EBU
OPERATING EUROVISION AND EURORADIO

BBC

News Integrity in AI Assistants

TOOLKIT

October 2025

EBU

BBC

Contents

Introduction

This Toolkit is a companion resource to the BBC/EBU report *News Integrity in AI Assistants: An International PSM Study*, evaluating how AI assistants answer questions about the news. In June-July 2025, Research participants from 22 Public Service Media (PSM) organizations analysed and evaluated more than 3,000 AI assistant responses to news-related questions, identifying hundreds of examples of how assistants get things wrong.

This Toolkit, which is intended to be a self-contained and evolving resource, has been developed by the BBC/EBU to help address two key questions raised by the research's findings: "What makes a good AI assistant response to a news question?" and "What are the problems that need to be fixed?

What is the problem we are seeking to address?

The BBC/EBU research clearly shows that AI assistant responses fall short of the standards of high-quality journalism, with 45% of responses having a significant issue – something which could materially mislead the user – of some form. The findings also show that this is a systemic challenge: The issues affect all four assistants we evaluated (Open AI's ChatGPT, Microsoft Copilot, Perplexity and Google Gemini), across all 18 countries, 14 languages and 22 participant PSM organizations.

The research further shows that the ways in which AI assistants can fail range widely, and span issues with accuracy, sourcing, providing context, editorialization and beyond.

Why does this need addressing?

Accurate, high-quality news is a cornerstone of democratic societies. As AI tools play a growing role in how people search for and obtain their news[1], it is increasingly important that AI assistant responses to news questions are fit for purpose and can be trusted by their users.

Media organizations, too, need to be confident that whenever their content is used as a source for AI assistant responses, it is represented

1. Newman et al (2025), Reuters Digital News Report

fairly and accurately, protecting their brand and preserving the trust of their audiences.

And it is essential, in an age of increasing disinformation, that members of the wider public are informed and empowered news consumers. Increasingly, this will mean having a greater understanding of how things can go wrong with AI assistant responses, and what to look for when using them.

What is in this Toolkit and how to use it?

This Toolkit presents a structured approach to categorizing the issues identified in the research to provide clear, granular, actionable answers to two key questions: "What makes a good AI assistant response to a news question?" and "What are the problems that need to be fixed?

What makes a good AI assistant response to a news question?

The Toolkit outlines four key components that are necessary in any good AI assistant response: accuracy, providing context, distinguishing opinion from fact, and sourcing. This defines basic standards of quality that AI systems should aim for, setting out the values that underpin trustworthy news.

Use this:

- to get a high-level understanding of what "good" looks like for AI assistant responses to news questions
- as a baseline reference point for evaluating how assistants represent news content

What are the problems that need to be fixed?

The Toolkit goes on to present a granular taxonomy of 'failure modes' – the specific and nuanced ways in which assistants get it wrong. Showcasing the rich breadth of examples identified in the BBC/EBU research, this is a structured guide to the issues that need to be addressed to build better AI news responses.

Use this:

- to get a deeper understanding of the specific problems that show up in AI assistant responses. Each section begins with an introduction to the broad category of issues, followed by detailed descriptions and examples of each of the specific "failure modes" within that category.
- as a diagnostic tool to trace issues, design evaluations and develop improvements
- as a guide for AI literacy initiatives or newsroom training on specific issues. The examples make the sections especially practical.

< Back to Contents

This Toolkit is designed to be flexible. You can read it from start to finish to get a complete understanding of what good and bad look like in AI responses to news questions, or dip into the sections most relevant to your work or interests. It can be used for deeper analysis, to inform and guide technical development, for media literacy and general newsroom training, or simply to build a clearer understanding of how AI assistants handle news.

It is important to highlight that this Toolkit is not intended to be a definitive or exhaustive 'last word', but rather a contribution to the conversation between PSM organizations, technology companies and other stakeholders around how we build AI tools that help rather than hinder the public's ability to obtain accurate, reliably sourced news.

Who is this Toolkit for?

We think this Toolkit can be a valuable resource for key audiences, including:

Tech companies: This Toolkit can help give a sharper focus to industry efforts to improve assistant responses. It offers a detailed list of key issues that tech companies need to track and address in order for assistants to offer consistently high-quality responses to questions about the news.

Media organizations: As AI assistants become a more established way for audiences to consume news, media organizations have an important role to play in building media and AI literacy. They are uniquely placed to identify how AI assistants can get things wrong and educate audiences on what to look out for. This Toolkit provides a foundation for building that understanding and education.

Media organizations and journalists can also use this Toolkit to help evaluate whether AI technology works for them, both in terms of how assistants represent our content and whether AI tools are good enough to use in newsrooms.

Research community: This Toolkit provides a valuable resource for informing further research efforts, particularly around AI evaluation and benchmarking in a news context.

General public: This Toolkit offers a useful guide for engaged or curious members of the public to discover and explore the key issues to look out for when using AI tools for news-related queries.

< Back to Contents

What makes a good AI assistant response to a news question?

Underpinning this Toolkit is an overarching PSM conception of what constitutes quality journalism, centred on editorial values such as accuracy and fairness. The Toolkit is grounded in the findings of the present BBC/EBU research, which spans 18 countries and 14 languages, as well as long-established research from industry and academia.

The Toolkit identifies four key components of a good AI assistant response, which reflect the range and depth of the issues identified in the BBC/EBU research, as well as the editorial values shared by public service and other media organizations.

1. **Accuracy:** is the information provided by the AI assistant correct?

   This includes the accuracy of statements but also of direct quotes. Key factual details and information, such as names, numbers, dates, locations, etc., should be accurate. Events and relations should be characterized correctly. Quotes, whether full or partial, should match exactly the words used in the cited source. The person who said the words should be identified correctly.

2. **Providing context:** is the AI assistant providing all relevant and necessary information?

   The assistant should provide the relevant information and points of view that users need to understand the issue in question. The assistant should also accurately convey the level of certainty that is warranted about a particular statement.

3. **Distinguishing opinion from fact:** is the AI assistant clear whether the information it is providing is fact or opinion?

   The assistant should clearly distinguish between whether information is a statement of fact or of someone's opinion or perspective. The latter should be clearly and accurately attributed. This includes being clear where the assistant may be adding its own editorialization, views or judgement to the response, and to avoid any editorializing that could undermine

the impartiality or trust of an organization cited elsewhere in the response as a source.

4. **Sourcing:** is the AI assistant clear and accurate about where the information it provides comes from?

   Since AI assistant responses involve no direct human editorial oversight, it is important that users are able to check and verify claims that assistants make in their responses. The assistant should always provide sources to support key claims in its response. The sources should be appropriate, relevant and accurately described. Any attributions of claims, statements or direct quotes within the body of the response should be accurate and accompanied by appropriate sourcing.

This Toolkit also presents an additional section presenting a fifth category of 'operational' failure modes. These are issues that affect how assistants operate and perform in general, rather than in specific responses. It is important to highlight that the above list is not intended to be exhaustive or definitive.

< Back to Contents

What are the problems that need to be fixed?

A taxonomy of failure modes

The sections below provide a detailed overview of what needs to be fixed to get to the sort of good response outlined above. They take each of the key ingredients – accuracy, providing context, distinguishing opinion from fact, and sourcing – and provide a structured breakdown of the ways assistants can get them wrong.

Each key ingredient has its own section (e.g. Accuracy), where we identify and define the relevant issues (e.g. Accuracy of direct quotes), before working through the detail of the different "failure modes" (e.g. fabricated quotes) within that category. Each section offers a definition of the key category, an overview of why it matters, and a "what good looks like" checklist.

For each specific failure mode, we provide a definition and one or two examples from a selection identified by our BBC/EBU evaluators as illustrative of the issue in question. These "failure modes" are not mutually exclusive, and a single AI assistant response may combine multiple issues across the four categories.

Each research participant evaluated the same 30 core news questions, asked to all four assistants. Some participants also evaluated additional custom questions, focusing on local and national issues relevant to the participant organization. Examples from both core and custom datasets are included in this Toolkit.

Examples are referenced in the form **Assistant / Participant organization (core news questions unless custom is specifically mentioned) / News question.**

< Back to Contents

1. Accuracy

Definition:
The assistant's answer should be accurate, whether with reference to known truth about any facts, opinions and other information it contains, or in how it relays content from a cited source.

Why it matters:
Factual accuracy is critical to a good AI response. Without it the AI response is not only not fit for purpose but potentially harmful.

Faithfully representing how a particular source presents a fact or opinion is key to the accuracy and quality of an AI response.

Accuracy about causal relations in a news context carries special importance due to the legal and ethical implications.

What good looks like:
- All facts presented by the assistant are accurate
- All presented facts are up to date
- Information taken from cited sources is presented accurately
- Chronological and causal relations are presented accurately
- Relations between entities are presented accurately

In this section: **How assistants get accuracy wrong**

< Back to Contents

1.1 Fabricated facts (including 'hallucinations')

Definition: The answer contains facts or statements that are inaccurate with reference to known facts and appear to be completely or partially fabricated by the assistant and not relayed from a cited source.

Examples:

- **ChatGPT / BBC / Should I worry about the bird flu?**
  The assistant stated that "regions like Shropshire and parts of Dorset have implemented Avian Influenza Prevention Zones, requiring poultry to be kept indoors." However, as the BBC evaluator notes, "The UK measures are wrong... England, Scotland and Wales are all in an avian influenza prevention zone and the specific places mentioned eg Shropshire did not have (on date of generating the answer) orders to keep birds indoors. This statement is not true and does not appear in the cited BBC piece."
- **ChatGPT / CBC / Is Türkiye in the EU?**
  ChatGPT linked to a non-existent Wikipedia article on the "European Union Enlargement Goals for 2040". In fact, there is no official EU policy under that name. The response hallucinates a URL but also, indirectly, an EU goal and policy.

1.2 Lack of fidelity to sources

Definition: The AI assistant's answer fails to faithfully represent information attributed to a cited source.

Example:

- **Gemini / LRT / How did Trump calculate the tariffs?**
  Gemini claimed that tariffs imposed by Donald Trump on Canada and Mexico 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 instead were described as punishment for the two countries failing to take action on fentanyl trafficking.
- **ChatGPT / RTVE / What did Marine Le Pen do?**
  ChatGPT mentioned Marine Le Pen's successor as leader of the National Rally party, Jordan Barella, as a potential candidate in the 2027 French presidential election. Talking of Barella, ChatGPT said he lacked "experience", but according to the RTVE evaluator, "what the [source] suggests is that he lacks 'preparation' ... a very relevant nuance."

< Back to Section 1

1.3 Out-of-date information or statement

Definition: The answer contains information or statements that may have been true in the past but no longer are, in a way that renders the answer inaccurate. This is often (but not always) linked to the assistant citing to out-of-date sources (see 6.5).

Examples:

- **ChatGPT / Yle Swedish / Who is the Pope?**
  The assistant claims that "Pope Francis is the current leader of the Roman Catholic Church". The correct information, that Pope Francis had died and the current Pope was (when the question was asked in May 2025) Leo XIV was available in the cited source.
- **Copilot / CBC / What does NATO do?**
  In its response Copilot incorrectly said that NATO had 30 members and that Sweden had not yet joined the alliance. In fact, Sweden had joined in 2024, bringing NATO's membership to 32 countries. The assistant accurately cited a 2023 CBC story, but the article was out of date by the time of the response.

1.4 Inaccurate representation of chronology

Definition: The answer fails to represent chronological facts and relations accurately. This includes:

- Confusing the timeline of events, incorrectly presenting the order in which they took place relative to each other or to some historical milestone
- Failure to correctly place events relative to the point of the AI response generation (not recognizing that events have already happened at the time of the AI response generation or, conversely, claiming that future events, e.g. the 2028 US election, have already happened)
- Conflating chronologically separate editions/versions of the same event (e.g. the 2017 and 2019 UK general elections).

Examples:

- **Gemini / Radio France / Can Trump run for a third term?**
  The response treated Trump's re-election as a hypothetical possibility, even though the answer was generated after the 2024 election.
- **Copilot / Suspilne / Why were NASA astronauts stuck in space?**
  The answer, generated in June 2025, said that "[The Astronauts] are expected to return to Earth in February" 2025, which is an

< Back to Section 1

inaccurate, misleading and confusing answer, since the astronauts had already returned by the time the response was generated.

1.5 Inaccurate representation of causal relations

Description: The answer makes causal connections between actors, events or facts in a way that is not supported by evidence or presents an inaccurate or misleading cause-effect relationship.

Examples:

- **Copilot / ZDF / Did Elon Musk do a Nazi salute?**
  The assistant links public reactions to an alleged "Nazi salute" by Musk with his resignation as special adviser to Donald Trump, implying the former caused the latter. There is no evidence in public record of a causal link, and no such evidence is provided by the assistant.
- **ChatGPT / Radio-Canada / Is Trump starting a trade war?**
  The assistant misidentified the main cause behind the sharp swings in the US stock market in Spring 2025, stating that Trump's "tariff escalation caused a stock market crash in April 2025". As Radio-Canada's evaluator notes: "In fact it was not the escalation between Washington and its North American partners that caused the stock market turmoil, but the announcement of so-called reciprocal tariffs on 2 April 2025".

1.6 Inaccurate scope or generalization

Definition: The answer contains a claim in which the assistant draws an unjustified conclusion or misrepresents the scope of a fact, e.g. by presenting a UK-specific law as applying to the EU.

Examples:

- **Perplexity / ČRo custom / Is surrogacy legal in the Czech Republic?**
  Surrogacy is currently not regulated by law in Czechia, meaning that it is neither explicitly prohibited nor permitted. However, the assistant incorrectly presents this legal vacuum as a complete ban.
- **ChatGPT / BBC / Should I worry about the bird flu?**
  The assistant stated that: "There have only been seven cases of bird flu in the UK and Europe." However, this is the number for the UK alone.

< Back to Section 1

1.7 Incorrect representation of entities and relations

Definition: The answer inaccurately represents entities or relationships between them. Examples include misrepresenting geographical, institutional or political entities and relations.

Examples:

- **ChatGPT / Suspilne / Why change to the Gulf of America?**
  The assistant conflated the Persian Gulf with the Gulf of Mexico, writing that: "Changing the name of the Persian Gulf to the 'American Gulf' is a political move aimed at demonstrating the geopolitical influence of the United States and supporting its allies in the region."
- **Perplexity / LRT / How long has Putin been president?**
  The assistant states that Putin has been president for 25 years. As LRT's evaluator notes: "This is fundamentally wrong, because for 4 years he was not president, but prime minister", adding that the assistant "may have been misled by the fact that one source mentions in summary terms that Putin has ruled the country for 25 years".

1.8 Failure of reasoning or logic

Definition: The assistant's response implicitly or explicitly contains errors of reasoning or logic, such as linking premises to invalid conclusions.

Example:

- **Perplexity / GBP / Did Elon Musk do a Nazi salute?**
  In its response, Perplexity stated that: "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 the GPB evaluator noted, "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."

< Back to Section 1

2. Accuracy of direct quotes

Definition:
A key subset of accuracy, it refers to AI responses that contain one or more direct quotes. The assistant should present the words and who said them exactly and accurately.

Why this matters:
Attributing a specific verbatim quote to a speaker carries significant implications, including legal liability, and implies a much higher expectation of exactitude than simple paraphrasing.

What good looks like:
- The quote in the response contains the exact words a person said, in the same order
- The quote is attributed to the right person or organization
- A verbatim quote is correctly presented as such

In this section:

How assistants get direct quotes wrong

< Back to Contents

2.1 Fabricated quotes

Definition: The answer contains a direct quote attributed to a particular source which is completely invented rather than incorrectly rendered.

Examples:

- **Gemini / ZDF / Is Viktor Orbán a dictator?**
  Gemini 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.
- **Perplexity / BBC Custom / Why did Birmingham bin men go on strike?**
  Perplexity fabricated two 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. One of these appears under the heading "Key quotes".

2.2 Altered quotes

Definition: The answer contains a direct quote where the quoted words only partially match those in the source or ground truth – i.e. somebody did say something, but they did not say it exactly the way the response claims.

Examples:

- **ChatGPT / Radio-Canada / Is Trump starting a trade war?**
  The assistant quoted Canada's then-PM Justin Trudeau as using the verbatim description "stupid trade war", yet his original phrasing was "It's a very stupid thing to do." This alters the tone of the quote in a way that can be considered misleading. Radio-Canada's evaluator notes that several media outlets that quoted him said he had denounced a 'stupid' trade war, which is probably where the assistant's mistake came from.
- **Perplexity / ZDF / Is Türkiye in the EU?**
  The answer 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."

< Back to Section 2

2.3 Inaccurate or misleading speaker attribution

Definition: The answer contains a direct quote which is incorrectly or misleadingly attributed, either through attribution to the wrong speaker or a misleading description of the speaker.

Examples:

- **ChatGPT/ LRT / What is the Ukraine minerals deal?**
  A direct quote by Ukraine's then-economic affairs minister Yulia Svyrydenko, in which she described the Ukraine minerals deal as "balanced and fair", was wrongly attributed by the assistant to Volodymyr Zelensky, even though the cited source carries the correct attribution.
- **Perplexity/ BBC Core / Why does Zelensky not wear suits?**
  The assistant attributed a statement to a "commentator," but the person it quoted is a cousin of US Vice President JD Vance, which the response did not mention. The assistant's description is misleading in the context of the answer.

2.4 Inappropriate signposting of direct quotes

Definition: The answer includes a direct quote from a source without clearly or adequately indicating that it is a verbatim quote, for example by omitting quotation marks.

Example:

- **Perplexity / NPR / Why does Zelensky not wear suits?**
  According to the NPR evaluator, the assistant "fails to use quote marks to distinguish between direct quotes and paraphrases, which is misleading and can be considered plagiarism."

< Back to Section 2

3. Providing context

Definition:
The answer should contain all the relevant information required for the answer to be informative and not misleading. This includes relevant facts as well as key opinions and views.

Why it matters:
A good answer is not just about whether facts or opinions included are accurate. It is also about making sure key facts or opinions are not missing. Without the appropriate contextual information an AI answer is likely to be inaccurate, incomplete, lacking impartiality or misleading.

What good looks like:
- The answer provides the relevant facts and opinions necessary to make the answer informative and not misleading
- The answer provides context at the right level of detail for understanding the answer
- The answer accurately conveys the appropriate level of uncertainty around claims or statements
- The answer accurately characterizes the substance of the opinion
- The answer accurately and appropriately provides explains key entities or concepts

In this section:

How assistants get context wrong

< Back to Contents

3.1 Omitting significant/material detail (lack of factual completeness)

Definition: The answer omits factual details relevant to the question or topic in a way that renders the answer inaccurate or misleading.

Examples:

- **ChatGPT / VRT / Is climate change getting better?**
  ChatGPT wrote: "Of the 219 analysed extreme weather events in 2024, climate change intensified 25 of them." However, the cited source clarifies that of those 219 only 29 events had enough relevant data, of which 25 were found to have been intensified by climate change. The assistant's response failed to provide this crucial context, resulting in a misleading framing of the results.
- **Copilot / Rai custom / What is happening with the recovery of the Bayesian?**
  According to the Rai evaluator, the response 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."

3.2 Omitting significant/material viewpoint or opinion

Definition: The answer omits relevant or important viewpoints or opinions, in a way that makes the answer misleading and/or lacking due impartiality. This includes failure to contextualize the relative strength or weight of opposing opinions and implying false balance or consensus.

Examples:

- **Copilot / ČRo / What does China export?**
  The assistant referenced an article based on a radio debate featuring two opposing views, but the assistant's response only mentioned one of those viewpoints and completely omitted the other.
- **Perplexity / SVT / Why can't Ukraine join NATO?**
  The response mentions different claims (that Ukraine cannot join NATO during a war and that potential NATO membership was a reason for the invasion) without making it clear these are distinct positions put forth by opposing sides of the conflict, which blurs who is asserting what in a way that can be misleading.

< Back to Section 3

3.3 Irrelevant or off-topic information or response

Definition: The assistant replies with information that relates to a different and irrelevant topic or aspect of the query. This is especially problematic if essential, relevant information is also omitted.

Examples:

- **ChatGPT / Suspilne / Why does East Germany vote AfD?** According to Suspilne's evaluator, the assistant, "instead of a response, provided a guide to restaurants in Kyiv".
- **Copilot / LRT / How did Trump calculate the tariffs?** Copilot provided no information on how the tariffs were calculated, which was the the subject of the query. Instead, it stated that 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."

< Back to Section 3

4. Distinguishing opinion from fact

Definition:
The AI assistant should accurately and appropriately indicate whether a statement is an opinion or a fact, as well as provide adequate attribution for any opinions contained in the response.

Why it matters:
Opinions are fundamentally different from facts, and maintaining this distinction is crucial in a news context. Failure to clearly signpost fact from opinion can lead to answers that are inaccurate or misleading.

What good looks like:
- Responses should be clear about whether information they present is fact or opinion
- Opinions should be clearly signposted and conveyed accurately, and should come with appropriate attribution
- Attributions of opinions to specific organizations or individuals should be accurate and not misleading

In this section:

How assistants get distinguishing opinion from fact wrong

4.1 Failure to adequately signpost opinion

4.2 Misleading or incorrect attribution of opinion

4.1 Failure to adequately signpost opinion

Definition: The answer contains an opinion but presents it without a clear indication that it is such.

Examples:

- **Copilot / Radio-Canada / How did Trump calculate the tariffs?**
  In response to a question about Trump's tariffs, Copilot 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 that "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." Copilot's response failed to make it clear that these were the administration's own claims.
- **Copilot / ČRo / What does NATO do?**
  Copilot states as fact that "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 statement is an interview with politician Alexandr Vondra. ČRo's evaluator notes that the assistant "takes quotes from the interviewee and transforms them into facts. The entire text is therefore highly misleading."

4.2 Misleading or incorrect attribution of opinion

Definition: An approach to attribution of opinions that is inaccurate or misleading, including:

- Attributing an opinion to the wrong person or organization (including to a media organization reporting someone else's opinion)
- Inaccurate or misleading characterisation of an opinion or view and/or of the person or organization expressing them
- Vague, generic or complete lack of attribution in a way that impacts the quality of the answer

Examples:

- **Gemini / SVT / Is Viktor Orbán a dictator?**
  The assistant stated: "Critics, including SVT and other news sources, argue that the reforms he has implemented have systematically undermined democratic institutions." However, this opinion was not

< Back to Section 4

SVT's, and its evaluator described the response as "deeply troubling … it wrongly states that SVT as a company have criticized Orbán".

- **Copilot / ZDF / How did the recent LA fires start?**
  Copilot 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 made the claim that climate change has increased the risk of forest fires.

< Back to Section 4

5. Inappropriate or misleading editorialization

Definition:
The answer introduces opinions or an editorial slant in its own voice in a way that is misleading or non-transparent.

Why this matters:
Any editorialization introduced by the AI assistant without clear signposting is likely to mislead readers. It is also liable to being wrongly attributed to an organization whose content is being used as a source elsewhere in the response, potentially undermining that organization's reputation and audience trust.

What good looks like:
- Any editorialization by the AI assistant must be clearly signposted to the reader
- The assistant should clearly demarcate statements attributed to or directly sourced from individuals or organizations from those in its own voice

In this section:

How assistants get editorialization wrong

< Back to Contents

5.1 Inappropriate or misleading editorialization

Definition: The AI assistant presents assertions without clearly signposting that they represent an editorial stance in its own voice, and in a way that misleads the reader and potentially undermines the perceived impartiality of any other sources cited elsewhere in the response.

Examples:

- **Gemini / ARD / Is climate change getting better?**
  According to the ARD evaluator, Gemini responded with an answer that "mixes expert opinion and editorial 'performance'". Gemini made a number of editorialized claims which it attributed 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".
- **ChatGPT / RTBF / How did Trump calculate the tariffs?**
  ChatGPT responded 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."

< Back to Section 5

6. Sourcing

Definition:
The AI assistant's answer should include appropriate sourcing for key claims in its response.

Why this matters:
Accurate, relevant and appropriate sources are essential elements of a high-quality answer to a news question, as they enable fact-checking and verifiability. It's also important for media organizations whose content is used as a source that any such sourcing is accurate and not misleading.

What good looks like:
- The assistant provides sources that support all key claims in the answer
- The sources provided are relevant to the topic and question
- The sources contain the specific information they are cited in support of
- The sources are reasonably easy to find, open and check
- The sources are correctly attributed in the response
- The response provides appropriate disclosure about the affiliations or partisanship of sources

In this section:

How assistants get sourcing wrong

The assistant fails to provide sources to back up key claims in the answer

Sources are not relevant or appropriate to the topic and question

< Back to Contents

Sources do not contain the specific information they are cited in support of

6.9 Source does not contain or support claim

Sources are not easy to find, open and check

6.10 Sources are not easily accessible for verification

6.11 Hallucinated sources or links

Source attribution is inaccurate or misleading

6.12 Inaccurate claim about source availability

6.13 Inaccurate or unverifiable sourcing claim

6.14 Incorrect attribution of secondary/syndicated content

< Back to Contents

The assistant fails to provide sources to back up key claims in the answer

6.1 No sources provided

Definition: The assistant does not provide any direct sources at all to support the claims in the response.

Examples:

- **Copilot / DW / Is Viktor Orbán a dictator?**
  The response includes several claims such as "his leadership style has sparked significant debate in Europe," but the assistant provides no sources at all.
- **Gemini / VRT custom / Are wild boars dangerous?**
  The response makes several claims about wild boars, including that encounters with them are "very rare". While the response notes that "VRT News has published various reports on this", the assistant does not provide any sources.

6.2 Key claims are not sourced

Definition: The answer does not provide appropriate sources for one or more key claims.

Examples:

- **Perplexity / CBC / How long has Putin been president?**
  Perplexity responded with biographical information, including the naming of five of Putin's 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]."
- **ChatGPT / Yle Finnish / Is Viktor Orbán a dictator?**
  In its response ChatGPT states that Yle reporter Janne Toivonen "writes that since 2010, Orbán has systematically concentrated power by, among other things, restricting media freedom and weakening the independence of the judiciary". However, as Yle's evaluator notes: "In the [cited] article, Toivonen does not talk about weakening the independence of the judiciary, so there is no basis for this claim in the response."

< Back to Section 6

Sources are not relevant or appropriate to the topic and question

6.3 Irrelevant sources

Definition: The answer cites sources that are outside the topic of the question.

Examples:

- **Perplexity / Suspilne / Why does Zelensky not wear suits?**
  The assistant linked to a Suspilne piece about the costumes of the Ukrainian band Kalush at Eurovision that is completely unrelated to the question of Zelensky's attire and was not used in the body of the answer.
- **Perplexity / VRT / Why change to the Gulf of America?**
  Perplexity lists nine VRT sources in its response, including some that are entirely unrelated to the topic of the query, such as articles on the abolition of first-class train seats, power plants in the Netherlands, and a 2012 article on a mumps outbreak.

6.4 Inappropriate number of sources

Definition: The number of sources provided by the assistant is detrimental to the quality of the answer. This could be too many, or too few.

Examples:

- **Perplexity / NRK / Multiple**
  In response to the question "How many people died in Myanmar earthquake?" Perplexity appended a sources block with 19 URLs but only referenced three of the sources in the body of the answer. Similarly, it provided nine 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."
- **Gemini / BBC / Should I worry about the bird flu?**
  Gemini provides a list of 11 symptoms, each with its own individual source - a total of eight unique sources for the symptom list, pushing the response to 20 sources overall. Most of these sources are different pages from the US Centres for Disease Control (CDC) – the entire list of symptoms could be cited to one of the pages

< Back to Section 6

which covers all the main symptoms (although for a UK audience, an NHS page covering six of the symptoms, also cited, would be more appropriate).

6.5 Out-of-date sources

Definition: The answer cites a source that is out of date in a way that makes it unsuitable for answering the question.

Examples:

- **Copilot / BBC / Should I worry about the bird flu?**
  Copilot states that "a vaccine trial is underway in Oxford" but cites a 2006 BBC News Health page. As the BBC evaluator notes: "This is grossly inaccurate as it draws from an article from 2006 that in no way shape or form represents the current state of vaccine research for H5N1. Everything in this section represents a two-decades-old viewpoint of the virus and vaccine development and states things are currently happening that are in fact not".
- **Copilot / RAI / What does NATO do?**
  Copilot provided a video from 2022 and an article from 2014 as sources. 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".

6.6 Inappropriate category of source (thematic appropriateness)

Definition: The answer cites a source that is thematically inappropriate for answering the question, such as citing a satirical website to support a factual claim.

Examples:

- **Perplexity / NOS-NPO / Did Elon Musk do a Nazi salute?**
  The response linked to De Speld, a Dutch satirical news website, as a valid source for answering the question of whether Musk performed a Nazi salute. Furthermore, as the NOS-NPO evaluator notes, the assistant did so "without explicitly mentioning that its content is satirical", making the use of the source even more inappropriate in this context.
- **Perplexity / Radio France / Why do people hate Tesla?**
  Perplexity based part of its response on a satirical column, claiming: "The rejection of Tesla goes beyond the simple issue

< Back to Section 6

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." This was not suitable for a factual answer, and the response did not carry any mentions that this was a satirical point of view.

6.7 Sources with inadequate editorial control

Definition: The assistant makes inappropriate use of a source from a brand, publisher or organization with a record of inaccurate, unreliable, unverifiable or insecure content. This is inherently a question of judgement. This may increasingly include websites entirely or partly containing AI-generated content with no human editorial oversight.

Examples:

- **ChatGPT / GPB / Why does East Germany vote AfD?**
  The evaluator notes that the assistant "bases the response on Reddit rather than journalistic or academic sources". More specifically, 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". This source is dominated by user-generated content, often anonymous and lacks the editorial oversight or expertise needed for answering this type of question.
- **Gemini / RAI / Why do people hate Tesla?**
  In its response, Gemini cites opentools.ai, an AI-tools promotional site that also publishes AI-generated "news" posts.
- **Gemini / BBC / Why does Trump want to buy Greenland?**
  In support of its statement that "both the Danish and Greenlandic governments have consistently and firmly stated that 'Greenland is not for sale'", Gemini cites a Google Gemini chat share link, i.e. another AI-generated response produced by Gemini itself.

6.8 Inappropriate use of partisan sources

Definition: The assistant offers support for factual claims by citing sources linked to some of the actors or stakeholders involved in the topic in question, or sources with an explicit or known partisan affiliation, and does so without appropriate transparency or contextualization of the partisanship or affiliation (see opinion vs fact).

< Back to Section 6

Examples:

- **Gemini / RTBF / How long has Putin been president?**
  The response cites three sources: two from the Kremlin and one from the Russian embassy in Ireland. The evaluator observed "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.
- **Gemini / DW / What is Nvidia famous for?**
  Half the sources provided by Gemini 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. The DW evaluator 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."

Sources do not contain the specific information they are cited in support of

6.9 Source does not contain or support claim

Definition: The answer cites a source in support of a claim, but the source does not contain or support that claim.

Examples:

- **ChatGPT / SRF / What does China export?**
  The assistant provides precise numbers that cannot be found in either of the two cited sources. SRF's analysis notes: "it's totally unclear where the information comes from [...] So, for the journalist, it was nearly impossible to verify if the facts in the response are true or part of GPT's hallucinations."
- **Perplexity / ARD / Is Viktor Orbán a dictator?**
  In its response, Perplexity stated that ARD's Tagesschau.de "describes Orbán's rule as authoritarian and illiberal". But this is not found in the cited Tagesschau source.

< Back to Section 6

Sources are not easy to find, open and check

6.10 Sources are not easily accessible for verification

Definition: The answer cites sources that are difficult for users to access, preventing easy verifiability. This includes:

- Paywalled sources
- Sources in a language that is different to that of the answer, including bias towards English-language sources
- Sources linking to home or landing pages rather than specific articles
- Sources linking to search engine result pages, including Google or Bing

Examples:

- **ChatGPT / SRF / What does China export?**
  The response cites paywalled website Statista. Without an accessible alternative, readers cannot easily verify the information provided by the assistant.
- **Copilot / DW / Is Viktor Orbán a dictator?**
  The assistant cites a five-year-old German-language documentary to substantiate its English-language response. As DW's evaluator notes: "it is not possible for an English-language speaker using the chatbot to check that, unless they can find the YouTube subtitles. All the prompting for DW was done in English and all the answers generated were in English too."

6.11 Hallucinated sources or links

Definition: A source is provided for a claim but either the website or the specific URL provided does not exist and has never existed. This can lead to media organizations being wrongly suspected or accused of removing previously published content.

Examples:

- **Gemini / NRK / Can Trump run for a third term?**
  In its response to the question, Gemini provided URLs for NRK articles that do not exist. NRK observed that "in answers on 'polarizing' topics … this might give the impression that we have removed content without explaining why to our readers.".
- **ChatGPT / RTP / Why do people hate Tesla?**
  The assistant cited links to RTP articles that RTP's online team confirmed do not exist. This undermines confidence in the rest of the sourcing provided by the assistant.

< Back to Section 6

Source attribution is inaccurate or misleading

6.12 Inaccurate claim about source availability

Definition: The assistant inaccurately claims that a requested source provider (e.g. a media organization) has not published content on the topic.

Examples:

- **Perplexity / RTP / Why do people hate Tesla?**
  The assistant claims that RTP has no information on the topic even though RTP has published relevant articles. RTP's evaluator called this "a resounding failure".
- **Perplexity / NRK / How many people died in the earthquake in Myanmar?**
  The response said there was no NRK coverage of the topic, then added "so I provided search results". But those results in fact included NRK's URLs, which could be confusing for the reader.

6.13 Inaccurate or unverifiable sourcing claim

Definition: The assistant asserts in its response that an organization's content is the source for claims or facts in the response but provides no link to that organization's content.

Examples:

- **Gemini / DW / Is Trump starting a trade war?**
  According to the DW evaluator, the assistant "repeatedly mentions "DW and other sources" in some form or other, without using a single DW source. It goes so far as to say in what months we reported on the introduction of specific tariffs, but then goes on to give CBS as a source. CBS do not appear to have cited any of our reports in their article".
- **Gemini / CBC / How did the recent LA fires start?**
  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..." However, the five source URLs provided by the assistant do not include any from CBC News, and CBC evaluators were unable to find any of these specific statements in their content, outside of expert interviews.

< Back to Section 6

6.14 Incorrect attribution of secondary/syndicated content

Definition: The assistant does not adequately capture secondary attribution of sources within responses, such as news-agency content published (via syndication) by other news organizations.

Example:

- **Perplexity / RTBF / How many people died in Myanmar earthquake?**
  The assistant used the phrase "According to RTBF" when citing RTBF articles that were almost entirely relayed from news agencies, and which carried a joint byline (e.g. "[RTBF] with AFP"). According to RTBF's evaluator: "Assistants often blurred the line between RTBF content and agency dispatches, presenting AFP or Belga material as if it were original RTBF reporting."

< Back to Section 6

7. Operational Issues

Definition:
Beyond the main categories presented above, there are several additional failure modes which are more operational in nature and relate to issues with the AI assistant's general approach. Such failure modes include AI assistants being too sycophantic or over-confident in their tone, breaching professional or legal codes, using inappropriate language or simply refusing to answer legitimate news questions.

What good looks like:
AI assistant responses should:

- use the appropriate tone and language
- be reasonably consistent in the face of variations in prompting
- adhere to professional and legal codes and standards
- adopt an appropriate tone when answering questions that involve a significant degree of uncertainty
- adopt an appropriate level of guard-railing that is not too restrictive

In this section:

How assistants get operational aspects wrong

< Back to Contents

7.1 Inappropriate sensitivity to prompt wording, including sycophancy

Definition: The response appears too sensitive to, and influenced by, how the prompt is worded, in a way that can result in the assistant inappropriately, inaccurately or misleadingly tailoring its response to the user.

It must be noted that the susceptibility to sycophancy (responding in a way considered most likely to please the user) is a widely noted issue with AI assistants[1] and is especially pronounced when prompts include incorrect or leading assumptions. The BBC/EBU research did not focus on questions of this type, and therefore the examples we provide here are suggestive and not intended to illustrate this failure mode precisely.

Examples:

- **Multiple / Multiple / Is Trump starting a trade war?**
  When asked this question, assistants echoed the non-neutral framing of the question and also appeared to tailor the answer based on whether the prompt (via the prefix used) identified the nationality of the user. When Radio-Canada asked ChatGPT, the assistant responded: "Yes, Donald Trump did indeed start a major trade war in 2025, targeting mainly Canada and Mexico." The same question asked to Perplexity by VRT in Belgium elicited the response: "Yes, Donald Trump is (again) starting or intensifying a trade war, mainly aimed at the European Union."
- **Multiple / NOS / Is climate change getting better?**
  NOS notes, "When a question is formulated rather subjectively, conveys a certain bias or steers in a certain direction, e.g. 'Is climate change getting better?', the assistants (ChatGPT, Gemini & Perplexity) seem to respond in the context of that same subjectivity."

7.2 Refusal to answer legitimate news questions

Definition: The assistant refuses to answer legitimate questions about new stories, invoking reasons such as the topics being "sensitive" or "off-limits". This is generally the result of guardrails introduced by the AI provider. This can lead to AI assistants preventing users from accessing legitimate news answers.

1. e.g. Fanous, Goldberg, Agarwal, Lin, Zhou, Daneshjou & Koyejo (2025), SycEval: Evaluating LLM Sycophancy

< Back to Section 7

Examples:

- **Copilot / NRK / What is the Frosta case?**
  Copilot declined to answer the question on the grounds of the topic being "off-limits" ("Forbudt område"). As the NRK evaluator noted: "We found it strange that Copilot would not generate a response on the topic "Frosta-saken". This is one of the biggest news stories in Norway the past year or more, with a doctor being accused of abusing patients in a small rural, Norwegian town. NRK have covered this story intensely, but the most surprising part was that Copilot said it could not answer because it was 'off-limits'."
- **Copilot / RTBF / Did Elon Musk do a Nazi salute?**
  RTBF noted that "the system simply blocked and refused to answer ... it only replied, 'I'll check that for you. One moment', and then it was impossible to push it to say more."

7.3 Not adhering to journalistic ethics or standards

Definition: The answer contains content which represents a breach of legal, journalistic or ethical standards with potential legal ramifications. This can include breaches of libel laws or ethical expectations around naming victims.

Examples:

- **ChatGPT / NRK Custom / What are the charges against Gjert Ingebrigtsen?**
  The assistant named a young victim in the "Ingebrigtsen-saken" trial case, whereas Norwegian outlets generally refrained because of the victim's age. The Norwegian Press Code (Vær Varsom-plakaten)[1] says: "As a general rule the identity of children should not be disclosed in reports on family disputes or cases under consideration by the childcare authorities or by the courts."

7.4 Irrelevant or inappropriate language

Definition: The assistant replies in the wrong language to that of the query, or switches language mid-response.

Examples:

- **Perplexity / Suspilne / Why is Trump imposing tariffs?**
  The question was asked in Ukrainian, but Perplexity replied in Bulgarian, which is inappropriate for the audience and input context.

1. Pressens Faglige Utvalg (2021) Code of Ethics of the Norwegian Press

< Back to Section 7

7.5 Inappropriate tone

Definition: The assistant's tone is inappropriate in the context of the question asked. This includes the assistant engaging in inappropriate assertions or speculation about the users – such as implying that they are wrong or confused – that are liable to make them uncomfortable.

Example:

- **Gemini / RTP / Why were NASA astronauts stuck in space?**
  Despite the fact that two NASA astronauts spent over nine months on the International Space Station after their spacecraft malfunctioned, Gemini challenged the user's question stating "this is a misconception" and then listing "possible reasons for your confusion", including science fiction films, misinterpretation of delays or technical issues on missions, and misinformation.
- **Gemini / NOS / Is Türkiye in the EU?**
  By way of justification for not providing links to NOS sources, Gemini responded with: "While the NOS is a reliable news source, the status of EU membership is a fundamental fact that is widely known and does not need to be specifically linked to a recent NOS publication for this basic information."

7.6 Over-confident tone

Definition: The assistant presents information with a tone of authority and certainty that is likely to mislead the user about the level of certainty warranted by the facts available to the assistant.

Examples:

- **ChatGPT / RTVE / What did Marine Le Pen do?**
  The assistant states in its own voice that "Le Pen's situation represents a turning point in French politics" – phrasing which suggests the AI assistant was an authoritative expert voice on French politics.
- **Perplexity / RTVE / Is Trump starting a trade war?**
  The response states that "Donald Trump is not only starting a trade war; he has already escalated it since his return to the presidency in 2025". The assistant presents a highly opinionated assessment in a tone that suggests greater certainty than the facts warrant.

< Back to Section 7

Authors

Hicham Yezza
Principal Data Scientist, BBC Responsible AI

James Fletcher,
Responsible AI Lead, BBC

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

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-EBU (2025), News Integrity in AI Assistants: An International PSM Study, https://www.ebu.ch/files/live/sites/ebu/files/Publications/MIS/open/EBU-MIS-BBC_News_Integrity_in_AI_Assistants_Report_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

Fanous, Goldberg, Agarwal, Lin, Zhou, Daneshjou & Koyejo (2025), SycEval: Evaluating LLM Sycophancy, https://arxiv.org/abs/2502.08177

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

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

Pressens Faglige Utvalg (2021) Code of Ethics of the Norwegian Press, https://presse.no/pfu/etiske-regler/vaer-varsom-plakaten/vvpl-engelsk/

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

< Back to Contents

EBU
OPERATING EUROVISION AND EURORADIO

BBC

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.

EBU
OPERATING EUROVISION AND EURORADIO

BBC

News Integrity in AI Assistants

TOOLKIT

October 2025

EBU
OPERATING EUROVISION AND EURORADIO

BBC

News Integrity in AI Assistants

TOOLKIT

October 2025

EBU
OPERATING EUROVISION AND EURORADIO

BBC

News Integrity in AI Assistants

TOOLKIT

October 2025