OUTCOMES ROCKET

AI'S SOURCES OF TRUTH: HOW CHATBOTS CITE HEALTH INFORMATION

GROWTH ACCELERATED

Large language models are now a common source for health questions—from basic symptoms to complex treatment guidance—but where their medical knowledge comes from remains unclear.

Our study, ***"AI's Sources of Truth: How Chatbots Cite Health Information,"*** analyzed 5,472 citations generated in response to health-related prompts across four web-enabled models: ***ChatGPT*** (GPT-4o with browsing), **Google Gemini** (2.5 Flash), ***Claude*** (Sonnet 4), and ***Perplexity*** (Sonar mode).

By examining the websites these chatbots cite—along with source recency, content characteristics, and paywall status—the study provides a clearer picture of how reliable today's AI-driven healthcare information really is.

Top Websites Cited by LLMs

When answering health questions, LLMs pull from a wide range of sources. Even the most frequently cited domain—PubMed Central, with 385 citations—accounts for less than 0.1% of all references.

The data reveals a strong preference for a small set of trusted domains. ***PubMed Central*** tops the list, underscoring AI's reliance on peer-reviewed, open-access research. ***Cleveland Clinic*** and ***Mayo Clinic*** follow closely, along with the government-backed ***NCBI database***, reflecting a consistent tendency to favor authoritative, medically reviewed sources.

Beyond these "big four," citations span academic publishers (ScienceDirect, Nature, arXiv), consumer health media (Healthline, WebMD, Medical News Today, Verywell Health), and major public health authorities (CDC, NHS, WHO, Heart.org).

Notably, YouTube appears in the top 20 with 47 mentions—indicating that video explainers and expert-led talks with robust transcripts sometimes rise to the level of usable sources, despite the platform's generally lower reliability as user-generated content.

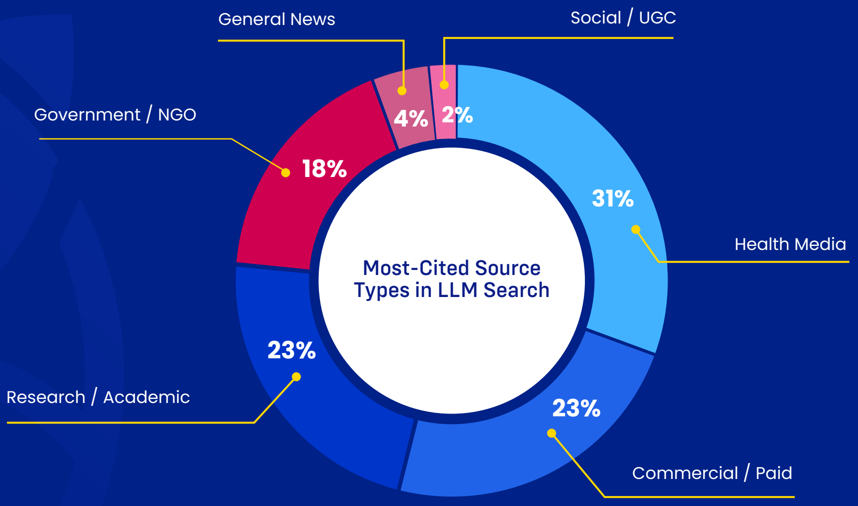
RANK	WEBSITE	TOTAL MENTIONS
1	pmc.ncbi.nlm.nih.gov	385
2	my.clevelandclinic.org	174
3	www.mayoclinic.org	163
4	www.ncbi.nlm.nih.gov	150

Nearly one-third of citations (30.7%) come from health media sources like Mayo Clinic, Cleveland Clinic, and Healthline. Commercial or affiliate sites follow at 23.1%, while academic and research sources make up 22.9%—showing that LLMs often lean toward accessible, consumer-facing content over technical literature.

General news (3.7%) and user-generated content (1.6%) appear rarely, indicating that journalism and anecdotal sources hold little influence in LLM responses.

Most-Cited Source Types in LLM Search

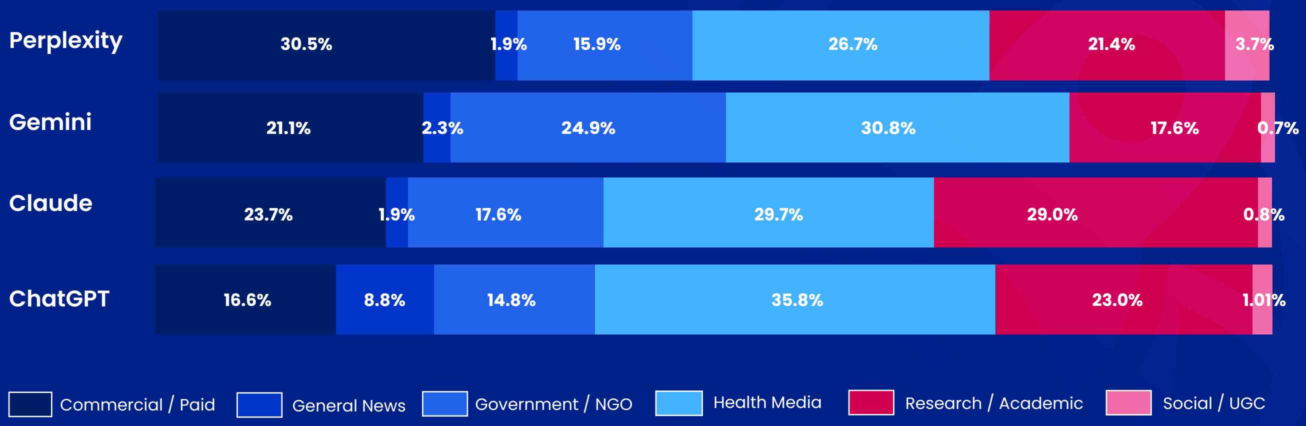
Revealing the Information Ecosystem Behind AI-Generated Health Advice



Model differences are clear. ChatGPT relies most on health media (35.8%) and cites academic research 23% of the time. Claude is more balanced, drawing similar shares from health media (29.7%) and academia (28.9%).

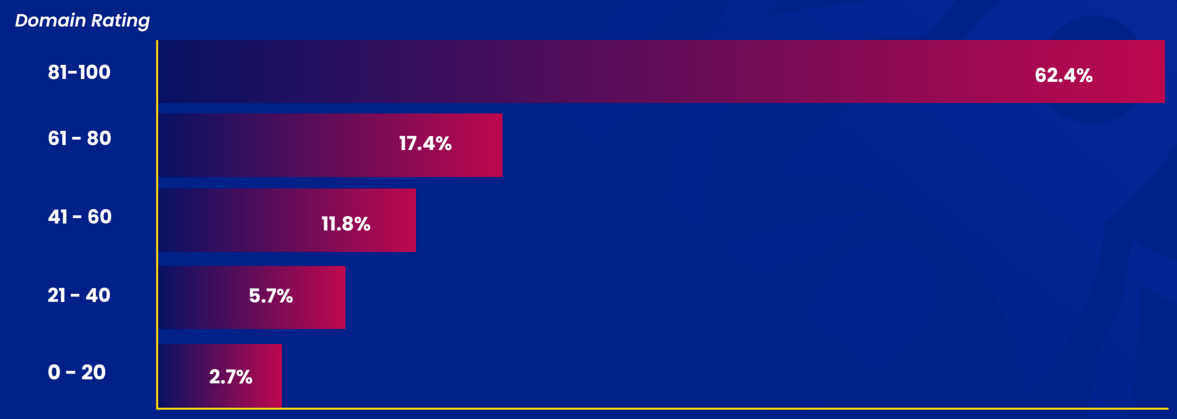
Gemini relies more on government and NGO sources than other models, while Perplexity leans most on commercial content and cites the most user-generated material.

Most-Cited Source Types by Chatbots



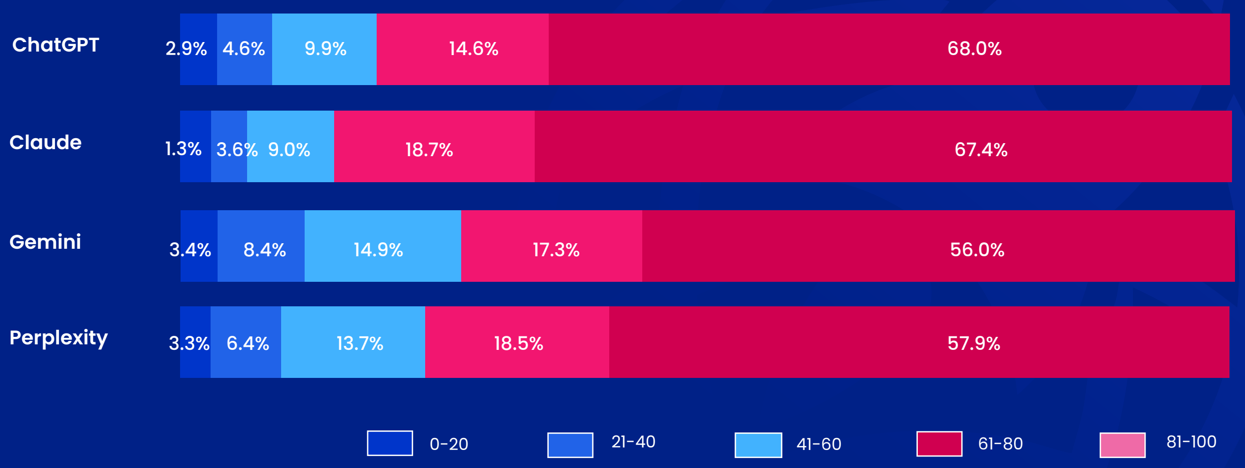
When it comes to credibility, Domain Rating (DR) scores provides a useful lens. The data reveals that 62.4% of all citations originate in domains with the highest authority ratings (DR scores 81-100). NIH, PubMed, and Mayo Clinic are some of the sites users tend to instinctively trust. By contrast, only 2.7% of citations are from the lowest tier of authority (DR 0-20), where reliability is most questioned.

Share of Citations by Domain Rating



Most chatbots rely heavily on high-authority sources, but they weight them differently. ChatGPT (68.0%) and Claude (67.4%) cite top-rated domains the most, while Gemini (56.0%) and Perplexity (57.9%) draw more from mid-tier sources. Perplexity also cites the most low-authority sites (3.3%), matching its greater use of commercial and user-generated content.

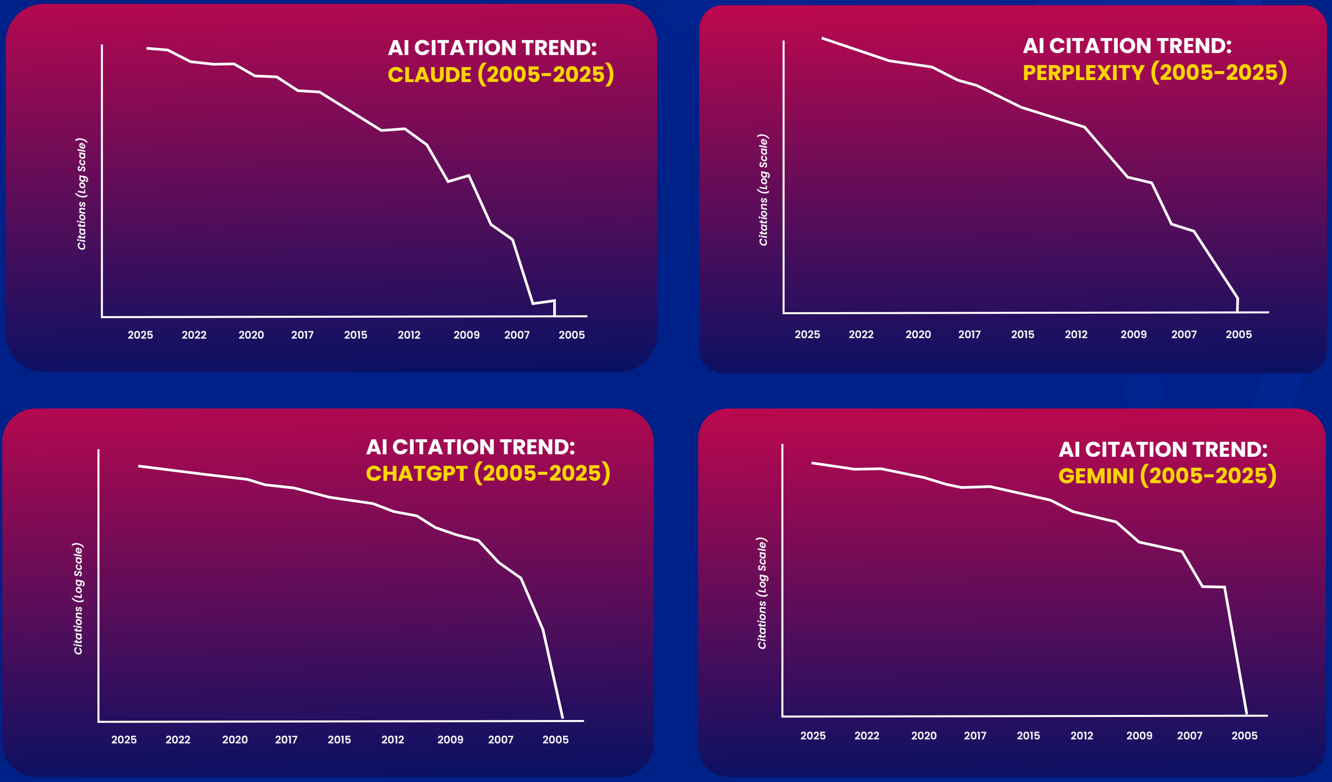
Share of Citationsby Domain Rating: Chatbot Breakdown



The Recency of Sources Referenced by LLMs

Analyzing the publication years of the 5,400+ citations reveals that chatbots predominantly reference recent research and articles, with almost two-thirds of citations dated either 2024 or 2025. **All four chatbots draw the most heavily from 2025 (40% of all citations), and decline precipitously after that.**

Main Reasons for Adopting an ABM Platform



The steep drop in citations from earlier years means older but still important research is referenced far less. While newer studies improve relevance, this pattern can unintentionally overlook long-standing evidence that still shapes medical standards and best practices. It raises an important question about balance: how often should models prioritize new findings over proven, foundational work?

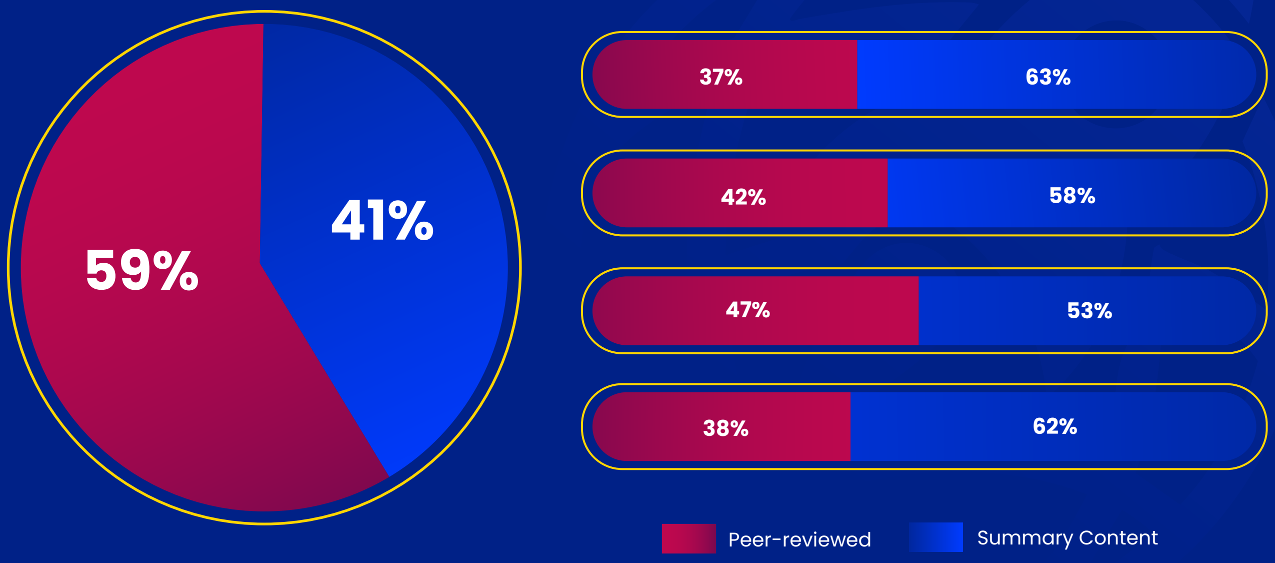
All four models follow this same trend, suggesting an industry-wide shift toward “present-tense” knowledge. This makes chatbot responses feel timely and aligned with current developments, but it also creates the risk of over-relying on short-term publications or early-stage data.

Ensuring that LLMs integrate both recency and rigor will be crucial for providing reliable, trustworthy health information moving forward.

Content Features

When it comes to the content of the health advice returned, chatbots lean more on interpreters of science than science itself. ***Across all four models, 59% of references include content drawn from summaries such as health media sites, explainers, and consumer guides compared to 41% from peer-reviewed research.*** In other words, it's more likely that LLMs will quote "what the science means" than the raw studies behind it.

Summaries vs. Original Research In LLM Citations



The preference is the most obvious in ChatGPT, which draws 62% of its citations from summaries and only 38% from research articles. Perplexity is almost the same, with 63% summaries and 37% peer-reviewed, showing a clear bias towards consumer-friendly explanations. Claude has the highest rate of peer-reviewed citations at 47%, while Gemini is positioned in a better balance, with 58% summaries and 42% research sources.

It is obvious why LLMs would favor summaries over studies, as it makes the answer more accessible to people's everyday lives. On the other hand, it could pose a threat to the original evidence if the cited interpretation is wrong and does not capture the full analogy, particularly in vital disciplines such as medicine, where nuance and accuracy are important.

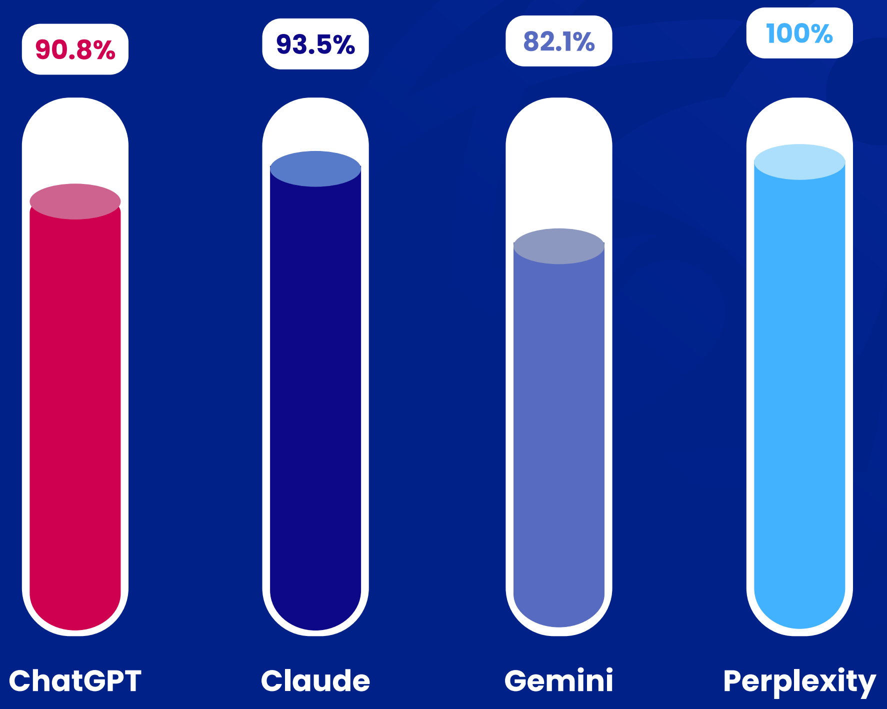
Overall, the data shows that LLMs rely more on ***summaries and secondary explanations*** than on original research, a trend clearly reflected across all four models in the graphic. While this makes health advice more accessible, it also means chatbot answers depend heavily on how others interpret the science—risking oversimplification or loss of nuance. Striking a balance between ***easy-to-understand content and scientific accuracy*** will be key to ensuring reliable AI-driven health information.

While the quality of the content returned is important, the number of supportive sources is also a matter of concern. On average, every answer a user gets to a ***health-related question is supported by 12 to 15 citations.***

Perplexity, which ranks last in terms of high-quality citations, ***ranks first in the volume of citations***, a***veraging 14.97 citations per query.*** Its approach represents a philosophy of abundance, drawing from a broad array of sources: commercial sites and even user-generated content are incorporated to give a more complete picture.

Coming second at 13.99 is Claude, which provides a more even balance of academic, government, and health media references. These are followed by ChatGPT with 13.59, and then Gemini with 12.29 citations per response.

On Average, LLMs Cite 13.7 Sources per Query



All four LLMs cite a high number of sources—***about 13 to 15 per health query***—showing that their answers draw from a broad mix rather than any single authority.

Where they differ is in how those sources are chosen. Perplexity cites the most and pulls from the widest range, including lower-tier domains. Claude and ChatGPT rely more on curated, higher-quality sources, while Gemini uses fewer citations but keeps a balanced mix. This shows that while quantity is similar, ***the type of sources*** each model favors plays a major role in the reliability of their health guidance.

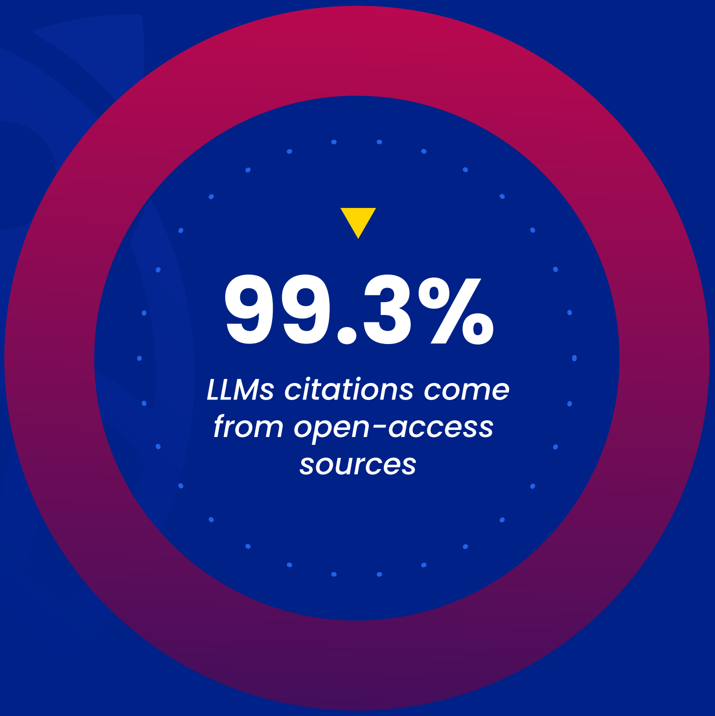
Content Accessibility

Of 5,400+ citations in our analysis, ***99.3% can be classified as "open access"***, meaning they were freely accessible. That makes LLMs highly "top of funnel" by focusing not on knowledge that's behind subscription walls, but on information anyone can click and read.

ChatGPT presents as the exception; it had 2.4% of its citations linking to paywalled material. While it's still a small share, it indicates that ***OpenAI is a little more willing to use gated research or premium media*** than its competitors. By contrast, Claude, Gemini and Perplexity almost never cite paywalled content, with less than 0.3%. Their outputs reflect a strong bias toward freely available sources, making responses more accessible, but may also limit depth when important research is found behind a paywall.

LLMs Overwhelmingly Avoid Paywalled and Gated Content

Most health information cited by LLMs comes from sources users can freely access, reflecting a strong bias toward open, verifiable content.



One of the greatest frustrations in online research is to click on a source, only to find oneself hitting a dead end. Luckily, chatbots appear to have largely solved that problem. ***Out of over 5,470 citations analyzed, only 0.2% were broken URLs.***

On all four models, the vast majority of the links worked out correctly, demonstrating that ***LLMs are surprisingly reliable*** in their ability to surface live, working sources. For users, that means fewer dead ends and more confidence that the evidence behind an AI's answer can actually be verified.

Conclusion

The data from this study paints a telling picture of how today's health chatbots construct their answers. They overwhelmingly prefer ***accessible sources that are recent and high-quality*** from an authorial perspective, and favor ***summaries*** over original research. Most of the answers have a lot of references, mostly from open sources, and almost never from broken links.

Each LLM shows distinct sourcing patterns. Perplexity leans most toward commercial and user-generated content, while Claude comes closest to balancing summaries and research. ChatGPT relies heavily on health media and is the most likely to cite paywalled studies. Gemini favors government and NGO sources, giving its answers a more policy-driven orientation.

How we interact with healthcare information is drastically changing due to AI chatbots. From a vast blackhole of URLs in Google Search, users are now able to pinpoint particular sources that are interesting and relevant to them.

From this study, we can be quite confident in the suggestions that AI models are producing; however, they are by no means immune to bias; thus, ***careful consideration and validation are always important and vital, especially with medical information.***

Methodology

This study ***analyzed 5,472*** unique citations generated by AI chatbots in response to health-related prompts.

The goal was to better understand not just how often these systems cite sources, but what kinds of sources they prioritize when providing medical and health information.

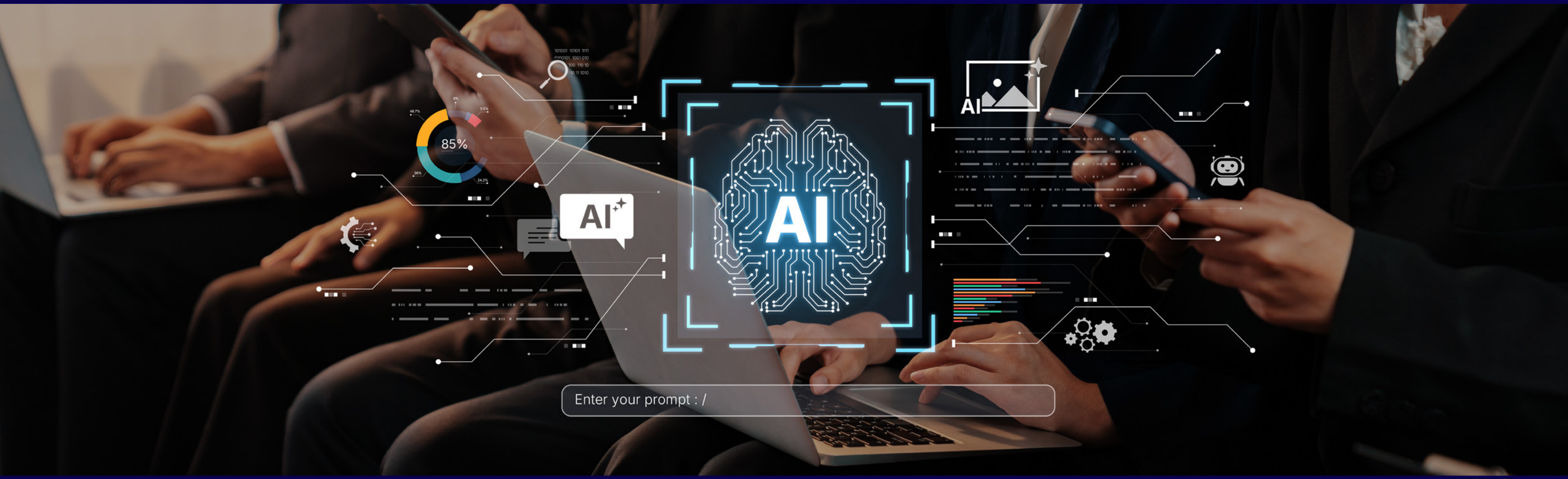
Data Collection

We built a prompt set designed to mimic real-world health queries, ranging from general wellness advice to more technical medical topics.

These prompts were run through four major web-enabled large language models during August 2025:

- ChatGPT (Web browsing mode, GPT-4o)
- Google Gemini (2.5 Flash)
- Claude (Sonnet 4)
- Perplexity (Sonar mode)

All links surfaced in chatbot responses were extracted and cleaned before classification. The final dataset comprised 1,497 citations from Perplexity, 1,217 from Gemini, 1,359 from ChatGPT, and 1,399 from Claude.



Source Categorization

Every citation was assigned to one of six groups to capture differences in authority, accessibility, and intent:

CATEGORY	DESCRIPTION
Research / Academic	Peer-reviewed journals, preprints, academic publishers, universities, scholarly databases.
Government / NGO	Official government health websites and nonprofit/global health organizations.
Health Media	Patient-facing resources from hospitals, medical centers, and health publishers.
General News	Mainstream media reporting on health issues and scientific developments.
Social / UGC	User-generated or anecdotal content (YouTube, Reddit, etc.).
Commercial / Paid	Corporate blogs, affiliate sites, e-commerce pages, or content with marketing intent.

Methodology

To evaluate rigor, citations were also tagged by evidence level:

- **Peer-reviewed research** – Primary studies, systematic reviews, and authoritative reports from government or NGOs.
- **Summary content** – Secondary or tertiary sources such as news articles, health guides, encyclopedias, or blogs that interpret or distill information.

Limitations

The findings reflect model behavior at a single moment in time. Because AI systems are regularly updated, citation patterns may change in future versions.

This analysis captures only the sources models choose to cite, not the full range of information they were trained on. Results are also based on English-language, publicly accessible responses, so behavior may vary across languages, regions, or platforms.

OUTCOMES
ROCKET

AI'S SOURCES OF TRUTH:
HOW CHATBOTS CITE HEALTH INFORMATION

GROWTH ACCELERATED

**SCAN THE QR CODE
TO VISIT OUR WEBSITE**

Discover how we help healthcare brands grow with smarter marketing and insights.