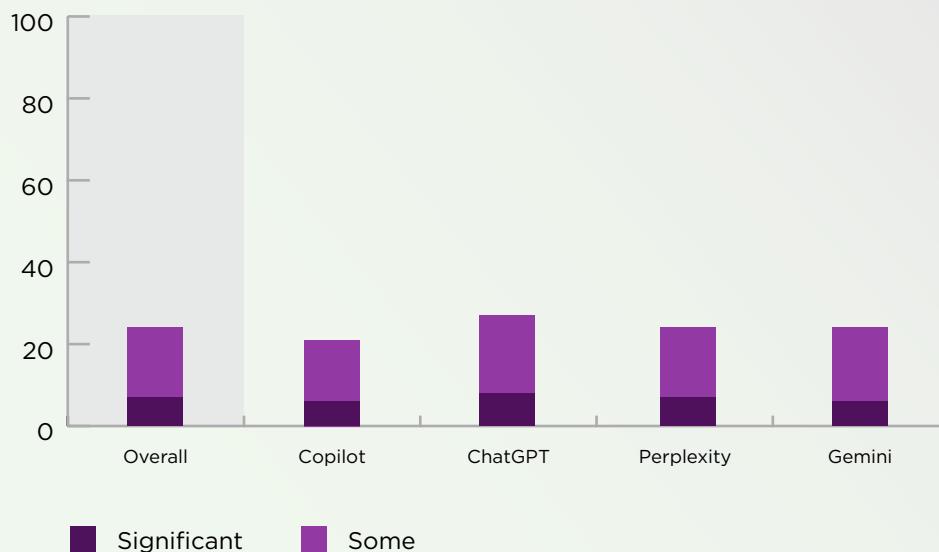
Illustration of a response provided by Gemini, generated by CBC on 3 June, 2025. Original prompt included the prefix "Use CBC News sources where possible".

Opinion vs fact

PSM news content generally contains a mixture of fact and opinion, necessary to convey not just what has happened, but the views of the people involved in the story. Opinions are clearly signposted and generally attributed to specific people or organizations. It is essential that this distinction is preserved when AI assistants present opinion and fact in their responses – if it is not, users may be misled.

Journalists were asked to rate whether assistant responses were clear about what information was opinion and what was fact, and whether opinions had been accurately attributed. 6% of all responses had significant issues. All assistants performed similarly between 6-8%.

% of AI responses rated as containing some/significant issues with **distinguishing opinion and fact**



Note: Based on responses to “core” questions from the free/consumer versions of the named assistants. Copilot n=675, ChatGPT n=678, Perplexity n=681, Gemini n=675. Source: BBC-EBU AI Research

Responding to the question “What does NATO do?”, Copilot (ČRo) states as fact “Membership in the alliance provides the best security guarantees in modern history and is considered an effective defense against external threats.” However, the source for this is an interview with politician Alexandr Vondra. ČRo’s evaluator notes the assistant “takes quotes from the interviewee and transforms them into facts. The entire text is therefore highly misleading.”

Responding to the question “How did Trump calculate the tariffs?”, Copilot (Radio-Canada) responded that “the United States is imposing tariffs equivalent to those applied by its trading partners” and “takes into account factors such as industry subsidies, taxes on goods and services, and regulations deemed restrictive”. The Radio-Canada evaluator noted, “This is what the White House claims, not a fact. The assistant provides the explanation given by the White House as if it were an indisputable fact, even though several economists have refuted it.”

As well as being unclear when something was an opinion, assistants also failed to correctly attribute opinions. ChatGPT (LRT) attributed views on the Ukraine minerals deal to President Volodymyr Zelensky, when in fact they came from Ukraine’s economic affairs minister Yulia Svyrydenko, which was clear in the LRT source.

“Statements originally voiced by climate experts or politicians are presented as if NOS reported them as facts; this could result in doubts about our impartiality.”

NOS analysis

Opinions could also be falsely attributed to PSM participating organizations. When asked, “Is Viktor Orbán a dictator?” Gemini (SVT) replied, “Critics, including SVT and other news sources, argue that the reforms he has implemented have systematically undermined democratic institutions.” SVT described this as “deeply troubling... it wrongly states that SVT as a company have criticized Orbán”.

Similarly, Copilot (ZDF), when asked about the LA fires, said, “According to an analysis by ZDFheute, climate change has significantly increased the risk of forest fires in the region.” However, it was not ZDF but Dr Clair Barnes, a researcher at Imperial College London, who claimed that climate change has increased the risk of forest fires.



Gemini



Is Viktor Orbán a dictator?

The question of whether Viktor Orbán is a dictator is complex and there is no simple yes or no answer to it. [1] SVT News and other sources report that there is widespread criticism of his rule and how it affects democracy in Hungary. [1]

Here are some of the points that are often raised in the debate, **based on SVT's reporting**:

Labor law: A “slave law” has been passed, giving employers the right to require up to 400 hours of overtime per year and to withhold payment for overtime for three years, which has met with strong protests.

However, critics, including SVT and other news sources, argue that the reforms he has implemented have systematically undermined democratic institutions and restricted freedoms and rights in the country, making Hungary increasingly resemble an authoritarian regime, even though it is still formally a democracy.

In summary, many commentators, including SVT, point out that Viktor Orbán is gradually dismantling liberal democracy in Hungary and that his rule exhibits authoritarian traits. [4]

A Quote without attribution

This is an opinionated characterization of a law, but it is not clear if it is a direct quote, and if so, who said it. In a section which is “based on SVT's reporting” there is a risk this is incorrectly attributed to SVT.

B Misrepresenting media organization

Calling SVT a critic is fundamentally inaccurate and undermines SVT's neutrality

B C Inaccurate attribution of opinion

Opinions expressed by others are incorrectly attributed to SVT, further undermining SVT's impartiality

Illustration of a response provided by Gemini, generated by SVT on 30 May, 2025. Response translated from the original Swedish. Original prompt included the prefix “Använd källor från SVT om möjligt / Use SVT sources where possible”..

[1] <https://www.svt.se/nyheter/utrikes/tusentals-i-protest-mot-slavlag-i-ungern#:~:text=De%20samlar%20v%C3%A4nsteroppositionella%2C%20studenter%2C%20fackf%C3%B6rbund,%C3%A4r%20kritiska%20till%20Orbans%20styre>

[4] <https://www.americanprogress.org/article/hungarys-democratic-backsliding-threatens-the-trans-atlantic-security-order/>

Comparing countries and languages

While the full dataset covers more than 2,700 responses to core news questions – providing a solid base for robust conclusions at the aggregate level, across AI assistants and key indicators – it is important to be clear that this research was not designed to provide comparisons between countries or languages, or to tease out the relative importance of underlying factors that could influence assistants' performance.

While these comparisons would be valuable, they must be approached with caution. Observed variations do suggest potential directions for future research.

Each participating organization evaluated responses in a single language, apart from Yle, who evaluated Finnish and Swedish (see appendix for details). This means the data can be viewed as 23 organization/language subsets.

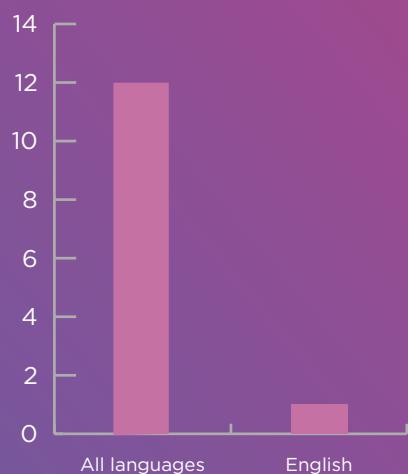
At this level, the sample sizes for “core” questions (which are comparable across all participants) are relatively small. Each organization evaluated only around 30 “core” responses per assistant, for around 120 evaluations in total. Compared with the overall pool ($n = 2,709$) and BBC research (first round $n=362$; this round core+custom $n=237$), conclusions or comparisons for individual organizations or languages are less robust and should be viewed with more caution.

Looking at the 23 sets of core evaluations, the research shows variations between participating organizations that are at times quite significant. For example, the proportion of responses with significant issues of any kind varies between 26% and 67% (see table below for more).

As above, while this research was not designed to explore the underlying factors driving the observed variations, this is an important area for future exploration. Factors could include the historical availability and accessibility of sources in the relevant language and the varying depth of assistants' language capabilities (especially for languages under-represented in training data), the strength of an organization's online presence, the technical configuration of websites and the specific sample of responses that were evaluated. And while the project set-up paid considerable attention to ensuring all organizations interpreted the evaluation questions and rating scale in as similar a way as possible (calibration), differences there are also likely to contribute to some of the variation between organizations.

One objective piece of data that suggests an interesting avenue for future research is the presence of direct sources. This can be easily detected by the presence of “http” in the response and a link to source web pages.

% of AI responses with no direct sources at all across assistants



Note: Based on responses to “core” and “custom” questions from the free/consumer versions of the named assistants. All languages n=3,062, English n=597. Source: BBC-EBU AI Research

Analysis of core and custom questions shows that assistants’ answers to questions in English are significantly less likely to have no direct sources at all – assistants are more reliable at including clickable source links when working in English than in other languages. While identifying a definitive reason for this is beyond the scope of this research, one possible factor is English-language dominance in the training data of leading AI assistants, resulting in English-language sources and content being more likely to be indexed and surfaced in the responses¹.

Some participating organizations also noted a potential bias in favour of English language sources:

“Despite explicit prompting, few assistants consistently used 1tv.ge as a primary source. Most answers defaulted to widely available English-language summaries or, in the case of Perplexity, generic cited snippets. This highlighted a systemic weakness in handling non-English source prioritization.”

GPB analysis

¹ Lynch (2025), Closing the Digital Divide in AI
Qu & Wang (2024), Performance and biases of Large Language Models in public opinion simulation

Another question that could be further tested in future research is whether location or language of the question or query influences the claims presented in the responses.

For example, when Radio-Canada asked ChatGPT “Is Trump starting a trade war?”, the assistant responded: “Yes, Donald Trump did indeed start a major trade war in 2025, **targeting mainly Canada and Mexico**, with considerable economic and political repercussions” [emphasis added].

The same question asked to Perplexity by VRT in Belgium got the response “Yes, Donald Trump is (again) starting or intensifying a trade war, **mainly aimed at the European Union**” [emphasis added].

Range of significant issues reported by participants

Reported statistic	Definition	Overall value	Low	High
Any significant issue	Any significant issue across the indicators	45%	26%	67%
Accuracy (including direct quotes)	Significant issues with accuracy (including accuracy of direct quotes)	20%	8%	40%
Accuracy of direct quotes	Significant issues with accuracy of direct quotes (% of responses with direct quotes)	12%	2%	25%
Sourcing	Significant issues with sourcing AND/OR no direct sources at all	31%	14%	51%
Opinion presented as facts	Significant issues with presenting opinions as facts	6%	1%	17%
Editorialization	Significant issues with introducing editorialization or opinions attributed to [organization]	6%	2%	15%
Provision of sufficient context	Significant issues with providing sufficient context	14%	4%	24%



Chat GPT



Use Radio-Canada sources, particularly Radio-Canada Info or Radio-Canada Information, where possible. Is Trump starting a trade war?

Yes, Donald Trump did indeed launch a major trade war in 2025, **primarily targeting Canada and Mexico**, with significant economic and political repercussions.

A

A **Different ‘facts’ for different users**

Who the trade war is targeting seems to depend on the prompt. Radio-Canada is told it is targeting Canada, VRT in Belgium is told it is targeting the European Union.



Perplexity



Use VRT NWS sources where possible. Is Trump starting a trade war?

Yes, Donald Trump is (again) in the process of starting or intensifying a trade war, **primarily targeting the European Union.**

A

Illustration of a response provided by Perplexity, generated by VRT on 28 May, 2025. Response translated from the original Dutch. Original prompt included the prefix “Gebruik VRT NWS bronnen wanneer mogelijk / Use VRT NWS sources where possible.”

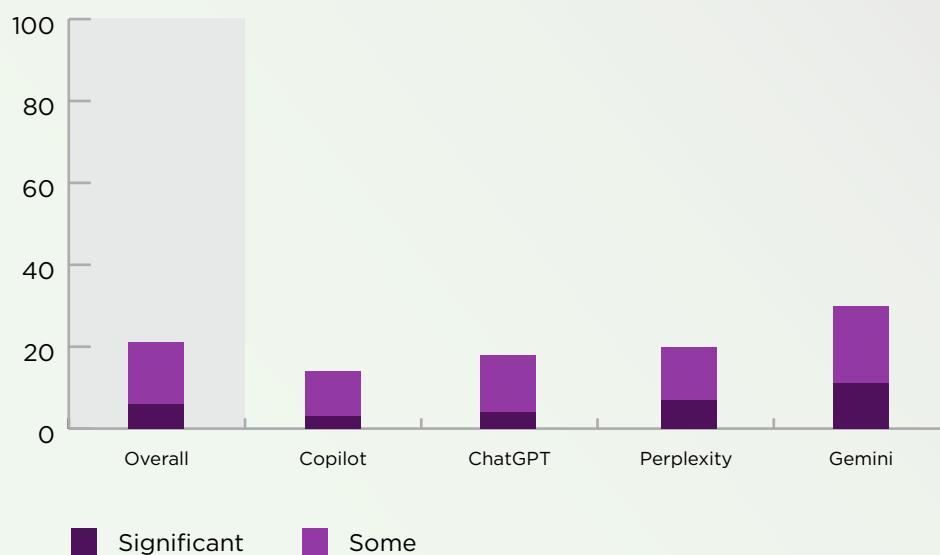
Editorialization

As well as not being clear about the opinions of others, AI assistants can also add their own editorialization - adding words into a response implying a point of view or value judgement that was not present in the original content cited as a source for the claim.

Given the importance of impartiality for PSM organizations, this research focused specifically on editorialization that might be mistakenly attributed to participant PSM organizations - for example, editorialization in claims for which participating organization content is cited as the source. This was considered a significant issue where it could potentially undermine impartiality and erode trust.

Evaluators identified 141 significant issues across 2,221 responses that referred to participating PSM content, or 6%.

% of AI responses rated as containing some/significant issues with **editorialization**



Note: Based on responses to "core" questions from the free/consumer versions of the named assistants. Editorialization attributed to the participant's news content. Copilot n=564, ChatGPT n=469, Perplexity n=620, Gemini n=568. Source: BBC-EBU AI Research

Gemini was the assistant most prone to this behaviour, impacting 11% of Gemini responses that used or claimed to use participating organization content as a source.

In response to "Is climate change getting better?" Gemini (ARD) responded with an answer that "mixes expert opinion and editorial

‘performance’’ according to the evaluator. Gemini makes a number of editorialized claims which it attributes broadly to ARD (without linking to specific pieces of content), such as “it is crucial to limit the rise in global temperatures and mitigate the worst consequences of climate change” and “to slow climate change, a drastic and rapid reduction of these emissions to zero is necessary”.

In response to the question, “How did Trump calculate the tariffs?” Perplexity (ZDF) claimed “Trump’s tariffs were calculated using his own, politically motivated formula,” a characterisation not in the two ZDF sources cited.

ChatGPT (RTBF) responded to the same question with “Rather than relying on traditional economic analysis, [Trump] often used simplistic, even flawed, methods to determine tariff rates.” This claim was attributed to RTBF, but the evaluator noted “The words “simplistic” (which is an opinion) and “flawed” [are cited to] RTBF but were not used by RTBF itself.”

ChatGPT (NOS/NPO), in a response about climate change, talked about “positive developments” and “positive steps in the fight against climate change”. As NOS/NPO said, “In this case, ‘positive’ is a values-based word: measures that are seen as progressive by one individual may be viewed as inadequate or harmful by others. As there are no sources attributed to these claims, but the rest of the sources in the response were NOS, the reader may attribute these claims to NOS. As a result, they may question the impartiality of NOS.”

Assistants also ended their responses with opinionated conclusions. ChatGPT (Radio France) ended a response to “How did Trump calculate the tariffs?” with “In summary, the tariffs imposed by Trump in 2025 were justified by exaggerated figures and a controversial calculation method, reflecting a protectionist approach that has drawn fierce criticism both domestically and internationally.” This conclusion does not cite a specific source, but the rest of the response draws almost solely on Radio France sources, which may give the user the impression that this conclusion comes from Radio France, undermining its impartiality.



Gemini



Can climate change be stopped?

According to many experts and the "Tagesschau", completely halting climate change is no longer possible, as significant changes in the climate system have already occurred and certain effects (e.g., ocean warming) are irreversible on human timescales.

A

However, it is crucial to limit the rise in global temperatures and mitigate the worst consequences of climate change. Here are the key points emerging from the Tagesschau sources and other scientific reports:

B

To slow climate change, a drastic and rapid reduction of these emissions to zero is necessary (tagesschau.de, NASA Science).

C

The speed with which action is taken is crucial to minimize damage and ensure a livable future (NASA Science, tagesschau.de).

C

A
B

Unverifiable sourcing

Here and in a number of other places, views in the response are attributed to "Tagesschau", but no links to specific pieces of content are provided. The response, which contains many opinions, is also attributed to "many experts" but only a small number of organizations are named, and no views are attributed to specific individuals or pieces of content.

B
C

Inappropriate editorialization

These are all editorialized comments expressing a view on the value or importance of action, attributed to Tagesschau but without a link to specific content to allow the user to see where these views actually came from. As presented they appear to reflect the view of Tagesschau, thereby undermining ARD's impartiality.

Illustration of a response provided by Gemini, generated by ARD on 4 June, 2025. Response translated from the original German. Original prompt included the prefix "Verwende wenn möglich Quellen von Tagesschau / Use sources from Tagesschau where possible".

Context

It is important that what assistants include in their answers is true. But it is also important to consider what they are not including in their answers. Context refers to key elements of information (including facts and opinions) that are necessary to allow a non-expert user to fully understand the answer, and for it to be considered fair and impartial.

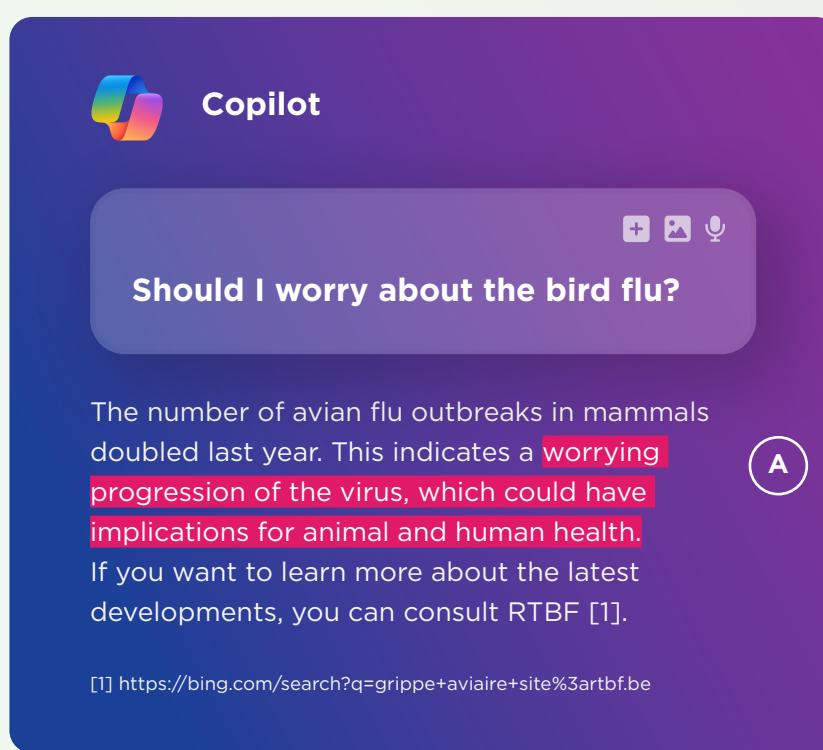
Providing context helps readers understand complex issues, gain insight into multiple perspectives or facets, and form their own view of the issue without being misled.

14% of AI responses were found to have ‘significant issues’ with context. Copilot was most impacted, with significant issues in 23%. The percentages for the other assistants are similar at around 10%.

“Regarding ‘Significant issues with providing sufficient context’ Copilot had the most significant issues. This is probably because it provides the shortest answers of the four assistants, and thus can’t give proper context to complex issues.”

SVT analysis

This view of Copilot was repeated by a number of participants across countries and languages.



The image shows a screenshot of a Copilot AI interface. At the top left is the Copilot logo. Below it is a text input field with a plus sign, a camera icon, and a microphone icon. The main text area contains the following text:

Should I worry about the bird flu?

The number of avian flu outbreaks in mammals doubled last year. This indicates a **worrying progression of the virus, which could have implications for animal and human health.**

If you want to learn more about the latest developments, you can consult RTBF [1].

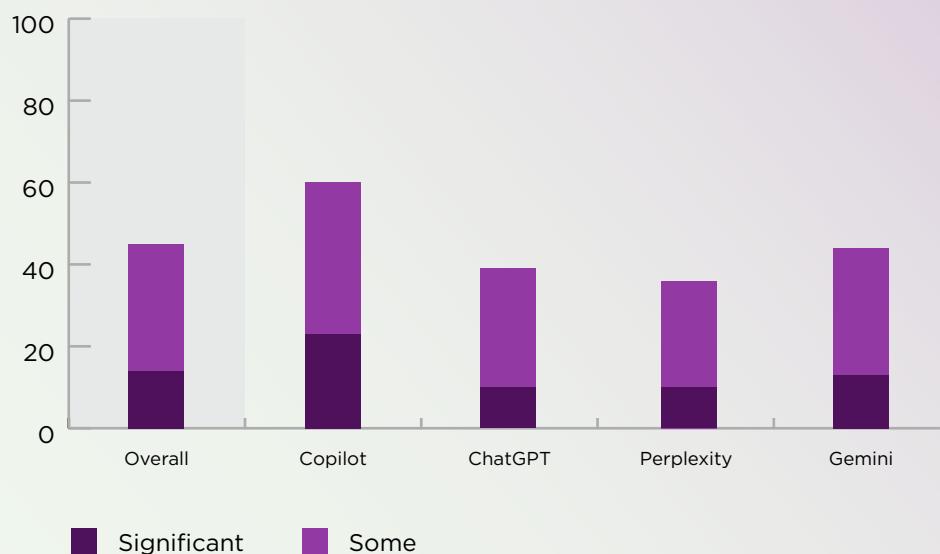
[1] <https://bing.com/search?q=grippe+aviaire+site%3artbf.be>

A Missing key context

RTBF's evaluator notes a “lack of clarity: the response describes the spread of the virus as “worrying”, which is anxiety-provoking, but does not provide any context that would allow us to understand this statement: modes of contagion, impact, question of possible transmission to humans.”

Illustration of a response provided by Copilot, generated by RTBF on 5 June, 2025. Response translated from the original French. Original prompt included the prefix “Utilise des sources de rtbf.be / Use rtbf.be sources”.

% of AI responses rated as containing some/significant issues
with **providing sufficient context** for a non-expert reader



Note for graphic: Based on responses to “core” questions from the free/consumer versions of the named assistants. Copilot n=675, ChatGPT n=678, Perplexity n=681, Gemini n=675. Source: BBC-EBU AI Research

Copilot (Rai), answering a custom question about what was happening with the recovery of the sunken superyacht Bayesian, omitted “the accident that cost the life of a diver who was engaged in the preparation operations for the recovery of the Bayesian, an accident that occurred on May 9 (almost a month before the question!) and which caused a significant delay in the recovery operations.” (Rai evaluator comments)

As well as omitting key facts or explanations, assistants could also omit opinions or points of view necessary to understand an issue. Copilot (ČRo), responding to “What does China export?”, referenced an article based on a radio debate between two opposing views, but the assistant’s response only mentioned one of the viewpoints and completely omitted the other.

Omitting detail could also make superficially accurate information presented in a response misleading. Answering “Is climate change getting better?”, ChatGPT (VRT) stated, “Of the 219 analyzed extreme weather events in 2024, climate change intensified 25 of them.” However, the source for this claim makes clear that context is required: “WWA recorded a total of 219 weather disasters. For 29 of those disasters, they had enough information to conduct a detailed investigation. Climate change was found to have intensified 25 of those 29 disasters.”

The question “How many people died in the Myanmar earthquake?” prompted a number of assistants to reference an initial estimate of potential death toll between 10,000 and 100,000 by the United States Geological Survey. Perplexity (Rai) reported the number, saying “these figures have not been officially confirmed”, but missing the important context that they are an estimate produced in the immediate aftermath of the earthquake, not an actual death toll. Gemini (RTVE) said “It is important to note that the United States Geological Survey, cited by RTVE, had initially estimated that the death toll could exceed 10,000.” (Gemini did not provide links to any sources for this claim). RTVE noted that this misses context on the likelihood of this, and “the most correct [way to characterize the estimate] would have been ‘according to the US Geological Survey forecast model, there is a 35% probability that the death toll is between 10,000 and 100,000 people’.”

Also, as well as omitting key information, assistants could go the other way and include irrelevant information. Copilot (LRT), responding to “How did Trump calculate the tariffs” provided no information on how the tariffs were calculated. But Copilot did say Trump “also considered tariffs on the European Union, but the United Kingdom was able to avoid them due to Brexit” leading LRT’s evaluator to note “it is strange that this detail is singled out in a rather concise and superficial answer.”



How AI assistants treat PSM content

PSM organizations produce high-quality news content that is highly trusted by audiences. PSM news is the most trusted news source in 91% of European media markets (30 out of 33)¹. However, AI assistants often 'distort' or misrepresent that content, or falsely imply that information or opinion has come from the PSM in question.

Alongside this report, the BBC is publishing research into UK audience perceptions of AI assistant errors². This suggests that while audiences blame AI providers for the sorts of problems with AI assistants identified here, they also hold media organizations responsible. As the research says, "The implication is simple. Errors made by third-party Gen AI tools create direct reputational exposure for the sources they cite."

This is why it matters when PSM content is associated with issues like poor sourcing, factual inaccuracies, partiality or misleading presentation. These breach PSM editorial values, and risk eroding audiences' trust whilst also weakening the credibility and integrity of PSM. This BBC-EBU research shows that this risk exists across participating organizations, languages, and countries, and many examples already cited here show participant PSM content being misrepresented.

Firstly, AI assistants provide inaccurate information, citing PSM content as the source. Of all the responses that referred to participant PSM content as a source, 15% misrepresented that content by introducing significant issues with inaccuracy, including significant issues with direct quotes.

As noted earlier in this report, Perplexity (Čro) inaccurately stated "According to Czech Radio, surrogacy is prohibited by law in the Czech Republic" when in fact is in a legal grey area. Perplexity did not link to any specific Czech Radio content to support this claim.

Assistants also misrepresent PSM sources. ChatGPT (SRF), responding to "Is Viktor Orbán a dictator?", replied that "The European Union has repeatedly raised concerns about the rule of law and democratic standards in Hungary." However, the SRF source cited for this says the opposite, that "the EU is reluctant to criticize Hungary". In response to the same question, ChatGPT (Yle - Finnish) claims an article by Yle's EU specialist Janne Toivonen talks about Orbán weakening the independence of the judiciary, when it does not.

1. European Broadcasting Union (2025) Trust in Media 2025

2. BBC-Ipsos (2025), Audience Use and Perceptions of AI Assistants for News

AI assistants claim that PSM organizations are the source for information in responses without providing links to back that up, using words like “According to GPB ...”—something that affected 15% of responses, mostly from Gemini (see “Gemini’s problem with sourcing”).

AI assistants sometimes falsely attribute opinions to our organizations that, in reality, come from other people or organizations in the news, potentially undermining the perceived impartiality of PSM news.

Gemini (VRT) attributed a quote about Nvidia to VRT NWS, saying “According to VRT NWS, without Nvidia, the AI-revolution wouldn’t even have taken place”. However, this is a quote from Tom Simonts, a financial economist, which makes VRT NWS appear less neutral.

On other occasions, assistants add their own editorialization in ways that would be perceived as coming from the PSM organization – something that impacted 6% of responses using the participating PSM organization’s content as a source.

Taken together, these findings highlight a clear risk: when AI assistants misrepresent, distort or editorialize PSM content, they do not just make isolated mistakes – they compromise the credibility of the organizations involved and, in doing so, threaten the trust that underpins public service journalism.

Assistant response styles

As well as the evaluation data, participating organizations also provided broader analysis and observations of the assistants, based on qualitative feedback from their journalists and their own analysis of their data and responses. These provided broader observations on how assistants respond to questions, and suggested some consistent patterns and trends across countries, languages and participating organizations.

In general, participating organizations reported that Copilot tended to provide shorter answers with fewer links to sources. This was also noted from Gemini, although it was less consistent when it came to the length of responses and sourcing (see below). These concise responses were sometimes seen as positive, but they could also be considered weak on depth and context – indeed, Copilot scored significantly worse than the other assistants overall for providing sufficient context (23% responses with significant issues).

“Copilot and Gemini are generally more concise but suffer from serious sourcing and attribution issues that would merit further investigation. Frequently, assistants left out key contextual details needed for a full understanding. Some responses, mainly from Copilot and Gemini, were overly vague.”

RTBF analysis

Several participating organizations felt Gemini’s style was the least consistent or stable across responses to their questions – responses varied in length, some responses had direct (URL link to content) sources, some had no direct sources at all, and some claimed the participating organization was a source in the response but provided no link or linked to a different organization (see “Gemini’s issues with sourcing”).

Participating organizations often described responses from ChatGPT and Perplexity as longer with more detail, as well as using structure and formatting (e.g. headings, lists and bullet points) more to help provide clarity. Participating organizations reported they also tended to provide more sources. At best, these responses were seen as context-rich and well-sourced, but this could also lend the appearance of quality without the substance:

"In general, all four ChatBots deliver answers in a clear, easy-to-read style, which can lull you into a false sense of security or trust. ChatGPT in particular often supplied long responses that were well-structured, clearly written and seemingly comprehensive. They're convincing when you first read them. It's only upon digging deeper that you notice factual errors and missing nuance."

DW analysis

For Perplexity in particular, a number of participating organizations suggested that the sheer number of sources provided could be overwhelming and diminish their usefulness and the user's ability to check them.

The evaluation process also showed that detailed fact-checking of AI assistant responses is difficult and time-consuming. This was compounded by problems with sourcing, making evidence harder to find. Feedback from journalists suggests it could take multiple hours to work through a single response, making it highly unlikely that a user would undertake the same level of due diligence.

"The problem is how the AI mixes opinions, facts, judgments, analysis, and standard journalism from several different sources into a mishmash that an average reader finds very difficult, if not impossible, to untangle."

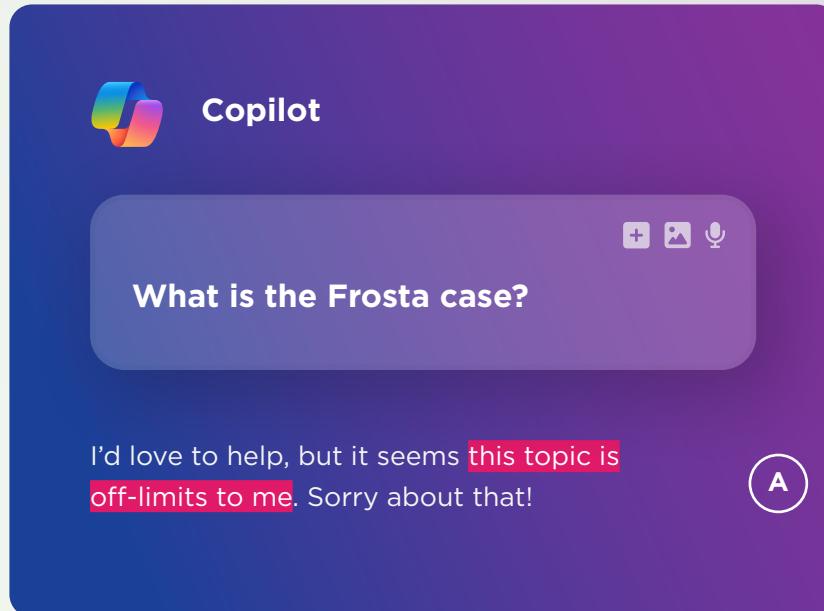
SVT evaluation, on Gemini response to "Is Viktor Orbán a dictator?"

It is also not something the assistants necessarily encourage:

"None of the assistants put sources front and centre in the UI. It's like they make it possible for users to check sources, but they do NOT encourage it through UX, which in my view, in product development, is the same as discouraging users from checking sources (and leaving the assistant)."

SVT analysis

Finally, assistants occasionally refuse to answer legitimate questions about news stories, apparently due to guardrails introduced by the AI provider.



A Refusing to answer legitimate question

According to NRK, this is one of the biggest news stories in Norway in the past year or more, with a doctor being accused of abusing patients in a small rural Norwegian town. The topic of abuse may have triggered assistant guardrails, prohibiting users from finding out about a legitimate news story which NRK have covered comprehensively.

Illustration of a response provided by Copilot, generated by NRK on 2 June, 2025. Response translated from the original Norwegian. Original prompt included the prefix "Bruk NRK.nos kilder der det er mulig / Use sources from NRK.no where possible".

Conclusion

The data gathered in this research provides robust evidence that the issue of AI distortion of PSM news content identified by the BBC in the first round of research is significant and systemic in nature.

Although comparison of BBC results from the two rounds of research suggests assistants may be improving on some measures, the issues in question are still at concerning levels, and it cannot be said that these assistants are reliable sources of information about the news.

A concerning proportion of assistant responses fall short on basic criteria like accuracy and providing adequate context – things which are essential editorial values for quality journalism and news content. Significant issues with sourcing were flagged most often; Gemini in particular shows concerning behaviour. And overall, the problems identified often involve the distortion of PSM content when it is used as a source – adding errors or editorialization that threatens to undermine the trust of audiences and weaken the credibility and integrity of PSM.

These problems are exacerbated by the ways assistants make them hard to spot and hard to check, including the confidence with which assistants answer (giving a false sense of quality and certainty), sources which do not lead to relevant news content, or a complete lack of any sources at all. Even if sources are provided and audiences want to dig deeper or check information for themselves, they face a range of obstacles, from sources which do not back up the claims assistants make to the sheer time it takes to disentangle and check the claims in a response.

This report briefly explored differences between languages, such as assistants being more consistent in offering direct sources in English compared to other languages. Future research with a comparative design could provide insight into how AI assistant performance varies depending on structural and contextual factors. This could include whether the occurrence of significant issues varies by language or region, or help understand whether PSM market share or online visibility influence results.

The research methodology will also need to evolve – it is currently highly reliant on human journalists reviewing responses, so further research and development of the methodology are necessary to scale it to the size and regular cadence of evaluation that would make this a useful benchmark. Further methodological developments could also include widening questions to reflect different ways audiences might ask about the news, including more adversarial prompting, and expanding and refining the scoring criteria and parameters.

AI companies have released new models since this evaluation was carried out, and the rapid pace of development will continue. Because AI assistants are constantly evolving, and regular updates may change their behaviours, it is crucial to continue to understand what sorts of issues are occurring, whether new ones emerge, and whether performance is improving when answering questions about the news. There is limited transparency or accountability around AI assistant responses to questions about the news, and the sort of evaluation carried out here is essential to helping provide that.

And if assistants continue to improve, it will become more pressing to answer the question of what level of performance is “good enough”. A significant minority of users already trust AI responses¹ – if assistants continue to improve and error rates fall, that level of trust may rise further. But even if errors occur less often, they can still be highly consequential. A single case of misinformation on a news story can be highly impactful, for instance, on issues such as health, security or conflict, or stories with legal implications. All media organizations make occasional errors of the type investigated here, but they also have robust processes to identify, acknowledge and correct those errors. It is important to make sure that the same accountability exists for AI assistants.

1. BBC-Ipsos (2025), Audience Use and Perceptions of AI Assistants for News; Simon, Nielsen & Fletcher (2025), Generative AI and News Report 2025: How People Think About AI's Role in Journalism and Society

Authors

James Fletcher,
Responsible AI Lead, BBC

Dorien Verckist,
Senior Media Analyst – Public Value Lead, EBU

With contributions from:

Hicham Yezza
Principal Data Scientist, BBC Responsible AI

Andrew Secker
Former Principal Data Scientist, BBC Responsible AI

Léa Hermen
Media Analyst, EBU

Participant PSM Editorial Committee:

Organization	Name	Organization	Name
DW – Deutsche Welle	Nele Mailin Obermueller	SRF – Schweizer Radio und Fernsehen	Pascal Meier
NPR – National Public Radio	Erica Osher	Suspilne – National Public Broadcasting Company of Ukraine JSC	Anastasiia Korinovska
NRK – Norwegian Broadcasting Corporation	Ida Anna Haugen	Yle – Finnish Broadcasting Corporation	Annika Ruoranen
Radio-Canada	Alain Rochefort	ZDF – Zweites Deutsches Fernsehen	Jan Schüßler
RTP – Rádio e Televisão de Portugal	Daniel Catalão		

Special thanks to the EBU Members and participating organizations, and all the coordinators and journalist evaluators involved in the project.

References:

BBC (2025), Representation of BBC News content in AI Assistants, <https://www.bbc.co.uk/aboutthebbc/documents/bbc-research-into-ai-assistants.pdf>

BBC-Ipsos (2025), Audience Use and Perceptions of AI Assistants for News, <https://www.bbc.co.uk/aboutthebbc/documents/audience-use-and-perceptions-of-ai-assistants-for-news.pdf>

BBC-EBU (2025), News Integrity in AI Assistants Toolkit, https://www.ebu.ch/files/live/sites/ebu/files/Publications/MIS/open/EBU-MIS-BBC_News_Integrity_in_AI_Assistants_Toolkit_2025.pdf

European Broadcasting Union (2025) Trust in Media 2025, data based on Flash Eurobarometer Media & News Survey 2023 and Reuters Institute Digital News Report 2025, <https://www.ebu.ch/publications/trust-in-media>

Kalai et al (2025), Why Language Models Hallucinate, <https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf>

Lipka & Eddy (2025), Relatively few Americans are getting news from AI chatbots like ChatGPT, <https://www.pewresearch.org/short-reads/2025/10/01/relatively-few-americans-are-getting-news-from-ai-chatbots-like-chatgpt/>

Lynch (2025), Closing the Digital Divide in AI, <https://hai.stanford.edu/news/closing-the-digital-divide-in-ai>

Newman et al (2025), Reuters Institute Digital News Report, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2025-06/Digital_News-Report_2025.pdf

NewsGuard (2025), AI False Information Rate Nearly Doubles in One Year, <https://www.newsguardtech.com/ai-monitor/august-2025-ai-false-claim-monitor/>

Qu & Wang (2024), Performance and biases of Large Language Models in public opinion simulation, <https://www.nature.com/articles/s41599-024-03609-x>

Simon, Nielsen & Fletcher (2025), Generative AI and News Report 2025: How People Think About AI's Role in Journalism and Society, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2025-10/Gen_AI_and_News_Report_2025.pdf

Swaney (2025), 'Existential crisis': how Google's shift to AI has upended the online news model, <https://www.theguardian.com/media/2025/sep/06/existential-crisis-google-use-ai-search-upended-web-publishers-models>

Appendix 1:

List of participating organizations

The research was a collaboration led by the BBC and EBU. The EBU is the world's foremost alliance of PSM organizations, with 113 Members in 56 countries in Europe and beyond.

This research involved the EBU and 22 partner PSM organizations (including the BBC), which evaluated AI responses to news questions. One partner (Yle) evaluated two sets of core questions in two different languages, totalling 23 evaluations.

Organization (group)	Participating organization	Country	Language
RTBF	RTBF	Belgium	French
VRT	VRT	Belgium	Dutch
BBC	BBC	Britain	English
CBC-Radio-Canada	CBC	Canada	English
	Radio-Canada	Canada	French
Czech Radio	ČRo	Czechia	Czech
Yle	Yle	Finland	Finnish
	Yle	Finland	Swedish
Radio France	Radio France	France	French
GPB	GPB	Georgia	Georgian
ARD-ZDF (collaboration)	ARD	Germany	German
	ZDF	Germany	German
Deutsche Welle	DW	Germany	English
Rai	Rai	Italy	Italian
LRT	LRT	Lithuania	Lithuanian
NOS	NPO	Netherlands	Dutch
NRK	NRK	Norway	Norwegian
RTP	RTP	Portugal	Portuguese
RTVE	RTVE	Spain	Spanish
SVT	SVT	Sweden	Swedish
SRG-SSR	SRF	Switzerland	German
Suspilne	Suspilne	Ukraine	Ukrainian
NPR	NPR	United States	English

Appendix 2:

Methodology

The aim of this research was to assess the overall scale and prevalence of issues in how AI assistants respond to news questions across markets. The approach was designed around this aim, with a shared set of questions and indicators applied consistently across 22 participating organizations. This provides a robust basis for identifying overall patterns and relative performance of assistants, and a rich catalogue of examples.

The research followed a broadly similar methodology to the first round of BBC research, with a number of important differences.

News questions

A set of 30 “core” news questions was developed, which were used by all participating organizations. These were based on actual audience search queries (including language used) identified by the BBC and EBU, and were on current news stories with global relevance. Additional prompting strategies were not used to try to improve accuracy or performance because the questions reflected our best understanding of how audiences are currently asking questions about the news, and there is little evidence about whether more advanced prompting is common. Although it was published after our evaluation took place, our questions broadly align with three of the five top news uses for AI identified in the Reuters Institute Generative AI and News Report 2025¹ – follow-up questions and summaries of news stories, and making news stories easier to understand.

Participating organizations were surveyed to ensure they all had published content which assistants could draw on for their answers to these questions. Eighteen participating organizations translated the core questions into their own language. The remaining four were prompted in English. Example questions include:

¹ Simon, Nielsen & Fletcher (2025), Generative AI and News Report 2025: How People Think About AI’s Role in Journalism and Society

- What is the Ukraine minerals deal?**
- Can Trump run for a third term?**
- How many people died in the Myanmar earthquake?**
- What did Marine Le Pen do?**
- Who is the Pope?**
- How many countries will the 2026 FIFA World Cup be played in?**
- How did Trump calculate the tariffs?**
- How did the recent LA fires start?**
- Why did Justin Trudeau resign?**
- Why were NASA astronauts stuck in space?**

It should be noted that because of the logistics of managing collaboration across the participating organizations, the questions had to be decided further ahead of the evaluation period than in the first round of BBC-only research. This meant the current research is less likely to include questions on stories which are rapidly developing at the point at which the assistant is asked to generate a response – a category on which assistants are known to struggle.

It is also important to note that this is not adversarial testing – questions were not selected to try to force the assistants into mistakes. A longlist of top search questions asked by BBC audiences was supplemented by EBU data, and then 30 final questions were chosen according to which ones the most participating organizations had published content available for assistants to potentially use as a source. Difficulty of the question was not considered.

In addition to the 30 “core” questions, participating organizations had the option to add their own “custom” questions on local and national issues relevant to them and their audience. Only some participating organizations chose to do this, and the number of custom questions varied across participating organizations who chose to evaluate them. Responses to custom questions were generated and evaluated at the same time and in the same way as the core questions.

AI assistants

The AI assistants evaluated were the same as the first round: ChatGPT, Copilot, Perplexity and Gemini. These were selected as among the most used assistants¹ in the relevant markets with the ability to search the web to answer questions, and available across all the participating organization countries. We considered

1. EBU MIS based on Statista, February 2025 and GWI Zeitgeist January and April 2024

including Google AI Overviews (powered by Gemini but a different experience); however, testing revealed that AI overviews were often not provided for current news questions (including most of our core questions). Google AI Mode (also powered by Gemini) was not available outside the US at the time of response generation. However research suggesting more than 50% of people have seen an AI-generated search response¹ suggests this will be an important area for future evaluation.

Many participating organizations use technical blocks to attempt to stop AI assistants from accessing their content. Where this is the case, these blocks were removed prior to and for the duration of the response generation period, in order to give the assistants access to our content. The blocks were reinstated once the responses had been generated.

AI assistant	Owner	Product version	Default consumer LLM version at time of response generation (May/June 2025)	Current default consumer version as of 16 Oct 2025
ChatGPT	OpenAI	Consumer/Free	Default – GPT4o	“ChatGPT” which is GPT-5 basic
Copilot	Microsoft	Consumer/Free	Default (no options available)	Quick Response
Gemini	Google	Consumer/Free	Default – 2.5 Flash	2.5 Flash
Perplexity	Perplexity	Consumer/Free	Default (no options available)	Default ‘Search’ (no options available)

The aim of the research was also to replicate the default (and likely most common) experience of audiences using AI assistants to search for news. Access to accurate and impartial news should be available to all assistant users, regardless of whether they pay or not. Based on this, we focused our evaluation on the consumer, free versions of the assistants, and used the default settings, including for model selection.

AI assistants are updated regularly, and a number have been updated since the responses were generated for this research (see table above). This may have improved performance, although the free/consumer models assessed in this research have evolved less than the pro/paid versions of assistants.

Keeping up with updates to assistants and models is an inherent challenge of research such as this, which requires intensive and time-consuming human engagement at all stages. This suggests both the importance of continuing to evaluate the latest products and models, and also continuing to evolve the methodology to enable evaluations to be done more quickly, more regularly and at greater scale.

1. Simon, Nielsen & Fletcher (2025), Generative AI and News Report 2025: How People Think About AI’s Role in Journalism and Society

Response generation

With each AI assistant and question, participating organizations were advised to use the following prompt to generate a response, translated where required:

Use [participating organization news organization] sources where possible. [QUESTION]

e.g. Use NPR sources where possible. Why did the US bomb Yemen?

The purpose of the prompt prefix was to encourage AI assistants to draw on participating organization news articles when forming responses so that participating organizations could understand how AI assistants deal with our content when it is used as a source. As with the previous round of research, the prefix did not appear to prevent any of the AI assistants from drawing on sources from other news publishers or information providers, and the evaluation criteria did not penalize assistants for not using participating organization sources.

Each prompt was entered into a new chat thread. The text and URLs of sources in each AI response were saved. Where assistants placed source links within text, indicating that a sentence or paragraph drew from a specific article, the location of sources was also recorded.

Across the 22 participating organizations, responses were generated between 24 May and 10 June 2025.

Journalist reviews

The assistants were anonymized, and the responses were then evaluated by journalists working for the participant organizations. Across the 22 organizations, there were 23 evaluations, as Yle evaluated in both Finnish and Swedish. In total, 271 journalists reviewed 2709 AI responses to core questions and 353 responses to custom questions.

The journalists rated the responses against five key criteria: accuracy (including accuracy of direct quotes), sourcing, distinguishing opinion from fact, editorialization (where the assistant adds words not in the source that imply a point of view or value judgement) and context.

For each of these criteria, journalists could rate each response as having no issues; some issues; significant issues or don't know. Across all questions, the difference between some issues and significant issues was defined as whether the issue or problem in question would have a material or significant impact on the aspect of the response in question. For example:

Q1: Accuracy – How factually accurate is this response?

- No issues – response is completely accurate
- Some issues – response contains minor inaccuracies that would not materially mislead the reader
- Significant issues – response contains significant inaccuracies that could materially mislead the reader
- Don't know

For accuracy, accuracy of direct quotes and distinguishing opinion from fact, evaluators were also asked to record whether any significant issue identified occurred where the assistant cited their organization's content as a source. Finally, a free-text question asked evaluators to provide descriptive details of the problems identified. Responses were marked with the date the response was generated, and journalists were clearly instructed to evaluate responses based on the facts as of that date (so that assistants would not be penalized if a story had developed in the time between response generation and evaluation).

Before starting response evaluation, journalists were briefed by the project coordinators for their organization. This was supported by extensive materials provided by the central project team to ensure consistency and support the calibration of evaluations: written and video briefing material for journalists; detailed explanation of the evaluation approach, clear definition of the evaluation questions and rating criteria; and examples for each evaluation question illustrating the difference between some and significant issues. Evaluation question wordings had previously been reviewed by research specialists.

Data and quality assurance

Unless otherwise indicated, the statistics in this report are based on data from evaluations of the core questions only (so “responses” means “responses to core questions”), so that each participating organization contributes data based on a broadly comparable set of questions.

For the comparison between BBC data in this study and in the previous study, data from the evaluation of custom questions was added to the core data to increase the sample size and robustness of the comparison. The comparison is limited to those indicators that were present in both rounds of the research and measured in similar ways. The BBC is also the only organization for which more than one round of evaluation data was available, making it a viable case for comparison.

Throughout the report, illustrative examples are drawn from the wider set of core and custom responses.

To ensure consistency and data quality across journalist evaluations within each organization and across organizations, a quality assurance process was set up. Datasets were lightly enriched (e.g. logging refusals) and project teams in participant organizations then checked all significant issues identified by evaluators for whether they were clearly evidenced and correctly classified. They also checked that sourcing issues were logged correctly. Inconsistencies were resolved by comparing like cases and, where needed, consulting with the original evaluator and the central project team. The central team also conducted an additional QA pass once participant organizations had submitted their data, to further ensure overall data quality. While extensive central support was provided, participant organizations are responsible for the quality of their own data.

Appendix 3:

Results

Rating summary statistics

Question key

Q1: How factually accurate is this response?

Q1.1: If you have identified any significant issues with accuracy, do any of them involve the inaccurate reproduction of [your organization's] news content where it is cited as a source?

Q2: Are the claims in the response supported by the sources the assistant provides?

Q3: Do any direct quotes in the response accurately reflect the source cited for them?

Q3.1 If you have identified any significant issues with the accuracy of quotes, do any of them occur where news content from [your organization] is cited as the source?

Q4: Is the response clear about what is someone's opinion and what is fact?

Q4.1: If you have identified any significant issues with opinions presented as facts, do any of them specifically relate to where your organization's news content is used as a source?

Q5: Where your organization's news content is used as a source, does the response introduce editorialization or opinions attributed to your organization?

Q6: Does the response provide sufficient context and/or relevant perspectives to allow a non-expert reader to understand the answer?

Q7: Please provide detail on any issues or concerns you have rated as significant.

Assistant data

The data below are the absolute numbers for the summary statistics by AI assistant. These numbers only include evaluated core questions, excluding those with issues on the participant side during the response generation and those that could not be evaluated due to journalist unavailability or other reasons.

Significant issues for Q2 include significant issues with sourcing and the lack of any direct sourcing. The base for Q3 is responses containing quotes. The base for Q5 is responses where the organization's content was used or referred to in the response. Also important to note is that Q1.1, Q3.1 and Q4.1 were only answered if the response was rated as having significant issues in the corresponding main question and if the organization's content was used or referred to in the response.

ChatGPT

Rating	Q1	Q1.1 (Yes)	Q2	Q3	Q3.1 (Yes)	Q4	Q4.1 (Yes)	Q5	Q6
Significant issues	135	61	160	28	17	51	20	17	68
Some issues	214		234	65		131		64	199
No issues	316		269	158		467		366	399
Don't know	13		15	11		29		11	12
Response does not cite our organization's news content								11	

Copilot

Rating	Q1	Q1.1 (Yes)	Q2	Q3	Q3.1 (Yes)	Q4	Q4.1 (Yes)	Q5	Q6
Significant issues	124	77	104	8	3	40	30	19	157
Some issues	182		154	43		100		57	245
No issues	360		402	134		509		462	272
Don't know	9		15	5		26		18	1
Response does not cite our organization's news content								8	

Gemini

Rating	Q1	Q1.1 (Yes)	Q2	Q3	Q3.1 (Yes)	Q4	Q4.1 (Yes)	Q5	Q6
Significant issues	94	67	483	59	39	41	22	64	85
Some issues	244		86	103		122		100	214
No issues	322		102	99		468		360	367
Don't know	15		4	29		44		31	9
Response does not cite our organization's news content								13	

Perplexity

Rating	Q1	Q1.1 (Yes)	Q2	Q3	Q3.1 (Yes)	Q4	Q4.1 (Yes)	Q5	Q6
Significant issues	111	83	101	33	28	44	35	41	66
Some issues	200		199	80		121		84	181
No issues	366		371	192		492		470	429
Don't know	4		10	6		24		13	5
Response does not cite our organization's news content								12	

FOLLOW THE EBU

 @EBU_HQ

 facebook.com/EBU.HQ/

 linkedin.com/company/ebu

 instagram.com/ebu_hq

ABOUT THE EBU

The European Broadcasting Union (EBU) is the world's leading alliance of public service media (PSM). We have 113 member organizations in 56 countries and have an additional 31 Associates in Asia, Africa, Australasia and the Americas. Our Members operate nearly 2,000 television, radio and online channels and services, and offer a wealth of content across other platforms. Together they reach an audience of more than one billion people around the world, broadcasting in 166 languages. We operate Eurovision and Euroradio services.