



competition commission
south africa



Media and Digital Platforms Market Inquiry (MDPMI) Final Report

ANNEXURE 5: GENERATIVE ARTIFICIAL INTELLIGENCE

NON-CONFIDENTIAL

PREPARED BY:

Donnavan Linley

Sivuyise Mtiki

Ntwanano Chauke

NOVEMBER 2025



a competitive, dynamic, deconcentrated and inclusive economy

TABLE OF CONTENTS

1.	INTRODUCTION.....	02
2.	GENERATIVE AI LANDSCAPE AND BUSINESS MODEL.....	03
2.1.	Generative ai business model.....	06
2.2.	Strategic partnerships between technology firms.....	09
2.3.	AI value chain and the development of llms.....	11
2.4.	The value of AI.....	14
2.5.	Licensing agreements between generative AI and large media houses....	15
3.	USE OF MEDIA CONTENT IN LLM TRAINING.....	18
4.	USE AND VALUE OF MEDIA CONTENT IN TRAINING DATASETS.....	19
5.	IMPACT OF AI POWERED CHATBOTS AND SEARCH ENGINES ON MEDIA.....	24
5.1.	AI chatbots powered with search.....	24
5.2.	Use of AI search for news and use of sources.....	27
5.3.	Potential to favour foreign contracted media by chatbots.....	31
6.	COMMENTS ON THE PROVISIONAL REPORT AND MDPMI VIEWS.....	32
6.1.	News engagement on AI chatbots.....	33
6.2.	Crawling of South African News content for training and opt-out.....	37
6.3.	Importance of news content on Generative AI Chatbots and licensing agreements.....	50
6.4.	Other issues raised by AI developers.....	55
7.	CONCLUSION.....	56

LIST OF APPENDICES	58
--------------------------	----

LIST OF FIGURES

Figure 1: Monthly unique visitors in South Africa, May 2025-September 2025.....	04
Figure 2: Generative AI website traffic from South Africa, January 2025-September 2025.....	05
Figure 3: Age distribution of South African AI chatbots user.....	05
Figure 4: Overview of the generative AI LLM value chain.....	11
Figure 5: Integration of search engine features within Gemini and Meta AI.....	26
Figure 6: South African user browsing habits on AI chatbots, January 2024 - October 2024....	28
Figure 7: News segment as a proportion of total web traffic in AI Chatbots.....	37

LIST OF TABLES

Table 1: List of generative AI developers and their respective models.....	03
Table 2: List of South African news publications that have blocked/ not blocked crawlers related to generative AI.....	21
Table 3: List of channels that Generative AI Chatbots can be accessed.....	35
Table 4: List of crawlers used by AI developers as of October 2025.....	44
Table 5: Dates AI developers launched their generative AI chatbots and introduced opt-out measures.....	47



MDPMI FINAL REPORT ANNEXURE 5

1. INTRODUCTION

1. Since the launch of OpenAI's ChatGPT in November 2022, there has been rapid technological advancements within the generative Artificial Intelligence ("AI") market segment, resulting in what is seemingly an AI arms race between AI developers to produce the most groundbreaking capabilities. Almost every month, generative AI developers are publicly announcing new features and capabilities of their AI models. For users, generative AI has revolutionised the functionalities of the internet, as within seconds generative AI chatbots can return a response to complex prompts from users, with an option to pose follow-up prompts without restating the context. This has made it easy for users to complete tasks such as conducting research, getting the latest news updates, writing and proofreading written text, coding, generating summaries on long written text, generating audio and visual content etc. Similarly for businesses and organisations, the integration of generative AI within internal firm systems, enables them to improve productivity in various areas such as customer care services, create content for marketing or communications, optimise internal processes, improve cyber security, organise and analyse internal data, conduct forecasting, draft internal reports etc.¹
2. Generative AI refers to the use of AI to create new content, that is powered by foundational models called large language models ("LLMs") that can multi-task, perform out-of-the-box as described above and be adapted for targeted use cases.² According to AI developers, generative AI is a significant shift, as it boosts creativity, productivity, and collaboration.³ What makes generative AI distinct from other types AI, is that it is focused on creating new content (i.e., similar to a human being) based on learned patterns and probabilities. This content can be delivered in multiple modalities, including text, images, videos, and 3-D representations (such as scenes and landscapes for video games).⁴
3. AI has become the new gold rush amongst venture capital (VC) with estimates of \$190bn worth of VC funding into generative AI investments, in the past five years, and investment in generative AI companies alone reaching \$29.4bn in 2023 growing to \$94.7bn in 2025.⁵ An AI race is currently on to see which firms emerge victorious with established tech giants such as Google, Apple, Meta and Microsoft all leveraging their tech capabilities from their respective market segments and investing heavily in AI.

1 Google Cloud. 2024. 321 real-world gen AI use cases from the world's leading organizations. Available online: <https://cloud.google.com/transform/101-real-world-generative-ai-use-cases-from-industry-leaders> [accessed 01 December 2024].

2 Google Cloud. 2024. Generate text, images, code, and more with Google Cloud AI. Available online: <https://cloud.google.com/use-cases/generative-ai?hl=en> [accessed 31 October 2024].

3 Google Workspace. 2024. Gemini for Google Workspace: A powerful and easy way of working with generative AI.

4 McKinsey Digital. 2023. Exploring opportunities in the generative AI value chain. Available online: <https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/exploring%20opportunities%20in%20the%20generative%20ai%20value%20chain/exploring-opportunities-in-the-generative-ai-value-chain.pdf> [accessed 19 December 2024].

5 Dealroom (2025). Generative AI. Available: <https://dealroom.co/guides/generative-ai#:~:text=During%20a%20rapid%20emergence%2C%20Generative,AI> [Accessed 4 November 2025].

4.

Though the emergence of generative AI has been widely welcomed, for the news media industry there are various concerns that have been raised about its potential impact on the sustainability of the industry.⁶ These concerns stem from the unauthorised use of high quality news media content amongst other sources in the training processes of LLMs, and the evolution in the capabilities of generative AI chatbots that have been enhanced now to respond to user queries with real time information and return topical and authoritative information such as news content. This in turn raises
- questions as to the value of news content for training Generative AI LLMs and the potential threat to the news media of further disintermediation using their content by AI Chatbots in addition to search and social media.
5.

In this Annexure Report, the original Provisional Report (with updated information) is presented in sections 1 - 5. In section 6, the Inquiry presents the views of stakeholders in response to the Provisional Report and then provides the Inquiry's own views by incorporating assessed evidence on hand.

2. GENERATIVE AI LANDSCAPE AND BUSINESS MODEL

6.

The generative AI market landscape in South Africa mirrors global trends, as it is made up of the same well-known global players that are currently at the forefront globally in developing generative AI. The relevant players and their respective models are
- listed in the table below, and it includes some of the leading global technology firms. The capabilities and features of these models is available for South African users and organisations to use and/ or deploy, in accordance to their needs.

Table 1: List of generative AI developers and their respective models

Generative AI Developer	Model
Google	Gemini
Meta	Meta AI
Microsoft	Copilot
OpenAI	ChatGPT
Perplexity	Perplexity
xAI	Grok
DeepSeek	DeepSeek -R1

Source: Market Inquiry's own compilation

7.

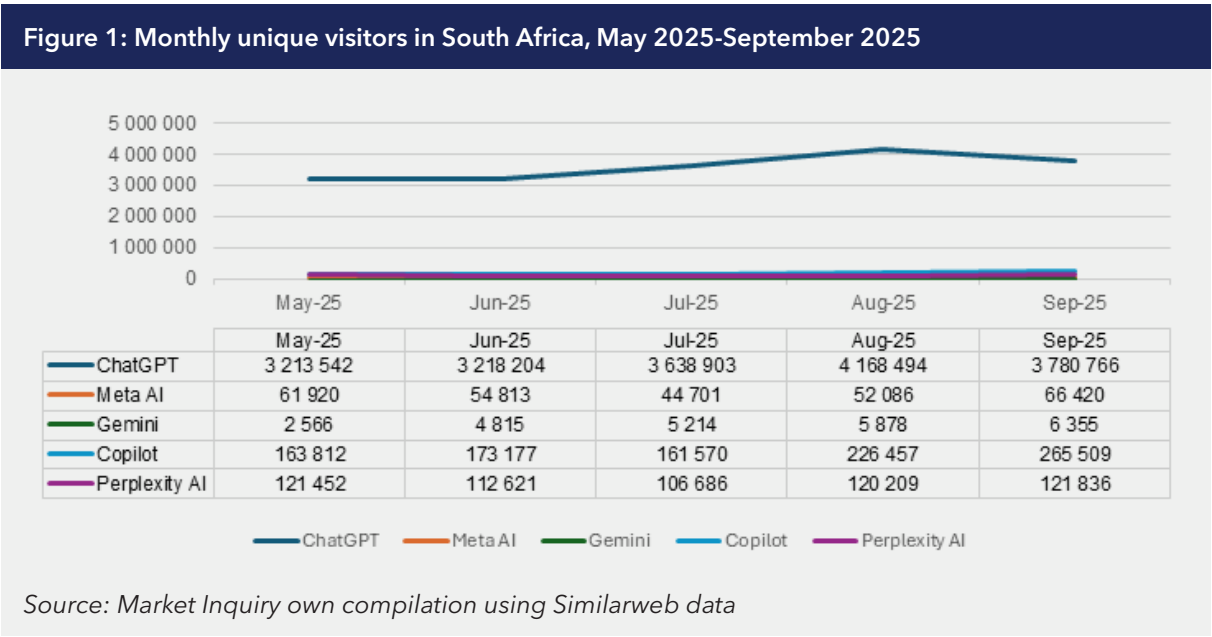
Data submitted by generative AI developers shows that there has been a significant uptake of AI chatbots from the inception stages of generative AI chatbots by South Africans,
- and the adoption trends have grown since with millions of South Africans actively using the various AI chatbots. By November 2023, Open AI already had [3] million monthly

6 Sherman & Rizzo. 2023. AI poses new threats to newsrooms, and they're taking action. Available at: <https://www.cnn.com/2023/06/06/news-organizations-ai-disinformation.html> (Accessed on 06 June 2023).

users while Bard (i.e., now Gemini) had [3X] daily active users.⁷ By April 2024 Copilot had already generated [3X] million active users with [3X] million prompts, while Meta AI generated [3X] million active users with over [3X] million prompts in the period April 2024 to June 2024 period.⁸

8. For a like-for-like comparison, Similarweb data shows the number of users using

generative AI chatbots are growing in South Africa, with data (i.e., as of September 2025) shown in the figure below, illustrating that ChatGPT is by far the most preferred chatbot for web-based visits.⁹ ChatGPT has over 3.7 million unique users in South Africa as of September 2025, followed by Perplexity AI and Copilot who have both generated less than 500 thousand users¹⁰.

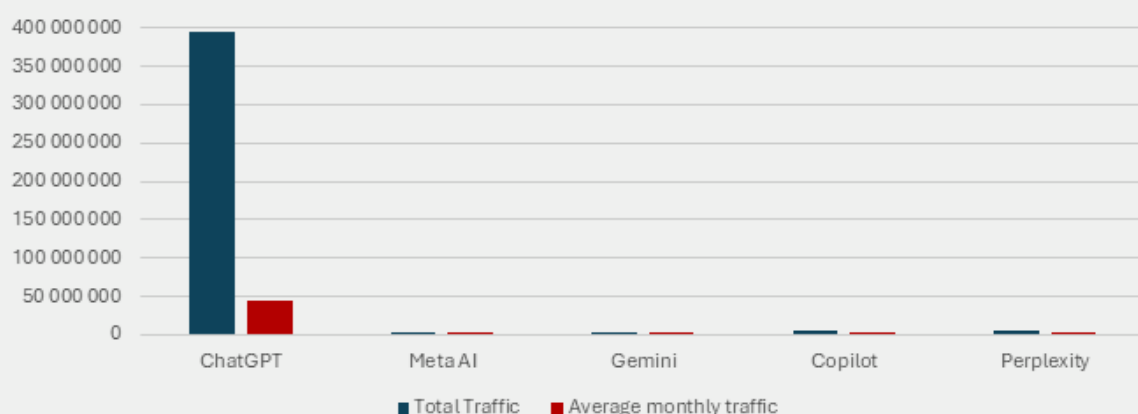


9. ChatGPT's popularity in South Africa is also reflected in the website traffic data. In the period of January 2024 to October 2024, ChatGPT generated a total of 137 million visits which accounts for 84% of the 163 338 302 overall total web-based visits that were generated by the AI chatbots in

South Africa¹¹. As shown by the figure below, the average monthly traffic (43.79 million) that ChatGPT generates in South Africa, is higher than the total traffic that has been generated by each of the other chatbots in the period of January 2025-September 2025.

7 Open AI submission by Chris Charter dated 06 December 2023. Written response to 1st RFI. Appendix B.
Google submission by Palesa Mpe dated 08 December 2023. Written response to 1st RFI tranche 2. Paragraph 28.
8 Meta submission by Claire Reidy dated 19 June 2024. Written response to 3rd RFI. Paragraph 9.
Microsoft submission by Xolani Nyali dated 13 June 2024. Written response to 3rd RFI. Paragraph 26.3
9 Using Similarweb data, ChatGPT accounted for 72.8% of web traffic in the period July to September 2025 in South Africa. Gemini accounted for 6%, openai.com (5.3%), Deepseek (3.4%), Grok (3.3%). Note this excludes app-based usage where, for example, MetaAI would feature more prominently.
10 A limitation of the Similarweb data is that it captures web-based usage of genAI chatbots and not app-based usage. For example, Meta AI which is found in Meta's suite of platforms (Facebook, Instagram, Messenger and WhatsApp) would have a significantly higher number of unique users if one included in-app usage.
11 In the period January 2025 to September 2025, overall visits to ChatGPT grew more than 3-fold to 428m visits accounting for 79% of web-based generative AI web traffic in September 2025 (using Similarweb data).

Figure 2: Generative AI website traffic from South Africa, January 2025-September 2025

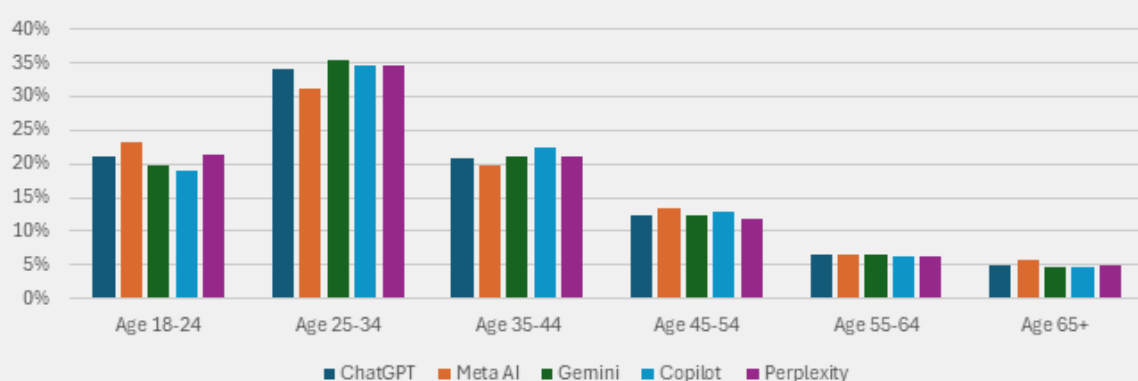


Source: Market Inquiry own compilation using Similarweb data

10. In relation to the average time users spend per visit on each chatbot, Similarweb data shows that most users have been spending more time on ChatGPT with an average of 08:51 mins, followed by Meta AI (06:00), Perplexity (04:57), Gemini (00:50) and Copilot (04:56). Regarding AI chatbots accessed via the web, 53% were accessed via desktop and 47% were accessed via mobile¹².
11. On demographics, there has been adoption across the various age groups. As shown by

the figure below, the largest portion of South African users (31%-36%) are within the age group of 25-34 years, followed by the 18-24 year age group and the 35-44 year age group with less users from the older +45 year age group. Similarly, the future use of generative AI chatbots will probably have more frequent use (70-73%) by the 18-50 year age group and less frequent use by the 51+ age group (46%) as per the Inquiry's Consumer Survey¹³.

Figure 3: Age distribution of South African AI chatbots users



Source: Market Inquiry own compilation of Similarweb data

12. In terms of gender parity, the distribution of users is close to being equally balanced

with a slight majority of users being females (51.83%).

¹² Similarweb data.

¹³ Inquiry Consumer Survey. Page 114-115.

2.1. GENERATIVE AI BUSINESS MODEL

13. The Generative AI business model is to develop foundational LLMs that can be adapted to a wide range of tasks and these are commonly released in various ways such as directly embedding the model into consumer facing applications (i.e., Chatbots), API integration, custom deployment and open-source release.
14. Chatbots are consumer facing and they are powered by foundational LLMs that have been fine-tuned or tailored to execute natural language processing tasks as such they are interactive, conversational and user-friendly, enabling users to interact with their functionalities to complete various tasks (i.e., coding, summaries, writing emails) as well as to search for information. There are mainly two access points for users to access Generative AI chatbots. There is the web browser option that users can use on their desktops or mobile devices by visiting the website domains of the Chatbots. There is also the applications option, which is for mobile devices. Users can also access generative AI through integrated means, where the Chatbot is integrated to other preexisting digital products for example on WhatsApp or Facebook.
15. When it comes to Chatbots, the business model has been to offer a free basic version of the chatbot to enable user trial and then sell enhanced versions which have more features and capabilities.
 - 15.1. Open AI has the free plan, as well as the premium plans (i.e., Go R149 per month, Plus R399 per month, Pro R3 999.99 per month, Business team R425 per month and Enterprise) which have unlimited model capabilities, as well as advanced features such as faster response rates, uncapped image generation, browse, advanced voice, memory, data analysis etc.¹⁴
 - 15.2. Google initially offered its premium chatbot under “Gemini Advanced”, which is bundled with other Google offerings under the Google One plan (R430 per month), offering Gemini Advanced, Gemini in Gmail and other applications, including 2TB of storage etc.¹⁵ But now has made new packages, Google AI Plus (R104.99 per month) Google AI Pro (R429.99) and Google AI Ultra (R2 699.99) with various offerings.¹⁶
 - 15.3. Perplexity offers Perplexity Pro at \$20 per month, which has more advanced capabilities and features, such as more searches a day, advanced reasoning and natural language processing capabilities, file uploads and analysis etc. In contrast to these business model, x.AI’s Grok is bundled with the X Premium (R200) and Premium+ (R400) subscription packages without any free basic version that users can access for free.¹⁷
 - 15.4. Meta AI currently does not have a premium product, but it has been reported that the firm is currently working on one.¹⁸

1.

-
- 14 OpenAI. 2024. Pricing. Available online: <https://openai.com/chatgpt/pricing/> [Accessed 05 November 2025]
 - 15 Google One. 2025. Secure storage and AI from Google. Available online: <https://one.google.com/about/ai-premium/> [accessed 24 January 2025].
 - 16 Google One. 2025. The best of Google AI in one membership. Available online: <https://gemini.google/subscriptions/>
 - 17 xAI submission by Derushka Chetty dated 15 January 2024. Written response to 2nd RFI. Paragraph 2.3.2.
 - 18 Meta AI. 2024. Expand your world with Meta AI. Available online: <https://ai.meta.com/meta-ai/> [Accessed 24 January 2025].

16. Aside from individual customers, the business model of Generative AI developers also includes paid plans as well as custom licensing contractual arrangements with third-parties. Through these arrangements, third-party organisations can access the LLMs directly through an API process or through custom deployment of the models where the model is fine-tuned for a specific use case. In return, Generative AI developers source revenue through a subscription plan and usage basis model where third-parties “pay as they go”.

16.1. The Generative AI developers have created various product offerings and avenues for third-parties to access their models. For instance Google has Gemini for Google Workspace, that is built into its Workspace products (i.e., Gmail, Docs, Sheets etc) with specialised plans (i.e., Gemini Business \$20 per user/month and Gemini Enterprise \$30 USD per user/month) and it enables institutions to collaborate with Gemini with workplace tasks such as writing, creating images, automate insights, optimise project work flows etc.¹⁹ Through its Google AI Studio platform, Google allows developers to experiment and deploy Gemini models into their applications.²⁰ On its Vertex AI platform, Google provides developers with the relevant infrastructure to customise and integrate Gemini models and others models into third-

party applications at scale.²¹ On both platforms, Google generates revenue based on the usage.²²

16.2. Open AI offers the ChatGPT Enterprise plan through a custom subscription offering model customisation options, model fine-tuning and API integration amongst other features.²³ OpenAI through OpenAI API also releases its AI models for third parties to integrate them into their applications, primarily through API access, but also provides them with avenues to fine tune the models for specific tasks.²⁴ OpenAI also utilises the usage based model to generate revenue from third parties that make use of their AI models.²⁵ Through GPT Store, OpenAI allows users to discover and create custom ChatGPT versions of their own choice for users that subscribe to ChatGPT Plus, ChatGPT Team and ChatGPT Enterprise.²⁶

16.3. Though it is still early stages, the recently announced custom partnership between OpenAI and Apple is an example of how generative AI developers are releasing their models to third-party firms, for large scale use as through the partnership Apple is integrating OpenAI’s model across various applications within its portfolio of devices globally.²⁷

19 Google Workspace. 2024. Gemini for Google Workspace: A new and powerful way of working.

20 Google Cloud. 2024. The Gemini era for developers and businesses. Available online: <https://cloud.google.com/ai/gemini?hl=en-ZA> [Accessed 09 December 2024].

21 Google Cloud. 2024. The Gemini era for developers and businesses. Available online: <https://cloud.google.com/ai/gemini?hl=en-ZA> [Accessed 09 December 2024].

22 Google AI for Developers. 2024. Pricing. Available online: https://ai.google.dev/pricing#1_5flash [Accessed 09 December 2024].

23 OpenAI. 2023. Introducing ChatGPT Enterprise. Available online: <https://openai.com/index/introducing-chatgpt-enterprise/> [Accessed 10 December 2024].

24 OpenAI. 2020. OpenAI API. Available online: <https://openai.com/index/openai-api/> [accessed 09 December 2024].

25 OpenAI. 2024. Pricing. Available online: <https://openai.com/chatgpt/pricing/> [accessed 06 December 2024]

26 OpenAI. 2024. Introducing the GPT Store. Available online: <https://openai.com/index/introducing-the-gpt-store/> [accessed 10 December 2024].

27 OpenAI. 2024. OpenAI and Apple announce partnership to integrate ChatGPT into Apple experiences. Available online: <https://openai.com/index/openai-and-apple-announce-partnership/> [accessed 12 December 2024].

17. One of the common trends by Generative AI developers is the integration of their generative AI models into their own commonly used digital services and applications which enhances the capabilities of these services, improves user experience and promotes user adoption.
 - 17.1. Google has deployed the Gemini model on Google Search for the "AI Overviews" features on its search results page which enables the search engine to return a response to more complex user prompts.²⁸ Google has also deployed the Gemini model in other products within its portfolio such as Google Workspace (Gmail, Docs, and Sheets). Similarly, Microsoft has also integrated Copilot model which is powered by OpenAI, into its search engine Bing for the "Browse with Bing AI" feature (i.e., now called Copilot Search in Bing), as well as in several other Windows applications such as Microsoft 365.²⁹
 - 17.2. Meta AI has integrated its model within its portfolio of social media platforms digital services (i.e., Facebook, Instagram, WhatsApp and Messenger) where users can easily access and engage with the Chatbot.³⁰ Grok is also integrated into X, the social media platform.
18. Similar to other forms of technology products or services (i.e., computer or mobile operating systems like Linux and Android), some of the generative AI developers also use the open source licensing model, to release their LLMs to the public. For instance, Meta AI's Llama 3 and Google's Gemma model, are both released through an open source licensing model which grants users and/ or developers with free access to the technical features of the model such as the weights it was trained on, less restrictions and access to customise and modify the model to meet their specific use case.³¹
19. AI LLMs are therefore typically monetised in other ways, other than the sale of advertising using user data, unlike search and social media. However, they are selling knowledge and capabilities, and to the extent they are integrated into search and social media, may contribute to the traditional means of funding those businesses.
20. It is no coincidence that the largest tech companies are part of the leadership in GenAI, as building and operating models requires considerable supercomputing power which these tech companies have to power their existing businesses, and training data which is available from the open web and their existing products. Even OpenAI is in a close partnership with Microsoft which has an investment in OpenAI's profit-orientated entity and provides it supercomputing

28 Google. 2024. Generative AI in Search: Let Google do the searching for you. Available online: <https://blog.google/products/search/generative-ai-google-search-may-2024/> [accessed 04 December 2024].

29 Microsoft. 2023. Copilot and Bing's AI features. Available at: <https://www.microsoft.com/en-us/bing/do-more-with-ai/bing-ai-features?form=MA13KP> (Accessed on 29 September 2023).

Microsoft. 2023. Introducing Microsoft 365 Copilot – your copilot for work. Available online: <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/> [accessed 04 December 2024].

30 Dang. 2024. Meta integrates Google & Bing search results into AI assistant. Available at: <https://megadigital.ai/en/blog/meta-integrates-google-bing-search-results-into-ai-assistant/> (Accessed on 22 April 2024).

Meta. 2024. Meet Your New Assistant: Meta AI, Built With Llama 3. Available online: <https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/> [accessed 26 November 2024].

31 Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. Available online: <https://ai.meta.com/blog/meta-llama-3/> [accessed 12 December 2024].

Google Open Source. 2024. Building Open Models Responsibly in the Gemini Era. Available online: <https://opensource.googleblog.com/2024/02/building-open-models-responsibly-gemini-era.html> [accessed 12 December 2024].

power in exchange for model use, with some competition authorities investigating whether this amounts to some form of control and a notifiable merger.³² The global tech companies also command most attention on

the web, and so distribution through these companies is an important use case for Generative AI. Hence the OpenAI deal with Apple.

2.2. Strategic partnerships between technology firms

21. Strategic partnerships between some of the large global technology firms has become one of the noticeable features within the generative AI space. To date, the Open AI and Microsoft partnership is one of the partnerships that stands out.

21.1. The partnership has resulted in over \$13 billion worth of investment in the development, expansion and commercialisation of OpenAI technologies since 2019.³³

21.2. The partnership entails joint development of specialised supercomputing technologies to train and operate AI models, and it includes the use of Microsoft Azure cloud computing to exclusively run OpenAI's services, as well as the deployment of OpenAI models on Microsoft consumer and business products, such as Copilot which

is also integrated within Microsoft ecosystem.³⁴

22. OpenAI has also established a partnership with Apple and Reddit.

22.1. With the Apple partnership, the two firms will be integrating ChatGPT into Apple's range of device operating systems (i.e., iOS, iPadOS and macOS), allowing users to access ChatGPT's capabilities.³⁵ The partnership enhances the user experiences for Apple device owners in tasks such as writing, image creation, proofreading etc. The partnership also includes ChatGPT powering Apple's virtual assistant Siri, and augmenting Siri's capabilities to respond to user prompts.³⁶

22.2. The OpenAI/ Reddit partnership grants OpenAI with access to Reddit's Data API and brings Reddit

32 Microsoft. 2023. Microsoft and OpenAI extend partnership. Available online: <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/> [accessed 13 December 2024].

Financial Times. 2023. How Microsoft's multibillion-dollar alliance with OpenAI really works. Available online: <https://www.ft.com/content/458b162d-c97a-4464-8afc-72d65afb28ed> [accessed 13 December 2024].

33 Microsoft. 2019. Press release: OpenAI forms exclusive computing partnership with Microsoft to build new Azure AI supercomputing technologies. Available online: <https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/> [Accessed 21 March 2024].

Microsoft. 2023. Microsoft and OpenAI extend partnership. Available online: <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/> [accessed 21 March 2024]

34 Microsoft. 2023. Microsoft and OpenAI extend partnership. Available online: <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/> [accessed 21 March 2024]

Microsoft. 2023. Introducing Microsoft 365 Copilot - your copilot for work. Available online: <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/> [accessed 03 December 2024].

35 Open AI. 2024. Open AI and Apple announce partnership to integrate ChatGPT into Apple experiences. Available at: <https://openai.com/index/openai-and-apple-announce-partnership/> (Accessed on: 10 June 2024)

36 Aljazeera. 2024. Apple unveils long-awaited AI strategy partnership with OpenAI at WWDC. Available at: <https://www.aljazeera.com/news/2024/6/11/apple-unveils-long-awaited-ai-strategy-partnership-with-openai> (Accessed on 11 June 2024).



- content to ChatGPT users to access and engage.³⁷ On Reddit's end, the partnership will grant it access to build news products on Open AI's models as well as develop features that are powered by the AI models.
23. Other generative AI developers have also entered into partnerships with other generative AI developers and/ or technology firms.
- 23.1. Meta has a collaborative partnership with Microsoft to further develop AI, scale it and make it accessible.³⁸ The firms have partnered on the development of Meta's open source LLM called Llama, which is available on Microsoft's Azure platform and locally on Windows.
- 23.2. Meta AI also has partnerships with both Microsoft Bing and Google that enables the chatbot to enhance its responses to users prompts.³⁹
- 23.3. OpenAI announced the development of the Stargate AI infrastructure project which will be a partnership between OpenAI, Softbank, Oracle, ARM, MGX and Microsoft, where Softbank will assume financial responsibility and OpenAI assume operational responsibility of the project⁴⁰.
- 23.4. There has also been unconfirmed reports that Google has recently entered into a \$2.7 billion licensing agreement with Character AI, an arrangement that includes technical access to Character AI's technology as well as the technical expertise of its founders.⁴¹
24. While the partnerships between the various tech companies delineated above are likely to generate synergies in the production of generative AI, they are also likely to raise serious competition issues including their entrenchment in the provision of generative AI services where the pre-existing digital barons can leverage their current dominant positions in digital markets into the newly forming AI and generative AI markets. Other competition issues may include the sharing of commercially sensitive information, abuse of dominant positions, price and other forms of anticompetitive discrimination, access to essential generative AI facilities etc. While these issues fall outside the scope of this Inquiry, they will play a significant role in shaping online and to a large extent offline markets, as well as their corresponding competitiveness. Therefore, it remains important for antitrust authorities to consider these relationships. The Inquiry will focus its attention on news media and content in relation to generative AI.

37 Reddit is an online community network, that allows users to post (i.e., text, links, images, and videos) vote, and comment in communities organized around their interests.

OpenAI. 2024. OpenAI and Reddit Partnership. Available online: <https://openai.com/index/openai-and-reddit-partnership/> [accessed 30 November 2024].

38 Meta. 2023 Meta and Microsoft Introduce the Next Generation of Llama. Available online: <https://ai.meta.com/blog/llama-2/> [accessed 01 December 2024].

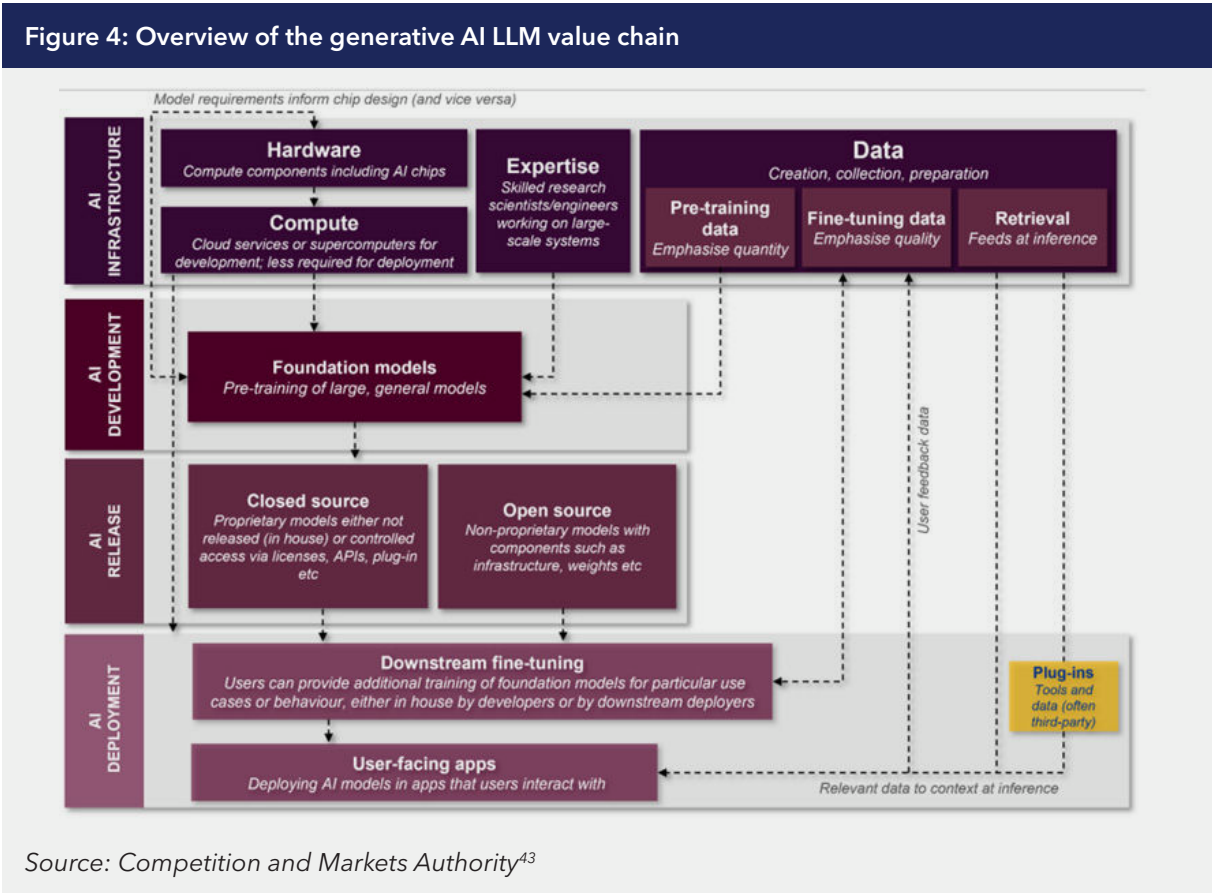
39 The Verge. 2024. Q&A: Mark Zuckerberg on winning the AI race. Available online: <https://www.theverge.com/2024/4/18/24134370/mark-zuckerberg-meta-interview-llama-3-ai-assistant-race> [accessed 30 November 2024].

40 Open AI (2025). Announcing the Stargate project. Available: <https://openai.com/index/announcing-the-stargate-project/> (accessed 22 January 2025).

41 The Wall Street Journal. 2024. Google Paid \$2.7 Billion to Bring Back an AI Genius Who Quit in Frustration. Available online: https://www.wsj.com/tech/ai/noam-shazeer-google-ai-deal-d3605697?mod=livecoverage_web [Accessed: 13 January 2025].

2.3. AI value chain and the development of LLMs

25. The development of foundational LLMs is a complex process that includes various interdependent stages. The process includes various key inputs and resources that the model depends on before it can be developed and be ready for deployment. Accordingly, the Market Inquiry considers it important to provide a brief overview of the value chain of generative AI models, in order to fully comprehend the developmental
- process that LLM models are subjected to before they are deployed for different use cases.
26. The key components of developing and deploying generative AI models is illustrated in the figure below and it includes various key inputs namely, technical expertise, the relevant computing infrastructure, as well as data.⁴²



27. Computing infrastructure includes bespoke high-end hardware and software that has the relevant power and capabilities to support
- the development and deployment process of LLMs and it includes components such as high-performance graphics processing units,

42 European Commission. 2024. Policy Brief: Competition in Generative AI and Virtual Worlds. Page 4. Available online: https://competition-policy.ec.europa.eu/document/download/c86d461f-062e-4dde-a662-15228d6ca385_en [accessed 19 December 2024].

43 Competition and Markets Authority. 2024. AI Foundation Models: Technical update report. Available online: https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf [accessed 18 December 2024].

- AI accelerator chips, high data storage (e.g. data centres, cloud storage) and processing capabilities (e.g. supercomputing), high bandwidth, low latency networking,⁴⁴ large amounts of power and infrastructure to support that power demand.⁴⁵
28. At the development stage of the value chain, generative AI LLMs require broad data (i.e., at scale), and this is for purposes of training the models so that they can sufficiently execute a wide range of tasks. The data includes text, images, music, videos as well as any other type of data that is deemed relevant.⁴⁶ This data is generally gathered from different sources such as public information that is on the internet, licensing arrangements, synthetic processes and it is converted into a usable format for extensive training, such as 'tokenising' text or breaking down images into a series of image region features.⁴⁷ For instance, as part of the data gathering process, some of the generative AI developers have either developed or used their pre-existing web crawlers to collect and store internet content (i.e., books, news articles, scientific journals, and user generated content on social media sites, blogs, and forums etc.) from various websites and combine with other data such as proprietary data to train their models.⁴⁸
 29. Once the foundational model has been developed, the model can either be released (i.e., through open source or closed source) for deployment or further developed through a process called fine-tuning, where the model is further trained and adapted to a targeted task.⁴⁹ The deployment of foundational models takes different forms such as powering consumer oriented applications and services like Chatbots (i.e. ChatGPT, Gemini, Meta AI) or integrated into existing applications by third-party institutions to enhance internal productivity.
 30. Many of the large global technology firms that have developed generative AI models are operating at multiple levels of the generative AI value chain. These firms have leveraged on some of their pre-existing operations, and they have also strategically expanded their activities across the value chain particularly at the upstream level. Market players such as Microsoft and OpenAI have collaborated extensively in the development process of OpenAI's models with Microsoft providing the relevant infrastructure for training and deploying the models, while OpenAI focuses on the actual research and development aspects of the process.⁵⁰
 - 30.1. At the infrastructure level, NVIDIA is considered to be the leading player in the production and supply of the

44 IBM. 2024. What is AI infrastructure? Available online: <https://www.ibm.com/think/topics/ai-infrastructure> [accessed 19 December 2024].

45 Duffy (2025). Trump announces a \$500 billion AI infrastructure investment in the US. Available: <https://edition.cnn.com/2025/01/21/tech/openai-oracle-softbank-trump-ai-investment/index.html> (accessed 22 January 2025).

46 Microsoft 2024. Written response to the European Commission's Call for Contributions on Competition in Generative AI dated 9 January 2024.

47 Competition and Markets Authority. 2024. AI Foundation Models: Technical update report. Paragraph 2.23. Available online: https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf [accessed 18 December 2024]. Commonwealth of Australia. 2024. Examination of technology: Multimodal Foundation Models. *Working paper*. Available online: <https://dp-reg.gov.au/sites/default/files/documents/2024-09/DP-REG%20-%20MFMs%20-%20Examination%20of%20Technology%20working%20paper.pdf> [accessed 19 December 2024].

48 [3<] Autoridade da Concorrência. 2023. Competition and Generative Artificial Intelligence. Issues paper. Available online: <https://www.concorrenca.pt/sites/default/files/documentos/Issues%20Paper%20-%20Competition%20and%20Generative%20Artificial%20Intelligence.pdf> [accessed 19 December 2024].

49 Microsoft 2024. Written response to the European Commission's Call for Contributions on Competition in Generative AI dated 9 January 2024.

50 Microsoft 2023. Microsoft and OpenAI extend partnership. Available online: <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/> [accessed 22 January 2025].

- key semiconductors globally (i.e., accounting for 70% of the global market), however firms such as Google, Microsoft, Amazon have also developed their own custom AI-optimised chips in order to self-supply some of their developments.⁵¹ Microsoft through its partnership with OpenAI has invested significantly in the development of specialised supercomputing systems and AI infrastructure to support and accelerate the development of OpenAI's models.⁵²
- 30.2. In terms of training data, firms such as Google and Microsoft have pre-existing data collection capabilities (i.e., Automated Bots) from their search engine operations that they can or have leveraged on to collect data from the internet at large scale and complement it with other sources of data.⁵³ Microsoft argues that for data related to the training and development of AI models, Google controls one of the largest proprietary data sets through the Google Search Index (estimated at between 500 and 600 billion pages) as well through YouTube (estimated at 14 billion publicly visible videos).⁵⁴ In the Inquiry's Public Hearings, Google did not deny the use of its search index data in the training of its LLMs when quizzed about its use.⁵⁵ Similarly, Microsoft also maintains a relatively scaled web index, containing 100-200 billion web pages that it can possibly turn to for training and development of AI models with OpenAI.⁵⁶ Both Meta AI and OpenAI have also developed web crawling and indexing capabilities.⁵⁷ Meta has confirmed that some of the posts on its Instagram and Facebook social media platforms formed part of the data that was used to train its models.⁵⁸
- 30.3. Many of the large technology firms also have presence at the release and deployment level of the value chain, where the models are optimised, scaled and made accessible. At this level, market players such as Amazon, Microsoft and Google have developed cloud computing platforms (i.e., Microsoft Azure, Amazon Web Services, Google Cloud Platform) where they provide a platform with the relevant computing services (i.e., online services of servers, storage, databases, networking, software, analytics, and intelligence) to deploy their own models as well as those of third-parties.⁵⁹ Through these platforms, the generative AI developers grant third party developers access to their AI infrastructure which enables them to experiment with generative AI models as well as develop their own models.⁶⁰ In addition, the cloud computing platform also grants the

51 Microsoft 2024. Written response to the European Commission's Call for Contributions on Competition in Generative AI dated 9 January 2024.

52 Microsoft. 2023. Microsoft and OpenAI extend partnership. Available online: <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/> [accessed 07 January 2025].

53 [36]

54 Microsoft 2024. Written response to the European Commission's Call for Contributions on Competition in Generative AI dated 9 January 2024.

55 Google Public Hearings- oral submission from Erin Simon dated 12 March 2024, page 100-101.

56 Competition and Markets Authority. 2020. Online platforms and digital advertising. Market Study final report. Paragraph 3.55.

57 [36]

58 Meta. 2023. Privacy Matters: Meta's Generative AI Features. Available online: <https://about.fb.com/news/2023/09/privacy-matters-metas-generative-ai-features/> [accessed 10 January 2025].

59 Microsoft. 2025. Cloud Computing defined. Available online: <https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-cloud-computing/> [accessed 10 January 2025].

60 Amazon. 2024. Amazon's response to the European Commission's call for contributions on Competition in Generative AI

public access to AI developers' own LLMs as well as to thousands of other third-party LLMs.

- 30.4. The generative AI value chain is considered to be evolving rapidly, as such it is prone to change over time. The structure of the value chain is also not static as it may be distinct in accordance to each type of model because the foundational model type dictates the computing inputs required as well as the processes that must be undertaken. The recent launch of Deepseek's R1 LLM is an example of this, where the apparent China/US "AI race" puts into question the pre-existing conception of the US-based AI value chain. DeepSeek's models

are significantly more cost-effective to train and develop than those from OpenAI or Meta.⁶¹ Their efficiency challenges the idea that advanced AI requires massive spending on high-end Nvidia chips. China managed to do this, notwithstanding, U.S. restrictions on advanced semiconductor sales to China, which aimed to slow AI progress but have not hindered DeepSeek's development and growth.⁶² Notwithstanding China's success in reducing costs, investments into the AI value chain for the production of generative AI services still remains high, and there's much speculation that the costs of Deepseek's models are understated.

2.4. The value of AI

31. As AI is seen as a transformative foundational technology, venture capitalists and markets have placed huge value on those companies with a leadership position in AI. OpenAI put the company's market value at \$157bn⁶³ in October 2024 and this grew to \$500bn in October 2025⁶⁴, while it made \$4.3bn in the first half of 2025⁶⁵ and is projected to generate \$29.4 billion in revenue in 2026.⁶⁶ OpenAI has now partnered with Softbank

and others in the Stargate AI infrastructure project where \$500bn of funding is expected to be channelled over the coming years⁶⁷. The AI chip maker Nvidia was propelled to the world's most valuable company with a market capitalisation of over \$3trn.⁶⁸ Perplexity AI has secured \$100 million in new capital, increasing its valuation to \$18 billion from \$9 billion in 2024.⁶⁹ x.AI's recently raised \$10bn, increasing its valuation to \$200bn.⁷⁰

- 61 Financial Post. 2025. What is China's DeepSeek and why is it freaking out AI competitors? Available online: <https://financialpost.com/technology/china-deepseek-ai-world> [Accessed on 30 January 2025].
- 62 Financial Post. 2025. What is China's DeepSeek and why is it freaking out AI competitors? Available online: <https://financialpost.com/technology/china-deepseek-ai-world> [Accessed on 30 January 2025].
- 63 TechCrunch. 2024. OpenAI raises \$6.6B and is now valued at \$157B. available online: <https://techcrunch.com/2024/10/02/openai-raises-6-6b-and-is-now-valued-at-157b/> [accessed 20 January 2025].
- 64 Gopalan (2025). OpenAI Is World's Top Startup With \$500B Valuation After Employee Share Sale. Investopedia. Available: <https://www.investopedia.com/openai-is-worlds-top-startup-with-500b-valuation-after-employee-share-sale-11822885#:~:text=OpenAI%20is%20World's%20Top%20Startup,Trade> [accessed 4 November 2025].
- 65 TechInAsia (2025). OpenAI's H1 2025: \$4.3b in income, \$13.5b in loss. Available: <https://www.techinasia.com/news/openais-revenue-rises-16-to-4-3b-in-h1-2025> [accessed 4 November 2025].
- 66 Lemkin (2025). OpenAI to Hit \$12.7 Billion in Revenue This Year. Available: <https://www.saastr.com/bloomberg-openai-to-hit-12-7-billion-this-year-but-wont-be-profitable-until-125-billion/> [accessed 4 November 2025].
- 67 Duffy (2025). Trump announces a \$500 billion AI infrastructure investment in the US. Available: <https://edition.cnn.com/2025/01/21/tech/openai-oracle-softbank-trump-ai-investment/index.html> (accessed 22 January 2025).
- 68 CompaniesMarketcap. 2025. Market capitalization of NVIDIA (NVDA). Available online: <https://companiesmarketcap.com/nvidia/marketcap/> [Accessed 24 January 2025].
- 69 Tech Funding News. 2025. AI Search disruptor Perplexity scores \$100M, soars to \$18B valuation. Available online: [AI search disruptor Perplexity scores \\$100M, soars to \\$18B valuation – TFN](https://www.techfundingnews.com/news/ai-search-disruptor-perplexity-scores-100m-soars-to-18b-valuation-tfn) [Accessed 18 July 2025]
- 70 CNBC. 2025. Elon Musk's xAI raising \$10 billion at \$200 billion valuation. Available online: [Elon Musk's xAI raising \\$10 billion at \\$200 billion valuation: sources](https://www.cnbc.com/2025/01/21/elon-musk-xai-raising-10-billion-at-200-billion-valuation.html) [Accessed 19 September 2025]

In addition, global funding trends by venture capitalists in the cloud technology market segment, indicates that around 45% of the funding in 2025 was directed to software and generative AI companies.⁷¹

32. In their respective SEC 10K filings, Google, Meta and Microsoft notes that it is expanding its investment in AI across the entire company which includes generative AI and continuing to integrate AI capabilities into its products and services⁷². This requires significant investment but is necessary to

meet the changing needs and expectations of its existing users and attract new users⁷³.

33. DeepSeek's models, such as the latest R1 model, are significantly more cost-effective to train and develop than those from OpenAI or Meta.⁷⁴ DeepSeek has claimed it took just two months to build its open source model at a cost of under \$6 million to build Nvidia's less-advanced H800 chips⁷⁵ while critics have noted that \$6 million understate the true cost of development. Either way, it appears that Deepseek's AI models remain cheaper than US-based AI models to develop.

2.5. Licensing agreements between generative AI and large media houses

34. The Inquiry has taken note of a trend by generative AI tech companies who are entering into licencing agreements with several large and often global media companies. For example, OpenAI has partnered with Axel Springer, a media and technology company which is active in more than 40 countries.⁷⁶ Through this partnership, ChatGPT users around the world will receive summaries of selected global news content from Axel Springer's media brands including Politico, Business Insider, and Bild and Welt.⁷⁷ Open AI has also established partnerships

with Le Monde and Prisa Media along with its publications such as El País, Cinco Días, As, and El Huffpost which target a Spanish-speaking market.⁷⁸ These partnerships will enable ChatGPT users to engage with Le Monde and Prisa Media's high-quality content on recent events in ChatGPT, and their content will also contribute to the training of the LLMs.⁷⁹ OpenAI also partnered with the Associated Press ("AP") including their text archive dating back to 1985 giving the platform access to quality content to train its model.⁸⁰ In exchange

71 Bain & Company. 2025. Global Venture Capital Outlook: The Latest Trends. Available online: [Global Venture Capital Trends: The Latest Industry Report | Bain & Company](#). [Accessed August 2025]

72 Microsoft Corporation SEC 10k filing for the year ended 30 June 2024, p.5; Meta Platforms Inc. SEC 10K filing for year ended 31 December 2024; Alphabet SEC 10k filing for the year ended 31 December 2023, p. 11.

73 Alphabet SEC 10k filing for the year ended 31 December 2023, p. 11.

74 Financial Post. 2025. What is China's DeepSeek and why is it freaking out AI competitors? Available online: <https://financialpost.com/technology/china-deepseek-ai-world> [Accessed on 30 January 2025].

75 Soni and Kachwala (2025). DeepSeek's Low-cost AI Spotlights Billions Spent by US Tech. *Reuters*. Available: <https://www.reuters.com/technology/artificial-intelligence/big-tech-faces-heat-chinas-deepseek-sows-doubts-billion-dollar-spend-ing-2025-01-27/> (accessed 4 February 2025).

76 Axel Springer. 2024. Company. Available at: <https://www.axelspringer.com/en/company> (Accessed on 01 January 2024)

77 Open AI. 2023. Partnership with Axel Springer to deepen beneficial use of AI in journalism. Available at: <https://openai.com/index/axel-springer-partnership/> (Accessed on 13 December 2023).

78 Open AI. 2024. Global news partnerships: Le Monde and Prisa Media. Available at: <https://openai.com/index/global-news-partnerships-le-monde-and-prisa-media/> (Accessed on 13 March 2024).

79 Open AI. 2024. Global news partnerships: Le Monde and Prisa Media. Available at: <https://openai.com/index/global-news-partnerships-le-monde-and-prisa-media/> (Accessed on: 13 March 2024)

80 O'Brien (2023). ChatGPT-maker OpenAI signs deal with AP to license news stories. Available: <https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a> (accessed 15 November 2024).

AP will leverage OpenAI's technology and product expertise^{81, 82}. Other partnerships include Vox Media, News Corp, Dotdash Meredith, Time, The Atlantic, and Financial Times.

35. These agreements involve prominent publishers, typically global publications, including News Corp, Axel Springer, The New York Times, and The Guardian, who are exploring potential agreements that may include subscription-style fees for their content. This model aims to support the development of AI technologies underpinning chatbots like OpenAI's ChatGPT and Google's Gemini.⁸³ Recently, Meta struck a multi-year AI deal with Reuters to use its news content to provide real-time answers to user queries in relation to news and current events on its Meta AI chatbot⁸⁴. Similarly, Google has also recently announced a similar licensing deal with the Associated Press to return news content for its Gemini chatbot.⁸⁵ To further support news publishers, Perplexity started its Publishers Program partnering with TIME, Der Spiegel, Fortune, Entrepreneur, The Texas Tribune, and WordPress.com⁸⁶.
36. As the landscape evolves, these agreements are seen as a crucial step in navigating the legal complexities surrounding copyright

related infringements and/ or disputes in the age of AI.⁸⁷ Establishing such agreements will be crucial for ensuring that media organizations are compensated for their intellectual property while allowing generative AI LLM developers to access high-quality content necessary for enhancing the accuracy and relevance of their models.⁸⁸

37. The New York Times filed a lawsuit against OpenAI and Microsoft, alleging copyright infringement. According to the suit, OpenAI used 'millions' of New York Times articles to train ChatGPT without authorization, potentially diminishing the value of the newspaper's subscription model.⁸⁹ The lawsuit highlights instances where ChatGPT allegedly generates exact phrases from New York Times content, enabling access without subscriptions or ad revenue.⁹⁰ Additionally, it points to Microsoft's Bing, which integrates ChatGPT, as displaying article content without referral links, thereby limiting traffic and revenue for the New York Times.⁹¹
38. In a similar vein, OpenAI has brokered deals with news media to avoid commercial copyright litigation. For example, in its deal with AP, it was noted that *"the AP deal is valuable to a company like OpenAI because it provides a trove of material that it can use for training purposes, and is also a hedge*

81 Associated Press (2023), Available: <https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/> (Accessed: 15 November 2024).

82 [36],

OpenAI Response to RFI from CCSA dated 6 December 2023, Pg 7.

83 Criddle, C., Murgia, M., & Thomas, D. (2023, June 17). AI and media companies negotiate landmark deals over news content. Financial Times. Retrieved from: <https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d>

84 Fischer, S. 2024. Meta strikes multi-year AI deal with Reuters.

85 Google. 2025. Working with The Associated Press to provide fresh results for the Gemini app. Available online: <https://blog.google/products/news/associated-press-gemini-app/> [accessed 20 January 2025].

86 Perplexity (2024). Introducing the Perplexity Publishers' Programme. Available <https://www.perplexity.ai/hub/blog/introducing-the-perplexity-publishers-program> (accessed 14 November 2024).

87 Rosenblatt, B. (2024, July 18). The media industry's race to license content for AI. Forbes. <https://www.forbes.com/sites/billrosenblatt/2024/07/18/the-media-industrys-race-to-license-content-for-ai/>

88 Criddle, C., Murgia, M., & Thomas, D. (2023, June 17). AI and media companies negotiate landmark deals over news content. Financial Times. Retrieved from: <https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d>

89 BBC. 2023. New York Times sues Microsoft and OpenAI for 'billions'. Available at: <https://www.bbc.com/news/technology-67826601> (Accessed on 08 November 2024).

90 BBC. 2023. New York Times sues Microsoft and OpenAI for 'billions'. Available at: <https://www.bbc.com/news/technology-67826601> (Accessed on 08 November 2024).

91 BBC. 2023. New York Times sues Microsoft and OpenAI for 'billions'. Available at: <https://www.bbc.com/news/technology-67826601> (Accessed on 08 November 2024).

against losing access to material because of lawsuits that have threatened its access to material". "In order to guard against how the courts may decide, maybe you want to go out and sign licensing deals so you're guaranteed legal access to the material you'll need".⁹² The reason AP was able to strike a deal like this is because it has the clout to do so notwithstanding that the news agency was in a stable condition before the deal while the rest of the industry was struggling⁹³.

39. A major concern, however, is that these partnerships can potentially reinforce an entrenchment of mainstream and global media houses, where the latter is remunerated for content provided to train large language models ("LLMs"). It also becomes the main source of "current" information that is displayed on generative AI chatbots, as well as the preferred linked source of information that will be available for users to click on when they are considering accessing the original news publication. This may come at the exclusion of other media publications, particularly the smaller and community-based publishers, from generative AI responses and the curation of news results and associated links.
40. Given that large publishers are in a position of superior bargaining power relative to smaller publishers and can leverage

this power through threat of litigation as discussed above, they are well positioned to extract favourable commercial terms relative to smaller publishers who typically do not receive a contractual agreement. Licensing arrangements that are only made up of a select group of large and global media publications will not only restrict monetisation opportunities of other (smaller) media, but it also limits and reinforces the views, perspectives and ideologies of these selected media houses. Accordingly, this perpetuates a media landscape lacking diversity and plurality of voices, particularly those of developing countries like SA and the communities within those countries.

41. Generative AI LLM developers have indicated that they support local media, for example, OpenAI has partnered with news organisations to create AI features that can support journalism and announced a partnership with the American Journalism Project, an organisation committed to restoring local news around America.⁹⁴ The partnership includes a \$10 million commitment from OpenAI to develop AI tools that can enhance the work of local journalists and address issues posed by the growing role of AI.⁹⁵ Nonetheless this level support is unlikely to address the Inquiry's concerns and does not include support for local South African journalism.

92 Nick Diakopoulos, professor of communications studies and computer science at Northwestern University, cited in O'Brien (2023), Associated Press, available: <https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a> (accessed 15 November 2024).

93 Ken Doctor, news industry analyst, cited in O'Brien (2023), Associated Press, available: <https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a> (accessed 15 November 2024).

94 Open AI. 2023. Partnership with American Journalism Project to support local news. Available at: <https://openai.com/index/partnership-with-american-journalism-project-to-support-local-news/> (Accessed on 18 July 2023).

95 [3<]

3. USE OF MEDIA CONTENT IN LLM TRAINING

42. The training process is an integral part of the development process of generative AI models, as highlighted in the previous section. The models are trained through a process that can be categorised into two phases, namely the pre-training phase⁹⁶ and the post-training phase of the model.⁹⁷

42.1. The pre-training phase is considered as the first step of the training process and it comprises the use of a wide range of data at scale, to develop the knowledge and broaden the capabilities of the foundational LLM for a wide use case.⁹⁸ At this phase of the development process, the model evaluates the proximity, order, frequency, and other attributes of portions of words called tokens from the training data, in order to establish the probabilities of relationships between the tokens.⁹⁹ Pre-training is a computation process that includes the development of the LLM's intelligence and understating of how words: (i) fit together grammatically, (ii) work together to form ideas, (iii) are sequenced to form structured thoughts or pose coherent problems.¹⁰⁰ Once trained, the base model is developed and comprises of

a large network of weights (i.e., large strings of numbers) that represent the probabilities of the relationships between the tokens. These weights are constantly changing and they enable the LLM to generate new responses to user prompts that reflect human reasoning and understanding instead of returning recited responses from the data used to train the model.¹⁰¹ Pre-trained models are not the final product of LLM development. Without any post-training they may generate biased, unverified and unreliable outputs to tasks which is why they are commonly referred to as foundational models.¹⁰² The pre-training process yields a base model (foundational) with adequate capabilities to solve novel problems that are unseen in its training data, however, the base model alone is not ready for use.¹⁰³

42.2. The post-training processes are undertaken to improve the model's performance in terms of latency, accuracy, operating costs as well as to make it user friendly, using various techniques.¹⁰⁴ Reinforced learning, is one of the techniques used to transform the model where targeted changes to the model are made

96 Pre-training is the process of training a model on a comprehensive dataset, thus allowing the model to learn the general principles relating to the task at hand. This preliminary training enables the model to detect patterns and concepts vital for generating high-quality outputs.

Kendurkar, A. 2024. Fine Tuning in Generative AI. Available at: <https://www.linkedin.com/pulse/fine-tuning-generative-ai-aniruddh-kendurkar-txgwf/> (Accessed on 14 January 2024).

97 Fine-tuning is the process of taking a pretrained machine learning model and further training it on a smaller, targeted data set. Craig, L. 2024. What is fine tuning in machine learning and AI? Available at: <https://www.techtarget.com/searchenterpriseai/definition/fine-tuning> (Accessed on 29 July 2024).

98 Competition and Markets Authority. 2023. AI Foundation Models: Initial Report.

99 [36]

100 OpenAI submission by Chris Charter dated 06 December 2023. Written response to RFI 1. Paragraph 12.9

101 OpenAI submission by Chris Charter dated 06 December 2023. Written response to RFI 1. Paragraph 12.9

102 Mozilla. 2024. Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI. Page 11.

103 OpenAI submission by Chris Charter dated 06 December 2023. Written response to RFI 1. Paragraph 12.10.

104 OpenAI. 2024. Introducing improvements to the fine-tuning API and expanding our custom models program.

through interactions of the model with human feedback.¹⁰⁵ Pre-trained LLMs may also be subjected to a fine-tuning process which is a technique of training the model further with new data that is relatively smaller than the data used for pre-training in order to improve the performance of the model on a specific task. Fine-tuning is a process that seeks to improve the effectiveness or performance of the model on specific tasks using smaller data that is relevant to the

specialised use case.¹⁰⁶ For instance a pharmaceutical company may fine-tune a pre-trained LLM using its own internal data for a drug discovery case use.¹⁰⁷ The Retrieval-augmented generation (RAG) is also another post-training technique where the developed LLM is combined with traditional information systems such as a search engine or a database to enhance the outputs generated by the model with information.¹⁰⁸

4. USE AND VALUE OF MEDIA CONTENT IN TRAINING DATASETS

43. To develop a high performance model, the volume of the pre-training data as well as its quality are considered to be important factors.¹⁰⁹ For instance OpenAI's GPT-3 was trained on 570 Gigabytes of text data.¹¹⁰ Accordingly, the majority of the highly capable models to date are pre-trained on diverse data that contains trillions of words and billions of images, videos as well as audio files.¹¹¹ Generative AI developers train their models with data that is sourced from various channels, and these include: (i). information that is publicly available from the internet, (ii). non-public information from third-parties through licensing or commercial arrangements and (iii). information from human feedback.¹¹²
44. It is considered common practice for generative AI developers to train their models with data that is solely or mostly comprised of public information that has been scrapped from the internet on various website pages.¹¹³
45. Given the substantial data requirements for pre-training, a foundational dataset for many AI models has been the Common Crawl, a corpus of approximately 9.5 petabytes of data collected from the public internet since 2008 with a monthly crawl.¹¹⁴

105 OpenAI submission by Chris Charter dated 06 December 2023. Written response to RFI 1. Paragraph 12.11

106 [3].
Microsoft. 2024. Written response to the European Commission's call for contributions on Competition in Generative AI. Page 3.

107 Meta. 2024. Written response to the European Commission call for contributions on Competition in Generative AI. Page 8.

108 Google Cloud. 2025. What is Retrieval-Augmented Generation (RAG)? Available online: <https://cloud.google.com/use-cases/retrieval-augmented-generation?hl=en> [accessed 15 January 2025].

109 Competition and Markets Authority. 2023. AI Foundation Models: Initial Report. Paragraph 3.5.

110 Microsoft. 2024. Written response to the European Commission's call for contributions on Competition in Generative AI. Page 3.

111 [3].

112 OpenAI submission by Chris Charter dated 06 December 2023. Written response to RFI 1. Paragraph 14.4
[3].

113 Competition and Markets Authority. 2023. AI Foundation Models: Initial Report. Paragraph 3.6

114 Common Crawl. 2024. Overview. Available online: <https://commoncrawl.org/overview>
Mozilla. 2024. Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI. Page 11.



- 45.1. The Common Crawl is not a copy of the entire Internet and has predominantly English content. South African content is a small part of the Common Crawl, with the .za domain accounting for 6.1m URLs from 124,000 domains representing 0.25% of the total URLs. However, some South African content also falls under the “.com” domain (42.5% of URLs) as well as the “.org” domains (5.6% of URLs), and so this figure is an underestimate.¹¹⁵
46. AI companies will also create their own corpus by scraping the internet with their automated bots for high quality and diverse content.
- 46.1. For instance, GPT-2 was trained on an OpenAI dataset called WebText using only 40 gigabytes of web pages text (i.e., 8 million web pages) curated by humans.¹¹⁶ As alleged in the New York Times lawsuit against OpenAI, this training dataset had a significant emphasis on news media website content.¹¹⁷ As confirmed by Github’s list of GPT-2 domains, 5 of the top 10 domains and 23 of the top 50 are news sites.¹¹⁸ In terms of local news publications, News24 comes in at 730 with 6494 tokens but there are no other prominent South African news site that seems to feature on the dataset.
- 46.2. GPT-3 was trained with 300bn tokens, of which c.62% were from the Common Crawl, c.19% from WebText2 and the remainder from books and Wikipedia.¹¹⁹ No information is available on the domains used in WebText2 but presumably this was also weighted heavily to news content. OpenAI has revealed nothing about GPT-4 training data, though unconfirmed reports have indicated that the pre-training data mixture included transcribed Youtube videos.¹²⁰
- 46.3. OpenAI is not alone, as the Google C4 dataset used to train Bard is reported to have 13% of tokens from news sites.¹²¹ Gemini has also used the Common Crawl¹²². Llama was trained with 3.3 terabytes sized data from Common Crawl, which constituted 67% of the pre-training data mixture of the model’s overall training data, followed by C4 (15%) which is a processed and cleaned version of Common Crawl’s web crawl corpus, Github (4.5%), Wikipedia (4.5%), Books (4.5%).¹²³
47. The Inquiry has requested the domains and tokens used in the training datasets from the generative AI LLM developers, to determine the extent South African news media content

115 Common Crawl. 2024. Statistics of Common Crawl Monthly Archives. Available online : <https://commoncrawl.github.io/cc-crawl-statistics/plots/tld/latestcrawl.html> [accessed 16 January 2025].

116 OpenAI. 2019. Better language models and their implications. Available online : <https://openai.com/index/better-language-models/> [accessed 17 January 2025].

117 The New York Times v. OpenAI and Microsoft. 2023. Complaint. United States District Court, Southern District of New York, Civil Action No. 23-cv-11195.

118 Github. 2019. OpenAI/GPT-2 domains txt. Available online: <https://github.com/openai/gpt-2/blob/master/domains.txt> [accessed 17 January 2025].

119 Lambda. 2020. OpenAI’s GPT-3 Language Model: A Technical Overview. Available online: <https://lambdalabs.com/blog/demystifying-gpt-3> [accessed 17 January 2025].

120 The Verge. 2024. OpenAI transcribed over a million hours of YouTube videos to train GPT-4.

121 The Washington Post. 2023. Inside the secret list of websites that make AI like ChatGPT sound smart. Available online: <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/> [Accessed 24 January 2024].

122 Gemini vs. ChatGPT: Which is better and What’s the Difference? Available online: <https://www.techopedia.com/gemini-vs-chatgpt-which-is-better> (accessed 22 January 2025).

123 Meta AI. 2023. LLaMA: Open and Efficient Foundation Language Models. Table 1. Papers with Code. Available online: <https://paperswithcode.com/dataset/c4> [accessed 17 January 2024].

has been used, but the Inquiry has thus far not been provided with that information.¹²⁴

48. Aside from the Common Crawl, OpenAI, Google and other AI developers have their own web crawlers that collect data from the web for training their AI LLMs.¹²⁵ The AI developers state that their bots adhere to the Robots.txt protocol which allows websites to opt out of allowing the web crawlers access to scrape their website.¹²⁶ However, the opt-out has typically been announced after their models have been released, and so only apply to future iterations of the model. Yet the prior data scrapped from the websites is presumably still available for training even if future website content is not. It has also been reported that AI developers make use of third-party suppliers of data and these are less scrupulous insofar as ignoring the robots.txt instruction from websites, and even collating information from behind paywalls.¹²⁷

49. The unauthorised use of copyrighted content has been raised by the news media and some litigation is taking place.¹²⁸ In response, some news publications have also taken steps to block certain crawlers from scrapping their content. Media24 confirmed that they blocked the automated web crawlers of generative AI developers from scrapping their publications' websites, once they became aware of the generative AI developers' actions.¹²⁹ This is also confirmed by the results from the News Homepages survey, which surveyed 1 158 news publishers across different jurisdictions to determine the extent news publications were allowing or blocking generative AI related bots to crawl their website content. For South Africa, the surveyed news publications are listed in the table below and it shows it is mainly the Arena Holdings and Media24 publication networks that have blocked the crawlers.

Table 2: List of South African news publications that have blocked/ not blocked crawlers related to generative AI

News Publication	OpenAI	Google AI	Common Crawl
amaBhungane			
Daily Maverick			
Daily Sun	X	X	X
eNCA			
Eyethu KwaZulu-Natal			X

- 124 [3].
 Meta submission by Claire Reidy dated 01 December 2023. Written response to RFI 1. Paragraph 24
 OpenAI submission by Chris Charter dated 06 December 2023. Written response to RFI 1. Paragraph 14.4-14.8.
 Google submission by Palesa Mpe dated 15 November 2023. Written response to RFI 1. Tranche 1. Paragraph 25.
 Meta submission by Derek Lotter and Claire Reidy dated 16 February 2024. Written response to RFI 2. Paragraph 6.
 xAI submission by Derushka Chetty dated 15 January 2024. Written response to RFI 2. Paragraph 2.8.3.
 OpenAI submission by Taigrine Jones dated 19 March 2024. Written response to RFI 2. Tranche 2. Paragraph 14.1.
 Google submission by Neil MacKenzie dated 01 February 2024. Written response to RFI 2. Tranche 1. Paragraph 29.1.
- 125 Google submission by Neil MacKenzie dated 01 February 2024. Written response to RFI 2. Tranche 1. Paragraph 29.1.
 Open AI. 2024. Overview of OpenAI Crawlers. Available online: <https://platform.openai.com/docs/bots/overview-of-openai-crawlers> [accessed 18 January 2025].
- 126 Google submission by Neil MacKenzie dated 01 February 2024. Written response to RFI 2. Tranche 1. Paragraph 29.1.
- 127 The Verge. 2024. Perplexity's grand theft AI. Available online: <https://www.theverge.com/2024/6/27/24187405/perplexity-ai-twitter-lie-plagiarism> [accessed 18 January 2025].
- 128 News Corporation v. Perplexity AI. 2024. Complaint. United States District Court, Southern District of New York, Civil Action No. 24-cv-7984.
 The New York Times v. OpenAI and Microsoft. 2023. Complaint. United States District Court, Southern District of New York, Civil Action No. 23-cv-11195.
- 129 Media24 - oral submission from Ms. Charlene Rolls, dated 05 March 2024, Page 19.

News Publication	OpenAI	Google AI	Common Crawl
Eyewitness News			
I'solezwe lesiXhosa			
Ilanga			
Independent Online			
Mail & Guardian			
Maroela Media	X		X
Netwerk24	X	X	X
news24	X	X	X
Oxpeckers			
thesanews			
Sowetan	X	X	
TimesLIVE	X	X	
Vrye Weekblad			

Source: Market Inquiry's own compilation sourced from News Homepages¹³⁰

50. The AI developers argue that there is no copyright infringement citing both the data mining and fair use provisions which permit transformative use not in competition with the original producer.¹³¹ They also emphasise that steps are taken to ensure there is no memorisation of training data to avoid duplication of copyright material in answers, and fair use has enabled the United States to take a lead in AI.¹³² In contrast, the New York Times, which initiated legal proceedings against OpenAI and Microsoft, argues that these companies knowingly violated Copyright Law in the arms race to get more and more data needed to train their models, and companies like Google have adjusted their terms and conditions to harvest more data.¹³³

51. AI developers also place emphasis on the need for trillions of words and the negligible role that a single news website contributes, as well as that the content does not matter to training LLMs and a news website is of equal use as a reddit conversation. However, the pursuit of deals with leading global news publishers for their volume of data for pre-training as they opt out of the web crawlers strongly suggests a different conclusion, namely that both the volume and content is particularly useful to AI developers as training data. Access to copyrighted material for training purposes have all been cited in press statements in which OpenAI has done publisher deals with the Associated Press¹³⁴, Axel Springer¹³⁵, Le Monde and

130 News Homepage. Who blocks OpenAI, Google AI and Common Crawl? Available online: <https://palewi.re/docs/news-homepages/openai-gptbot-robotstxt.html> [accessed 20 January 2025].

131 OpenAI. 2024. OpenAI and journalism. Available online: <https://openai.com/index/openai-and-journalism/> [accessed 18 January 2024].

132 OpenAI. 2023. Comments to the U.S. Copyright Office. Available online: <https://www.regulations.gov/comment/COLC-2023-0006-8906> [accessed 18 January 2025].

133 New York Times. 2023. How tech companies cut corners to harvest data for AI. Available online <https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html> [Accessed 18 January 2025].

134 New York Times. 2023. How tech companies cut corners to harvest data for AI. Available online. <https://www.ap.org/media-center/ap-in-the-news/2023/chatgpt-maker-openai-signs-deal-with-ap-to-license-news-stories/> [Accessed 18 January 2025].

135 Axel Springer. 2023. Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism. Available online. <https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism> [Accessed 20 January 2025].

- Prisa Media¹³⁶, the Financial Times¹³⁷, News Corp¹³⁸, the Atlantic and Vox Media¹³⁹, Time¹⁴⁰ and Conde Nast.¹⁴¹
52. The use of news content as well as the pursuit of news media publisher licensing deals by generative AI developers such as Open AI, Meta and Google, highlights its extraordinary value in training LLM models.
- 52.1. Besides the volume aspect of the training data (i.e., that news content also meets), the quality of the data that is used to train LLMs is essential in developing a high performing model that will generate the accurate output.¹⁴² Accordingly for text data, news content meets the high quality criteria given it is written by professional journalists that invest time as well as substantial resources in collating and editing news content that effectively captures the complexities of human realities.
- 52.2. Moreover, as news media complies with press codes, content will typically not contain offensive language or harmful content which must otherwise be extracted, unlike general content from the internet, saving costs. Furthermore news media contains a diversity of content which is useful for AI training.
- 52.3. Most news media publishers also have a large archive of content that is not readily available on their website which also offers additional volume of high quality material for training LLMs as well as a reputable source of reference for historical events.¹⁴³ For instance, OpenAI's agreements with News Corp and the Associated Press, grants OpenAI with access to not only the current news content, but also the archived content from the various publications.¹⁴⁴
- 52.4. Non-English South African language news content provides an opportunity for generative AI tech companies to train models embodying languages such as Zulu, Afrikaans, isiXhosa, etc. This can be used in, for example, translator services or translated into English to provide more content to further train existing LLMs.
53. These deals with the large global publishers are likely prompted by their array of publications and substantial archives which are not accessible through the website. Furthermore, large publishers are also able and likely to have excluded the AI web

136 OpenAI. 2024. Global news partnership: Le Monde and Prisa Media. Available online. <https://openai.com/index/global-news-partnerships-le-monde-and-prisa-media/> [Accessed 20 January 2025].

137 Financial Times. The Financial Times and OpenAI strike content licensing deal. Available online. <https://www.ft.com/content/33328743-ba3b-470f-a2e3-f41c3a366613> [Accessed 20 January 2025].

138 OpenAI. 2024. A landmark multi-year global partnership with News Corp. Available online. <https://openai.com/index/news-corp-and-openai-sign-landmark-multi-year-global-partnership/> [Accessed 20 January 2025].

139 Reuters. 2024. OpenAI signs content deal with The Atlantic and Vox Media. Available online. <https://www.reuters.com/business/media-telecom/openai-signs-content-deals-with-atlantic-vox-media-2024-05-29/> [Accessed on 20 January 2025].

140 TIME. 2024. Time and OpenAI Announce Strategic Content Partnership. Available online. <https://time.com/6992955/time-and-openai-announce-strategic-content-partnership/> [Accessed on 20 January 2025].

141 Reuters. 2024. OpenAI signs content deal with Conde Nast. Available online. <https://www.reuters.com/technology/openai-signs-deal-with-cond-nast-2024-08-20/> [Accessed 20 January 2025].

142 Microsoft. 2024. Written response to the European Commission's call for contributions on Competition in Generative AI. Page 3.

143 [38]

144 OpenAI. 2024. A landmark multi-year global partnership with News Corp. available online: <https://openai.com/index/news-corp-and-openai-sign-landmark-multi-year-global-partnership/> [accessed 20 January 2025].

The Associated Press. 2023. Press release: AP, Open AI agree to share select news content and technology in new collaboration. Available online: <https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/> [accessed 20 January 2025].

crawlers from current content, requiring deals if AI companies wished to access their copyrighted content. It was reported by the Reuters Institute that by the end of 2023, 48% of the top news publications websites across 10 different countries had blocked OpenAI crawlers and that half were blocking Google AI crawler.¹⁴⁵

54. It is also highly likely that deals with the largest global publishers are designed to strategically erode any potential copyright infringement related lawsuits as it is the large global publishers that are more likely to pursue and bankroll such litigation.¹⁴⁶ The real risk with this divide and conquer strategy is that it further bifurcates the news media market by strengthening the few global publishers in stronger financial position that already win at the subscription game at the expense of smaller publishers that are not compensated for their copyrighted material and are more likely to have their content scraped. This adversely affects media diversity, plurality and democracy.

55. Absent any regulatory action, it is unlikely that the South African media will be beneficiaries of publisher deals with AI companies. Concerning website content, news sites have to actively opt out, which means they require knowledge and skills to do so. Due to resource constraints, the majority of smaller news sites are unlikely to act, exposing their content to AI developers for free. In South Africa, the Inquiry established during the Public Hearing proceedings, that it was only Media24 of the mainstream media, now joined by Arena, that had opted out of the AI crawlers - doing so with OpenAI, Google and the Common Crawl.¹⁴⁷ The Reuters Institute study also found the results varied significantly across countries with 79% of top publishers in the US opting out but only 20% in Mexico and Poland.
56. As publisher deals reveal that news content has value for training on content that is subject to copyright, then it seems that all publishers should be compensated as even past models have laid the foundation for rapid innovation and commercial success that has been experienced by generative AI companies.

5. IMPACT OF AI POWERED CHATBOTS AND SEARCH ENGINES ON MEDIA

5.1. AI chatbots powered with search

57. Generative AI chatbots have also enhanced their capabilities by integrating search engines in the process of responding to user

queries. The integration of search engines through Application Programming Interface (API), enables generative AI chatbots to

145 Reuters Institute. 2024. How Many News Websites Block AI Crawlers? Available online : <https://reutersinstitute.politics.ox.ac.uk/how-many-news-websites-block-ai-crawlers#header--3> [accessed 19 January 2025].

146 The Conversation. 2024. OpenAI's content deal with the FT is an attempt to avoid more legal challenges - and an AI 'data apocalypse'. Available online : <https://theconversation.com/openai-content-deal-with-the-ft-is-an-attempt-to-avoid-more-legal-challenges-and-an-ai-data-apocalypse-229215> [accessed 19 January 2025].

147 Media24 - oral submission from Ms. Charlene Rolls, dated 05 March 2024, Page 19. [3<]

ground their responses to user queries with the information that they source from search engines. Grounding is the process of using LLMs with information that is use-case specific, relevant, and not available as part of the LLM's trained knowledge.¹⁴⁸ When users enter a query on chatbots, the chatbot through the back end: (i) analyses the query and sends a search query to the search engine through an API, (ii) the search engine's API processes the query on behalf of the chatbot and then returns a response to the chatbot (web retrieval), (iii) subject to the query the chatbot reviews and selects from the search engine response, for the most relevant and authoritative web sources (iv) the chatbot draws the relevant information from the chosen sources and formulate a response back to the users, (v) the response to the user is generally accompanied by links to the website pages that the chatbot sourced the information from.¹⁴⁹ This all happens in real time, from the time the user inputs the prompt to the time that the chatbot generates a result.

58. Most of the generative AI chatbots have integrated search engines into their operations for web retrieval. Meta AI formulates some of its responses to user queries with real time information sourced from both Microsoft Bing and Google Search.¹⁵⁰ Similarly, Gemini accesses topical information through its integration with Google Search's API.¹⁵¹ Both Open AI and Microsoft Copilot are integrated with

Microsoft Bing API.¹⁵² By grounding their responses with information sourced through search engines, generative AI chatbots are able to return enhanced responses to user queries, with information that is up to date (i.e., real time information), accurate and authoritative.¹⁵³

59. This helps the generative AI chatbots to address some of the inherent LLM limitations. For instance, chatbots knowledge becomes outdated due to the cut-off date on the data that is used to train the LLMs, as such they are not able to respond to topical queries. For example, OpenAI's GPT-3.5 was launched on the 30th of November 2022, but it had a cut-off date of September 2021.¹⁵⁴ OpenAI addressed this limitation through the "Browse with Bing" feature that enabled the chatbot to return real time responses to subscribed user queries sourced through Microsoft Bing.¹⁵⁵ There are also other limitations with the responses to user queries such as inaccuracies or hallucinations, ambiguity etc. that are addressed through integration with search engine API.¹⁵⁶
60. In addition to grounding responses, both Meta AI and Gemini have also developed features (highlighted in red in the figure below) that redirect users to the relevant SERP. On Gemini a click on the "double-check response" uses Google Search to find content that's likely similar to or likely different from the initial response provided by the chatbot and suggests search queries

148 Microsoft. 2023. Grounding LLMs. Available online: <https://techcommunity.microsoft.com/blog/fasttrackforazureblog/grounding-llms/3843857> [accessed 01 November 2024].

149 [30].

150 Microsoft Bing Blogs. 2023. Expanding Our AI Partnership with Meta. Available online: [https://blogs.bing.com/search/september-2023/Expanding-Our-AI-Partnership-with-Meta-\(1\)](https://blogs.bing.com/search/september-2023/Expanding-Our-AI-Partnership-with-Meta-(1)) [accessed 14 November 2024].

151 Google. 2024. Gemini API and Google AI Studio now offer Grounding with Google Search. Available online: <https://developers.googleblog.com/en/gemini-api-and-ai-studio-now-offer-grounding-with-google-search/> [accessed 14 November 2024].

152 Microsoft. 2024. Overview of Microsoft Copilot. Available online: <https://learn.microsoft.com/en-us/copilot/overview> [accessed 14 November 2024].

153 [30].

154 OpenAI. 2022. Introducing ChatGPT. Available online: <https://openai.com/index/chatgpt/> [accessed 19 November 2024].

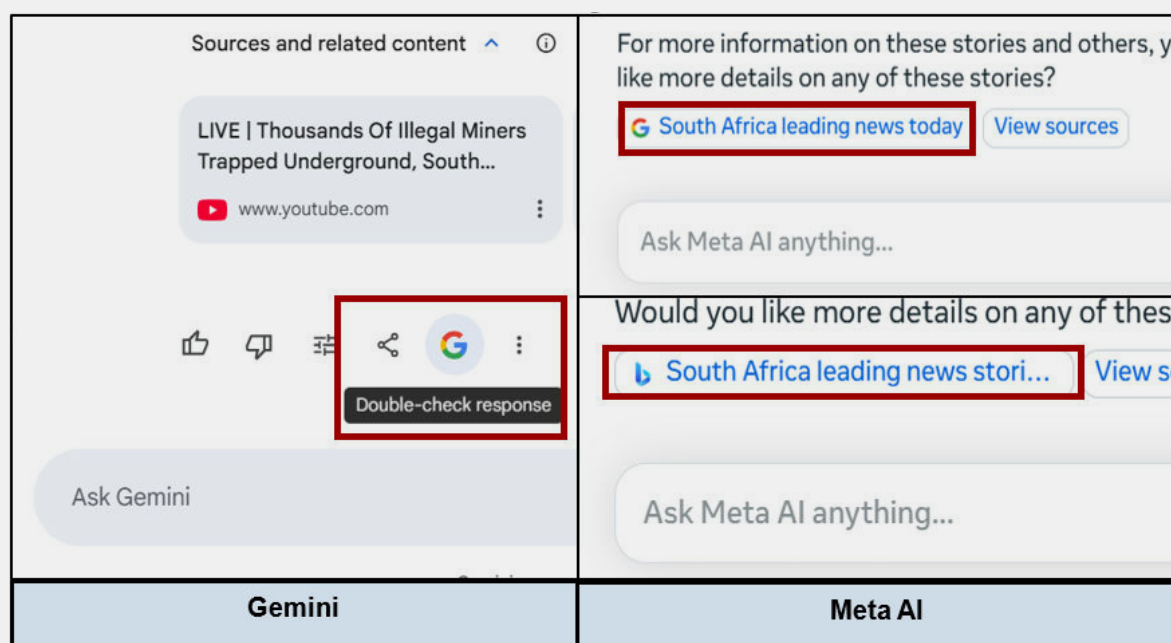
155 The Verge. 2024. ChatGPT can now search the web in real time. Available online: <https://www.theverge.com/2023/9/27/23892781/openai-chatgpt-live-web-results-browse-with-bing> [accessed 19 November 2024].

156 [30].

that users can select, and a click on the suggested queries redirects users to a new tab that returns Google's SERP in response to the selected query.¹⁵⁷ On Meta AI a click

on either the Google Search or Bing favicon automatically opens a new tab and returns SERP of the relevant search engine.

Figure 5: Integration of search engine features within Gemini and Meta AI



Source: Author's own compilation of screenshot of Gemini and Meta AI returned responses

61. Open AI has taken web retrieval further with its recent launch of ChatGPT search.¹⁵⁸ According to OpenAI, ChatGPT Search is fine-tuned version of its GPT-4o that has been post-trained with novel synthetic data generation techniques and it leverages from third-party search providers as well as from content provided to it by its partners. News publishers involvement and contributions in the development of ChatGPT Search was singled out by OpenAI, "we collaborated extensively with the news industry and carefully listened to feedback from our global publisher partners".¹⁵⁹

62. Though Open AI has not outrightly characterised ChatGPT Search as a search engine, the firm's description of its capabilities are similar to those that are carried out by a search engine. According to Open AI, ChatGPT Search enables users to: (i) search the web faster and better than before, with the relevant links that users would have previously had to go to a search engine for (ii) access the blended benefits of a natural language interface with the value of up-to-date sports scores, news, stock quotes, and more, (iii) automatically search the web depending on the query but users can also

157 Gemini Apps. 2024. Double-check responses from Gemini Apps. Available online: <https://support.google.com/gemini/answer/14143489?hl=en&co=GENIE.Platform%3DDesktop&oco=0> [accessed 18 November 2024].

158 OpenAI. 2024. Introducing ChatGPT search. Available online: <https://openai.com/index/introducing-chatgpt-search/> [accessed 15 November 2024].

159 OpenAI. 2024. Introducing ChatGPT search. Available online: <https://openai.com/index/introducing-chatgpt-search/> [accessed 20 January 2025].

manually choose to search by clicking the web search icon, and (iv) access up to date news and information such as weather, stock, sports and maps through partnerships. ChatGPT search is currently available to ChatGPT Plus and Team users, as well as

SearchGPT waitlist users, but the firm plans to have a wider roll out for enterprise, Edu users and all free users overtime. There have also been unconfirmed reports that Meta AI is also developing a “search engine”.¹⁶⁰

5.2. Use of AI search for news and use of sources

63. Generative AI chatbots have capabilities to respond to a wide scope of user queries, as a result users can use it to complete various tasks and source information. The emergence of generative AI chatbots has therefore created an additional channel that users can turn to for news content. Understanding the extent of this is important for the Inquiry, including the interactions chatbots have with news publication webpages. The Inquiry requested various data points from generative AI developers, and the common response initially was that news is not a common use case, and that they do not track or measure user engagement.¹⁶¹

64. The way generative AI chatbots have evolved, points towards a strong use case of AI chatbots for news content. Firstly, news content is accessible on generative AI chatbots as users can access the latest news updates, and generate news summaries and explanations about current events. Topics distribution on Similarweb data that tracks the browsing habits of users in accordance to the topics that are most searched for by users on each AI Chatbots web domain shows that news is one of the most popular topics that users are interested in across various chatbots as shown by the figure below.

160 The Verge. 2024. Meta is reportedly working on its own AI-powered search engine, too. Available online: <https://www.theverge.com/2024/10/28/24282017/meta-ai-powered-search-engine-report> [accessed 31 October 2024].

Reuters. 2024. Meta builds AI search engine to cut Google, Bing reliance, the Information reports. Available online: <https://www.reuters.com/technology/artificial-intelligence/meta-develops-own-ai-search-engine-cut-reliance-google-bing-information-reports-2024-10-28/> [accessed 31 October 2024].

161 [3<]



objective of providing people the ability to make informed choices based on reliable information and news sources.”¹⁶³

“OpenAI will receive access to current and archived content from News Corp’s...”

- 65.3. In the statement announcing the partnership, Axel Springer indicated that the agreement with OpenAI will enrich users’ experience on ChatGPT as it will add recent and authoritative content across a wide array of topics on ChatGPT and users will receive summaries of selected news content from across Axel Springer brand portfolio including paid content.¹⁶⁴
- 65.4. Similarly, Google also noted the value of news content on its Gemini chatbot.¹⁶⁵
“...the AP will now deliver a feed of real-time information to help further enhance the usefulness of results displayed in the Gemini app. This will be particularly helpful to our users looking for up-to-date information.”
- 65.5. Other AI developers have also entered into agreements, with limited to no information shared about the partnership arrangements. MetaAI has recently entered into an agreement with Reuters, that will enable the chatbot to use the publication’s content in its operations.¹⁶⁶ Microsoft Copilot through partnerships with Reuters, Axel Springer, Hearst Magazines, USA TODAY Network and Financial Times draws on news

content for the recently launched Copilot Daily.¹⁶⁷

66. Thirdly, the Inquiry survey sought to understand the extent to which South Africans are using AI Chatbots for news queries and the likelihood of them doing so in the future.
 - 66.1. The results confirm that AI Chatbots are already being used extensively for news queries, with 65% of the respondent confirming that they have used it. Of this, 25% of respondents indicated that they use AI Chatbots frequently to access news and a further, 20% occasionally and the other 20% less frequently (i.e., having used chatbots once or twice).¹⁶⁸
 - 66.2. The survey also confirmed that usage is likely to increase as 57% of respondents indicated they are likely or very likely to use AI Chatbots to access news in future.¹⁶⁹
 - 66.3. The Inquiry survey also established that a high proportion of South Africans already feel snippets can provide sufficient information and are a convenient time-efficient way to consume news that avoids having to pay for it, then it follows logically that a more extensive AI summary of news is likely to exacerbate the habit of not clicking through.
67. Fourthly, the Inquiry also experimented with each of the AI chatbots and searched on how often users search for news related

163 OpenAI. 2024. A landmark multi-year global partnership with News Corp. Available online: <https://openai.com/index/news-corp-and-openai-sign-landmark-multi-year-global-partnership/> [accessed 20 November 2024].

164 Axel Spring. 2023. Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism. Available online: <https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism> [accessed 20 November 2024].

165 Google. 2025. Working with The Associated Press to provide fresh results for the Gemini app. Available online: <https://blog.google/products/news/associated-press-gemini-app/> [accessed 23 January 2025].

166 Reuters. 2024. Meta Platforms to use Reuters news content in AI chatbot <https://www.reuters.com/technology/artificial-intelligence/meta-platforms-use-reuters-news-content-ai-chatbot-2024-10-25/> [accessed 20 November 2024].

167 Microsoft. 2024. An AI companion for everyone. Available online: <https://blogs.microsoft.com/blog/2024/10/01/an-ai-companion-for-everyone/> [accessed 20 November 2024].

168 Redflank Competition Commission of South Africa Survey (“Inquiry survey”) observations. Slide 62.

169 Inquiry survey report slides 112-113

- queries and how the news content fares in comparison to other types of queries. Surprisingly, the chatbots provided “self-introspective” responses based on their own internal data, observation of query trends and engagements with users (i.e., see appendix A-D). The general response from the chatbots is that news queries account for a significant portion of the overall search queries conducted by users on their respective platforms, and that the frequency of news related search queries relative to other query types is affected by factors such as global events and breaking news stories which are also influenced by the users location, interest etc. This is unsurprising as AI chatbots are query-based like search engines and the experience in search is that news queries rank highly.
68. The prominence of news content in AI chatbots adds to and is a continuation of a long going trend in digital markets. In the early years of development and scaling, news content has been the “go to content” for bolstering user engagement by newly established online platforms. For instance, the Google domain was registered in 1997 and in 2002 Google News was established.¹⁷⁰ Facebook became operational in 2004, and in 2006 the platform introduced News Feed.¹⁷¹ Platforms such as X have also developed dedicated news feeds or news tabs. The Inquiry has also established in Annexure 3 that news content features substantially and drives user engagement in recently established platforms such as Google Discover which currently consists of [50-60]% news publisher results in the feed.
 69. The importance of news content as a key use case in generative AI chatbots is evident by the recent announcements by AI developers of their latest features. When launching the ChatGPT search feature, OpenAI announced that the feature will include a dedicated news tab that will generate the latest news for users.¹⁷² Microsoft announced in the recent Copilot update that the platform will now include a new feature called “Copilot Daily”, which Microsoft describes as a “personalized audio briefing with news and weather updates, perfect for your morning routine”.¹⁷³
 70. Similar to AI powered search engines, the main concerns with AI chatbots is whether users are clicking on the source links that accompany responses to user queries or there is cannibalisation of traffic as users only end their search journey with the summaries that are returned by the AI chatbots. As highlighted in the introductory paragraph, the Inquiry requested various data from all the generative AI powered search engines and chatbots, and this included click through rates for the source links provided in response to news queries and it was not provided. Astonishingly, one company that indicated it had no reliable information continued to state that showing links within the summary was driving higher traffic to publisher websites.¹⁷⁴
 71. Accordingly, the Inquiry can only draw one conclusion from the evasive behaviour by AI developers in failing to provide the most basic information on CTRs for news query source links, and that is the results confirm the concerns of the news media, that users are rarely clicking through to the news websites

170 Google Search MDPMI Public Hearings Presentation dated 12 March 2024. Slide 33.

171 CNET. 2014. Facebook then and now (pictures). Available online: <https://www.cnet.com/pictures/facebook-then-and-now-pictures/> [accessed 21 November 2024].

172 OpenAI. 2024. Introducing ChatGPT search. Available online: <https://openai.com/index/introducing-chatgpt-search/> [accessed 15 November 2024].

173 Microsoft. 2024. Microsoft Introduces a More Personalized Copilot with Voice and Vision Features. Available online: <https://news.microsoft.com/source/asia/2024/10/09/microsoft-introduces-a-more-personalized-copilot-with-voice-and-vision-features/> [accessed 20 November 2024].

174 [3<]

where news publishers can monetise that information used in the AI summaries. Web data from Similarweb shows that a collective total of 405.3 million traffic visits to ChatGPT, Copilot and Perplexity in South Africa, the chatbots only referred 0.3% (i.e., 465 thousand) of traffic to news media.¹⁷⁵ The

variety of deals with the largest global news publishers is also indicative of the lack of indirect monetisation through referral traffic and the need to compensate these news publishers for the content that will be shown by AI Chatbots.

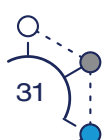
5.3. Potential to favour foreign contracted media by chatbots

72. The licensing agreements that generative AI developers have entered into with news publications to date have all been with a few large foreign publications. These selective content agreements pose a risk where the large news publishers will be favoured as the most relevant news sources and as such they stand to benefit the most from any referral traffic that does occur. One of the main reasons cited for content deals is to ensure that AI Chatbots are using reliable sources of information that will enhance the existing knowledge of the AI chatbot so as to prevent ‘hallucinations’, inaccuracies and outdated responses by AI Chatbots.¹⁷⁶ If these content deals provide that reliable source, then the AI Chatbots will rely more on those sources for news queries. Whilst no South African-specific publisher content deals have been concluded, if these were to occur then AI Chatbots may similarly favour a specific South African source for news content.

73. As with the compensation for news media training data, the selective deals with the largest global publishers to the exclusion of all others will further strengthen those large publishers whilst weakening the smaller publishers and those in developing countries like South Africa. This is because those deals will reduce the litigation of copyright infringement by those publishers capable of bankrolling such litigation. It means that generative AI relies on large global media for output and by extension be biased to the views and ideology of these big publications. This will reduce media diversity and plurality, particularly the voice from the Global South and marginalised communities and undermine their ability to generate revenue from generative AI tech.

¹⁷⁵ The data is for the period of January 2025 to September 2025.

¹⁷⁶ OpenAI. 2024. A landmark multi-year global partnership with News Corp. Available online: <https://openai.com/index/news-corp-and-openai-sign-landmark-multi-year-global-partnership/> [accessed 25 November 2024].
Axel Springer. 2023. Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism. Available online: <https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism> [accessed 25 November 2024].



6. COMMENTS ON THE PROVISIONAL REPORT AND MDPMI VIEWS

74. On the Generative AI segment, the MDPMI received submissions from various stakeholders, namely generative AI developers, news media as well as other stakeholders with an interest in sustainability of the news media industry.

75. From the end of AI developers, Google is the only AI developer that comprehensively engaged the Provisional Findings, while other AI developers made submissions on selected sections of the Provisional Report. The submissions are highlighted in detail in the sections that follow below, these sections also capture the MDPMI's views.

76. From the end of AI developers, all the AI developers that made submissions, raised various arguments which were mainly in disagreement with the Provisional Report's findings.

76.1. Google described the Provisional Report's findings in relation to the role of news in the large language model ("LLM") powering Gemini and the design of Gemini Apps, as incorrect. According to Google the MDPMI's findings are based primarily on unfounded speculations about what may or may not happen in the future, second-guessing future developments in this evolving and unpredictable space.¹⁷⁷

76.2. OpenAI downplayed its own market size, arguing that many of the findings and remedies are intended

for entrenched technology firms that are dominant in other digital market segments (i.e., search, advertising, and social media). According to OpenAI, such remedies are not suitable for OpenAI, and they may hinder product design. OpenAI also argues that the provisional findings and remedies, mandate particular commercial relationships, constraining OpenAI's and other companies' ability to respond to a dynamic environment at a time where numerous smaller firms (i.e., including OpenAI) are competing fiercely with established big tech companies.¹⁷⁸

76.3. Meta also raised their own issues, citing various issues which Meta considers to be failures by the MDPMI.¹⁷⁹ According to Meta the report fails to define a relevant market and establish whether it holds any market power in a relevant market. Meta also argues that the report does not establish a causal link between Meta's position and alleged conduct and the findings of adverse effect on competition. Meta also submits that the report does not explain how the proposed remedies address any of the news industry's fundamental societal and other causes of alleged decline and that it provides no evidence that Meta's generative AI products have had any impact on the news industry nor could they have at such an early stage when this space is still developing. Lastly, Meta argues

177 Google written submission on the publication of the MDPMI Provisional Report submitted by Neil Mackenzie dated 25 April 2025.

178 OpenAI written submission on the publication of the MDPMI Provisional Report submitted by Chris Charter dated 02 May 2025.

179 Meta written submission on the publication of the MDPMI Provisional Report submitted Thandeka Masimla dated 05 May 2025.

that there is no legitimate grounds to conclude that any or its practices in any markets have adverse effect on competition in South Africa.

- 76.4. [3X] submitted that the usage of [3X] is very low in South Africa relative to its larger rivals (i.e., Google or Meta AI), as such it is highly

unlikely that news publishers rely on them for searches in the same manner they would for its larger rivals.¹⁸⁰

77. Submissions from news publishers were mainly focused on the Provisional Report's proposed remedies, which they were mainly in support of.¹⁸¹

6.1. News engagement on AI chatbots

Stakeholder views on news engagement on AI chatbots

78. The MDPMI made a provisional finding that in South Africa, OpenAI's ChatGPT, Meta AI, Google's Gemini and Microsoft's Copilot will have the largest collective impact of all AI products on the news media, due to their extensive distribution through standalone sales and partnerships, social media, Android devices, and desktops respectively.
79. Google has submitted that this assertion is unfounded, raising two main issues.¹⁸² Firstly, Google argues that the finding is premised on assumptions, as it is not substantiated by any evidence that Gemini has any significant impact on South African news publishers. The AI developer also added that the MDPMI failed to establish a connection between Gemini use and the news industry in South Africa, and that the Provisional Report does not demonstrate the Gemini App causes harm to news publishers nor isolate the effect of Gemini from broader developments in the market.¹⁸³

80. Secondly, Google argues that the finding is in contradiction to several factors, namely that news content is not a significant source of the data used to train Gemini and is entirely substitutable. According to Google the LLM underpinning Gemini is trained on a very broad range of datasets, with news content representing only a portion of training data and that news content is not unique or essential, as the model is designed to learn statistical patterns in language and not absorb or reproduce the substantive content of individual sources. Google also argues that Gemini is not inherently designed to rely on news content as it is not primarily optimised for news-seeking queries, and it does not function as a platform for news consumption. According to Google, (i) Gemini does not engage with news content in a way that could be deemed as impactful or disruptive to news publishers, (ii) Gemini processes and generates responses based on linguistic patterns, not specific media sources, (iii) Gemini employs a grounding mechanism to enhance the accuracy and recency of the information it generates and news publishers have an option to opt-out and (iv) Gemini is not a news aggregator or

180 [3X].

181 Arena written submission on the publication of the MDPMI Provisional Report submitted by Robin Wheatley dated 07 April 2025.

Capital News written submission on the publication of the MDPMI Provisional Report submitted by Terrence Lane dated 25 April 2025.

PSS written submission on the publication of the MDPMI Provisional Report submitted by Terrence Lane dated 25 April 2025.

182 Google written submission on the publication of the MDPMI Provisional Report submitted by Neil Mackenzie dated 25 April 2025. Paragraph 5.5

183 [3X].



a platform that redistributes content from news websites as its primary purpose is to create new content.

MDPMI's Views on news engagement on AI chatbots

81. The MDPMI finds that while news media is a relevant use case for Generative AI training, the overall use of South African news in training AI models is small. The Inquiry also finds that AI chatbots are used for obtaining news, however, this constitutes a small and declining proportion relative to total AI chatbot usage. Accessing of news through AI chatbots also depends on the market positioning and distribution channels of the chatbot.
82. Adoption of Generative AI is growing rapidly in South Africa, and it is disrupting the way South African users are searching for information, navigating the internet and consuming news content which has an impact on South African news media industry. For instance, in June 2025 OpenAI's ChatGPT had a daily average of [X] users in South Africa.¹⁸⁴ Similarly, Meta AI had over [X] monthly users, and processed over [X] within the same period.¹⁸⁵ Microsoft Copilot had a total of [X] monthly active users in November 2024, and by August 2025 the number grew by over 200% to a total of [X] monthly users.¹⁸⁶ Other Meta AI data points also show similar trends, for instance the monthly sessions totalled to [X] in November 2024 which increased by 311% to [X] by August 2025, highlighting the rapid adoption of AI in South Africa. Though the rise of Generative AI has been rapid, search engines (including AI integrated search) will

most likely remain an integral part of the user search journey particularly in terms of navigating the internet, and AI chatbots will serve a complementary role for the time being. It is uncertain, however, how AI chatbots might evolve over time and could potentially become the predominant tool for searching for information. For example, in the recent judgement at the *US District Court for the District of Columbia*, Judge Mehta found that the rapid development of Generative AI can potentially offer competitive constraints to Search¹⁸⁷.

83. The distribution channels at the disposal of Google, Meta, Microsoft and OpenAI are currently unmatched and it is not easily replicable. Microsoft, Meta, and Google are all large multinational conglomerate players within the technology space. They have substantial resources and a large ecosystem made up of a range of digital products and services that have global reach. These companies have leveraged their resources (i.e., finances, data centres, cloud systems, processing chips, human resources and general ecosystem) to build and fast-track their generative AI capabilities both at the model development and the deployment phase.
84. Evidence shows that AI developers have aggressively integrated their AI models and chatbots across their digital services, enabling users to access these capabilities on various channels. As highlighted by the table below, the digital platforms and digital systems that the generative AI developers have deployed their chatbots in are some of the most popular digital products across different sets of devices, and these span

184 Own calculations using OpenAI Annexure RFI4 submission dated 12 September 2025, in response to the request for information, submitted by Taigrine Jones.

185 Meta AI written submission by Thandeka Masimla dated 15 August 2025, in response to the request for information. Paragraph 28.

186 Own workout using Microsoft Annexure submission by Anneri Coetzee dated 05 September 2025.

187 Godoy (2025). Google Ruling Shows How Tech Can Outpace Antitrust Enforcement. *Reuters*. Available: <https://www.reuters.com/legal/litigation/google-ruling-shows-how-tech-can-outpace-antitrust-enforcement-2025-09-04/> [accessed 22 October 2025].

across a wide range of segments. The distribution channels include a dedicated application, website domains, operating

systems, website browsers and various other pre-existing digital services like social media platforms etc.

Table 3: List of channels that Generative AI Chatbots can be accessed

Chatbot	Dedicated APP	Website	Operating System	Web Browser	Other access channels
ChatGPT	Yes	Yes	iOS iPadOS macOS**	ChatGPT Atlas	-
Copilot	Yes	Yes	Windows	Microsoft Edge	• Bing (custom version) • Copilot on Telegram, Viber, WhatsApp, and GroupMe** • Microsoft 365 Apps (Excel, Outlook, OneNote, PowerPoint and Word) • Xbox
Gemini	Yes	Yes	Android	Chrome	• AI Mode (Custom version) • AI Overviews (Custom version) • Gemini For Home • Google Workspace (Gmail, Google Docs, Google Sheets, Google Slides, etc) • Gemini Code Assist
Meta AI	Yes	Yes	-	-	• Facebook • Instagram • Messenger • WhatsApp • Ray-Ban Smart Glasses

Source: Own compilation using various sources ¹⁸⁸

85. The data submitted by AI developers clearly highlights that there is value in having an extensive distribution network. Approximately [80-90%] of Meta AI prompts are generated by users through WhatsApp, but Meta also has five other access channels that users can use to access Meta AI which are likely generate more user engagements over time.¹⁸⁹ Microsoft access channels are more spread out in terms of user intensity, and

these are across four channels accounting for [60-70%] of the user prompts with five other channels accounting for the rest.¹⁹⁰ Though its rival AI competitors submitted the requested data above, OpenAI submitted that providing the requested data would require a manual, resource-intensive effort.¹⁹¹

86. While OpenAI can be considered as an outlier in comparison to these conglomerate

188 Microsoft. 2025. Where can you access Copilot? Available online: <https://www.microsoft.com/en-us/microsoft-copilot/for-individuals/do-more-with-ai/general-ai/where-can-you-access-copilot?form=MA13NB>

OpenAI. 2024. OpenAI and Apple announce partnership to integrate ChatGPT into Apple experiences. Available online: <https://openai.com/index/openai-and-apple-announce-partnership/>

Google. 2024. Made by Google 2024. Available online: <https://blog.google/products/platforms-devices/made-by-google-2024-collection/>

Google. 2024. Android is reimagining your phone with Gemini. Available online: <https://blog.google/products/android/android-gemini-google-ai/>

Microsoft. 2025. Find Copilot on WhatsApp, Telegram, and more. Available online: <https://www.microsoft.com/en-us/microsoft-copilot/for-individuals/do-more-with-ai/ai-for-daily-life/find-copilot-on-whatsapp-telegram-and-more?-form=MA13NB>

189 Own workout using from Meta written in response to request for information, submitted by Thandeka Masimla dated 15 August 2025. Paragraph 69.

190 Microsoft Annexure A submission by Anneri Coetzee dated 05 September 2025, in response to a request for information.

191 OpenAI written submission by Taigrine Jones dated 12 September 2025 in response to request for information. Paragraph 12.4.2.

tech firms, given that it is a new player without a large portfolio of digital products etc. OpenAI is currently valued at \$500 billion enjoying a staggering year on year growth of 218% from a valuation of \$157 billion in 2024.¹⁹² OpenAI has also developed various channels of its own to distribute its ChatGPT Chatbot, however the partnership with Microsoft is significant as the ChatGPT model powers Microsoft's Copilot which spans across Microsoft products.¹⁹³ OpenAI also enjoys massive global reach with the Apple partnership which covers hundreds of millions of users globally spanning across the various Apple's devices.¹⁹⁴

87. On the issues raised by Google, the MDPMI provides the following evidence:

87.1. Firstly, several AI developers have withheld information that could shed more light on the true impact of AI chatbots on the news media industry. Besides this, Google has not backed its own rebuttal of the MDPMI's provisional findings with any data or evidence to substantiate its claims. As highlighted above the use of AI is growing rapidly in South Africa and its growth is linked to the broad use case that it offers to users with.

87.2. Secondly, the connection of Gemini and other AI Chatbots with South African news is established because of the use of publicly available

information in training AI models which is also made up of South African news publications, albeit reflecting a relatively small proportion of training material.¹⁹⁵ Another connection is the grounding of user prompts with South African news content.

87.3. News media represents a small and declining proportion of total user engagement on Generative AI chatbots. For example, on ChatGPT the number of current affairs prompts has gradually increased in absolute terms but is still a far lower share than on search and is a declining share of total prompts as other use cases grow much faster.¹⁹⁶ Similarly for Meta AI, news prompts are a much smaller share of total prompts as other use cases for the AI chatbot predominate.¹⁹⁷ Using Similarweb data, monthly average user segmentation of AI chatbots and tools puts the news and media industry segment higher for web traffic at 6% of total usage AI chatbot usage over the period January 2025 to September 2025¹⁹⁸ (See Figure 7 below). Over the same period, the news and media segment has declined by 16% of total usage, consistent with the trend seen in data submitted by GenAI companies.

192 OpenAI. 2024. New funding to scale the benefits of AI. Available online: <https://openai.com/index/scale-the-benefits-of-ai/>

Bloomberg. 2025. OpenAI Valuation Reaches \$500 Billion, Topping Musk's SpaceX. Available online: <https://www.bloomberg.com/news/articles/2025-10-02/openai-completes-share-sale-at-record-500-billion-valuation>

193 Microsoft. 2025. Microsoft and OpenAI evolve partnership to drive the next phase of AI Available online : <https://blogs.microsoft.com/blog/2025/01/21/microsoft-and-openai-evolve-partnership-to-drive-the-next-phase-of-ai/>

194 Apple. 2024. Introducing Apple Intelligence, the personal intelligence system that puts powerful generative models at the core of iPhone, iPad, and Mac. Available online: <https://www.apple.com/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/>

OpenAI. 2024. OpenAI and Apple announce partnership to integrate ChatGPT into Apple experiences. Available online: <https://openai.com/index/openai-and-apple-announce-partnership/>

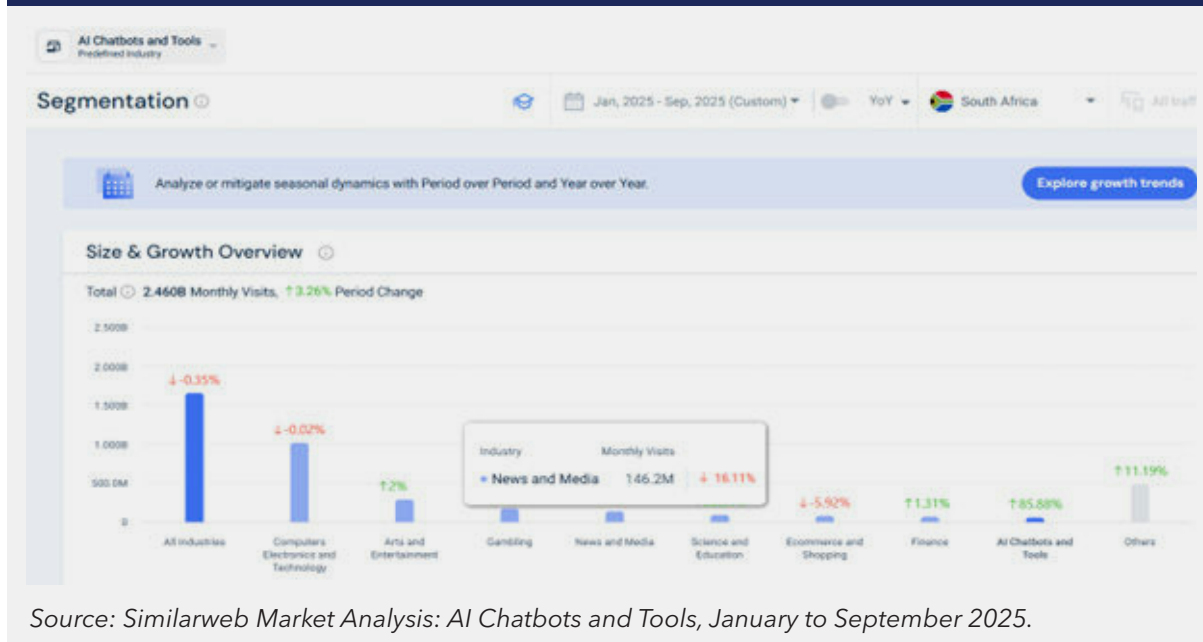
195 For example, South African content is a small part of the Common Crawl, with the .za domain accounting for 6.1m URLs from 124,000 domains representing 0.25% of the total URLs. When including .com urls, this number is larger but still very small.

196 Prompts for current affairs constituted [3<] of total prompts in July 2025 and declined from [3<] in January 2025. Own calculations from data submitted in OpenAI RFI4.1.

197 Meta AI RFI 4. Foreign news prompts were [3<] of total prompts in June & July 2025.

198 A major limitation of this data is that it reflects mobile and web usage but not app-based usage.

Figure 7: News segment as a proportion of total web traffic in AI Chatbots



88. The harm on the news media industry that the MDPMI is mainly concerned about includes the unauthorised use of copyrighted news content to train AI models without any compensation to the original producers of this content as well as the lack of value sharing in terms of the news content being

used to ground responses as the publishers are not generating any equivalent traffic or compensation in return. Furthermore, news websites are potentially disintermediated and must compete with its own or similar news content surfaced within AI chatbot platforms. This is discussed later.

6.2. Crawling of South African news content for training and opt-out

89. The MDPMI made a provisional finding that South African news media continues to provide access to AI web crawlers to scrape their website content for training purposes, and as a result appears either ill-informed or ill-equipped to protect their content from AI web crawlers, due to the opt-out rather than opt-in requirements. Restricting access places SA news media in a better position to negotiate content deals with AI developers, whilst still providing access for public interest GenAI projects that deploy their own web crawlers.

Stakeholder views on content used for training AI Chatbots

90. In their submissions AI developers have emphasised on the point that their models are trained on broad range of datasets and as such news content represents only a portion of the training data.¹⁹⁹ OpenAI argues that no one site or series of sites in South Africa or elsewhere play a significant role in the training data as a whole or influence the model's ability to learn language or otherwise meet what consumers want across

¹⁹⁹ Google written submission on the publication of the MDPMI Provisional Report submitted by Neil Mackenzie dated 25 April 2025. Paragraph 5.8.1.

a wide array of use cases.²⁰⁰ Google noted the MDPMI's provisional finding that "... South African news media is likely to be a small fraction of the total training dataset based on information currently available to the Inquiry..". Google has argued that the MDPMI Provisional report has not taken into account that Gemini has limited engagement with South African news content and that the MDPMI contradicts itself by suggesting that this minor fraction has had a meaningful impact on the development of AI models such as Gemini.²⁰¹ Google has also pointed out that the Provisional Report does not fully reflect how minor this contribution is, given that Gemini is trained a broad range of data sources, with limited interaction with South African news content and minimal revenue generated.

MDPMI's Views on content used for training AI Chatbots

91. The MDPMI maintains its view that news content plays an important role in the training of AI models, and this role is more important than what the AI developers are willing to acknowledge.

91.1. None of the AI developers have denied that they have used news content which may include South African news publishers' content to train their AI foundational models.

91.2. The objective of AI developers is to expose their models to a substantially large and diverse dataset sourced from various sources, as such any categorisation of the dataset according to content type will undoubtedly result in the various categories accounting for a small portion of the overall training data of AI models.

91.3. If news content was offering little to no value in the process of developing foundational AI models, none of the AI developers would include news content in their datasets let alone sign commercial licensing agreements with news media publishers.²⁰²

91.4. The MDPMI has on numerous accounts requested the relevant information from AI developers in relation to their training data and they have not submitted this critical evidence to the MDPMI to date.²⁰³ For instance, the last request for information letter issued to AI developers, the MDPMI provided a list of news website domains for AI developers to confirm whether any of the listed domains formed part of the training dataset of their latest or previous models that power their chatbots. Microsoft submitted that it is not in possession of the data used to train the model that powers Copilot, given that Copilot is powered by OpenAI's models.²⁰⁴

200 OpenAI written submission on the publication of the MDPMI Provisional Report submitted by Chris Charter dated 02 May 2025. Paragraph 13.1

201 Google written submission on the publication of the MDPMI Provisional Report submitted by Neil Mackenzie dated 25 April 2025. Paragraph 5.20.

202 Microsoft submission by Amanda Mfuphi dated 08 April 2025. Written response to the Provisional Report. Page 2.

203 Competition Commission South Africa Fourth Request for Information by the Media and Digital Platforms Market Inquiry to Google (Gemini) dated 16 July 2025. Paragraph 3.

Competition Commission South Africa Fourth Request for Information by the Media and Digital Platforms Market Inquiry to Meta (Meta AI) dated 04 July 2025. Paragraph 3.

Competition Commission South Africa Fourth Request for Information by the Media and Digital Platforms Market Inquiry to Microsoft (Copilot) dated 21 July 2025. Paragraph 3.

Competition Commission South Africa Fourth Request for Information by the Media and Digital Platforms Market Inquiry to OpenAI dated 15 July 2025. Paragraph 3.

Competition Commission South Africa Fourth Request for Information by the Media and Digital Platforms Market Inquiry to xAI dated 21 July 2025. Paragraph 3.

204 Microsoft submission by Xolani Nyali dated 05 September 2025. Written response to a request for information letter. Page 5.

Meta submitted that they use publicly available data, and that the exact data mix used is subject to decisions made by researchers and its training systems and that any explanation of this is beyond the scope of the MDPMI terms of reference.²⁰⁵ OpenAI submitted that they are not able to respond to this request, as training data is commercially sensitive and that the request requires manual and resource intensive efforts to respond to.²⁰⁶ xAI submitted that they do utilise publicly available information to train Grok but it does not specifically seek out news content from South Africa.²⁰⁷ If this evidence was provided, it would enable the MDPMI to satisfy itself about the true extent of the use or non-use of South African news content in the training and development of foundational AI models.

Stakeholder views on opt-outs

92. In terms of the Inquiry's findings in relation to news media protecting their content, AI developers have made submissions countering the MDPMI's finding.

92.1. Google submitted that the MDPMI is incorrect on the ability of South African media News publishers to opt-out.²⁰⁸ According to Google, news media has full control over the use of their content for AI training through Google-Extended, which allows website publishers to fully exclude their content in the training

of future generations of foundational models powering Gemini Apps or Vertex AI generative APIs as well as from grounding responses to user prompts.²⁰⁹ Google also acknowledged that training and development of the older models previously did not filter for Google-Extended, meaning website publishers could not exclude their content from model training. Google also argues that the decision to not opt-out suggests news website publishers are not objecting to their content being used for training or grounding. Lastly, Google submits that the MDPMI misstates that news publishers have no real scope to opt out of AI Chatbot search. According to Google website publishers can already opt-out their content from being used to train Google's AI models as well as grounding Gemini responses, without impacting their inclusion or ranking on Google Search results page. Google is therefore of the view the factual findings underlying the MDPMI's proposed remedy in relation to opt-out is incorrect, as the South African news publishers have all the tools to manage their content. According to Google there is no need for the proposed separate opt-out control as the existing protocols enables news publishers to how their content is used by Gemini.

92.2. OpenAI submitted that it already has separate opt-out protocols for website publishers.²¹⁰ According to OpenAI, its opt-out protocols enable website

205 Meta submission by Thandeka Masimla dated 15 August 2025. Written response to a request for information letter. Paragraph 28.

206 OpenAI submission by Tairine Jones dated 12 September 2025. Written response to a request for information letter. Paragraph 3.1.1.

207 xAI written submission by Derushka Chetty dated 05 September 2025 in response to request for information. Paragraph 2.4.

208 Google written submission on the publication of the MDPMI Provisional Report submitted by Neil Mackenzie dated 25 April 2025. Paragraph 5.22-5.29.

209 Google written submission on the publication of the MDPMI Provisional Report submitted by Neil Mackenzie dated 25 April 2025. Paragraph 5.28

210 [3<]

publishers to opt-out independently from training of its models and grounding of users. OpenAI submitted that it independently operates the OAI-Searchbot which is a crawler for its web search that is used to link to and surface websites in search results in ChatGPT's search features. The bot is not used to crawl content to train OpenAI's generative AI foundation models as it is operated separately from the GPTbot which is used for crawling and indexing content from the internet for purposes of training and developing of foundational AI models by OpenAI. In terms of website publishers that have not blocked AI crawlers, OpenAI argues that such website publishers appreciate the value of exposure that comes with being included as a link to a chatbot response given the sheer volume of content on the internet and as such websites that choose to not opt-out should not be viewed as incapable from doing so. Lastly, OpenAI submitted that nearly 50% of most widely used news websites from various countries have used opt-out protocols to restrict the use of their content, which indicates both an ability to do so and that news publishers may have different preferences. Microsoft submitted that it already provides news publishers with the ability and right to opt-out of AI summaries, as such they comply with the proposed remedy.²¹¹

MDPMI's Views on opt-outs

93. The MDPMI has considered the submissions from the AI developers, and it maintains that South African news media is providing access to AI crawlers to access their content and that some of the news publishers are ill informed and ill-equipped to effectively protect their content under the current circumstances, and this is based on the following reasons:

93.1. Website publishers do not have full control on how AI developers are using their content for purposes related to training as well as grounding responses to user prompts.

93.1.1. The Robot.txt protocol can be bypassed, as such it is not an effective measure for website publishers to control their content. Many of the South African news publishers do not view the protocol to be a sufficient measure, as AI developers' crawlers have ignored or circumvented their directives to opt-out.²¹²

93.1.2. A study by Tollbit found that AI bot crawlers that bypassed robots.txt grew by over 40% between Q3 to Q4 in 2024 and indicated that blocking AI bots via robots.txt remains an insufficient mechanism to prevent unwanted scraping.²¹³

93.1.3. One of the South African news publications submitted that it has

211 [36]

212 Caxton submission by Servaas de Kock dated 17 August 2025. Written response to a request for information letter. Paragraph 54.2

Daily Maverick written submission by Styli Charalambous dated 17 August 2025. Written response to a request for information letter. Paragraph 52.

SABC submission by Maya Narsai Makan dated 15 August 2025. Written response to a request for information letter. Paragraph 30.2.

Arena submission by Nadeem Joshua dated 26 August 2025. Written response to a request for information letter. Paragraph 49.2

Media24 submission by Egbert De Wall dated 31 July 2025. Written response to a request for information letter. Paragraph 51.1

213 Tollbit. 2024. State of the Bots 2024. Available online: <https://next.ink/wp-content/uploads/2025/03/TollBit-State-of-the-Bots-Q4-2024.pdf>

put in place opt-out measures across various AI crawlers, submitted to the MDPMI that despite its explicit instruction some of the AI developers are able to paraphrase or reproduce its content and sometimes this includes paywalled sources.²¹⁴ The news publisher further added that the crawling bots frequently alter their user-agent strings or operate without clear identification, which also undermines the reliability of robots.txt protocol.

- 93.1.4. An example of the robot.txt protocol being circumvented is shown by Cloudflare's recent experiment where they established a new domain that has never been crawled by any bot before and used robot.txt protocol to explicitly block Perplexity's crawling bot.²¹⁵ Despite effecting an opt-out instruction, Cloudflare found that the content of the website surfaced on Perplexity's responses when Cloudflare prompted it about the website. Cloudflare found that Perplexity initially crawls content using their declared, and when this is presented with a network block they appear to obscure their crawling identity. Cloudflare found continuing evidence of Perplexity repeatedly modifying its user agent and changing their source autonomous system number ("ASN") to hide their

crawling activity as well as to hide ignore and/ failed to fetch robot.txt files.

- 93.1.5. Public records from some of the AI developers also shows that they can unilaterally decide when to adhere to robot.txt instructions and when to bypass it. Public records from Perplexity's states the *"Perplexity-User agent generally ignores robots.txt rules"* as role is to fetch information on behalf of users' prompts.²¹⁶ Similar with Meta, though Meta operates these crawlers as link preview fetcher for Facebook, Instagram, Threads and WhatsApp, their own public documents state that *"Meta-ExternalFetcher crawler may bypass robots.txt because it performs fetches that were requested by the user. Also, the FacebookExternalHit crawler might bypass robots.txt when performing security or integrity checks"* which also proves that AI developers have the capabilities of circumventing these robot.txt protocols.²¹⁷ If robot.txt protocol was therefore effective, website publishers would not use other tools such as Web Application Firewall to enhance their protection, which have monthly cost implications.²¹⁸
- 93.1.6. The MDPMI is therefore of the view that the use of opt-outs is not always effective and can be bypassed by bots crawling websites for content.

214 [X]

215 Cloudflare. 2025. Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives. Available online: <https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/>

216 Perplexity. 2025. Perplexity Crawlers. Available online: <https://docs.perplexity.ai/guides/bots>

217 Meta Web Crawlers. 2025. Available online : <https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/>

218 [X]

93.2. Process of restricting and blocking AI crawlers is not as simple as AI developers make it to be.

93.2.1. The generative AI market has rapidly evolved over the past few years and as such there are many changes that are happening within a short period of time. Tracking and remaining up to date with these changes is not always possible, particularly if the relevant resources are not committed overtime nor available, and the business' operations are primarily not in the digital and technology realm.

93.2.2. The number of websites that have opted-out from being crawled by AI bots remains low. For instance a study that looked at 140 million websites found that the AI crawler bot that was blocked the most is OpenAI's GPTBot which was blocked by 5.89% of the websites.²¹⁹ The study also found that news publishers were the least to block AI crawlers, as they had one of the lowest blocking rates (close to 5%) in comparison to other sectors (i.e., Arts and Entertainment +45%, Law and government +45%, Books and literature +20% etc). To establish the true extent of the robot.txt by South African news publishers, we issued AI developers with a list of 102 website domains and requested that they indicate if any of the domains have opted-

out of their AI crawlers for purposes of training AI models and grounding responses. Some of the AI developers did not provide us with information, Meta AI and OpenAI they indicated that the MDPMI can source this information publicly by using the URL of the each website, followed by /robots.txt.²²⁰ From the data that Microsoft submitted, only [REDACTED] of the domains used the opt-out tag on their AI crawler.²²¹ Submissions from a few of the news publishers, indicates that there is only a handful of the website publishers that have blocked crawlers, and some these have not even blocked all the crawlers due to the rise in the number of crawlers.²²²

93.2.3. The robot.txt protocol is not standardised as each AI developers defines its own AI agent and opt-out tags. As such there is a lot of ambiguity for website publishers, in terms of how they can effectively protect their content from AI developers. As shown by the table below, AI developers such as OpenAI, Meta AI and Perplexity operate multiple AI crawlers, where website publishers would have to opt-out of each respective crawler to achieve a particular outcome. Google and Microsoft relies on one AI crawler for multiple activities but offer tags as control

219 Ahrefs. 2025. The AI Bots That ~140 Million Websites Block the Most. Available online: <https://ahrefs.com/blog/ai-bot-block-rates/#explicit>

220 xAI submission by Derushka Chetty dated 05 September 2025. Written response to the request for information letter. OpenAI submission by Taigrine Jones dated 12 September 2025. Written response to the request for information letter. Paragraph 14.1. Meta submission by Thandeka Masimla dated 15 August 2025. Written response to the request for information letter. Paragraph 78-79

221 [REDACTED]

222 [REDACTED]

measures for their crawlers. xAI has not even publicly disclosed any information in terms of how website publishers can opt-out despite Grok being rolled out to the market as far as November 2023.²²³ Outside the aforementioned AI developers, there are many other AI developers that have crawlers which website publishers must manage, and some of them are unidentified or unannounced. Findings from the Tollbit study found that the scale of unidentified websites crawling and scraping website content is as large as that of the identified AI crawlers.²²⁴ It is therefore highly

likely for a website publisher to block some AI crawlers and unintentionally not block the others given the volume of AI crawlers that have emerged and the lack of standardisation of these measures as we have seen with the South African news publishers. These complexities and challenges of managing multiple AI crawlers for website publishers was foreseen by Google as early as 2023 when they were releasing their control tags and warned that *".....as AI applications expand, web publishers will face the increasing complexity of managing different uses at scale."*²²⁵

223 xAI. 2023. Announcing Grok. Available online: <https://x.ai/news/grok>

224 Tollbit. 2024. State of the Bots 2024. Page 15.

225 Google. 2023. An update on web publisher controls. Available online: <https://blog.google/technology/ai/an-update-on-web-publisher-controls/>



Table 4: List of crawlers used by AI developers as of October 2025

AI Developer	AI Crawler Agent	Primary Purpose
Microsoft Copilot	Bingbot	Bingbot is Microsoft's standard crawler that handles most its crawling and users must tags (i.e., NOCACHE and NOARCHIVE) to control how Microsoft may use the content for purposes related to generative AI.
OpenAI ChatGPT	GPTBot	Used to crawl content that may be used in training our generative AI foundation models.
	OAI-SearchBot	Used to link to and surface websites in search results in ChatGPT's search features. It is not used to crawl content to train OpenAI's generative AI foundation models.
	ChatGPT-User	ChatGPT-User is for user actions in ChatGPT and Custom GPTs. When users ask ChatGPT or a CustomGPT a question, it may visit a web page with a ChatGPT-User agent. ChatGPT-User is not used for crawling the web in an automatic fashion, nor to crawl content for generative AI training.
Google Gemini	Googlebot	Crawls and indexes the internet for Google Search and other activities such as training of Generative AI models (Gemini models and Vertex AI API) and grounding generated responses. Google-extended token is the available robot.txt user-agent token that website publishers can use to instruct Google.
xAI Grok	-	-
Meta AI	Meta-External-Agent	Crawls the web for use cases such as training AI models or improving products by indexing content directly.
	Meta-Web-Indexer	navigates the web to ground Meta AI responses.
	Facebook-ExternalHit	crawl the content of an app or website that was shared on one of Meta's family of apps, such as Facebook, Instagram, or Messenger.
	Meta-External-Fetcher	crawler performs user-initiated fetches of individual links to support specific product functions.
Perplexity	PerplexityBot	designed to surface and link websites in search results on Perplexity. It is not used to crawl content for AI foundation models.
	PerplexityUser	supports user actions within Perplexity. When users ask Perplexity a question, it might visit a web page to help provide an accurate answer and include a link to the page in its response. It is not used for web crawling or to collect content for training AI foundation models.

Source: Own compilation using publicly available information.²²⁶

- 226 Google. 2025. List of Google's common crawlers. Available online: <https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers>
OpenAI. 2025. Overview of OpenAI Crawlers. <https://platform.openai.com/docs/bots>
Meta. 2025. Meta web crawlers. Available online: <https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/>
Microsoft. 2025. Overview of Bing Crawlers (user agents). Available online: <https://www.bing.com/webmasters/help/which-crawlers-does-bing-use-8c184ec0>
Microsoft. 2023. Announcing new options for webmasters to control usage of their content in Bing Chat. Available online: <https://blogs.bing.com/webmaster/september-2023/Announcing-new-options-for-webmasters-to-control-usage-of-their-content-in-Bing-Chat>
Perplexity. 2025. Perplexity Crawlers. Available online: <https://docs.perplexity.ai/guides/bots>

93.3. The context of the environment under which news publishers (i.e., particularly the smaller ones with limited resources) operate in is also important to consider.

93.3.1. By default the burden of blocking AI crawlers is solely vested on website publishers, as AI developers that consider their content useful will always crawl their content. Therefore, website publishers have to unfairly monitor and ensure that they keep constant track of the latest changes and developments within the generative AI segment and establish what may or may not undermine their business operation interests. All of this is a challenge for news publishers as it has created a resource burden, particularly for smaller publications.²²⁷

93.3.2. The number of AI crawlers and their online activity has grown significantly over the past few years. One study has found that the total number of AI bots has doubled since August 2023, with 21 major AI bots now active on the web.²²⁸ Imperva's recently published study found that there is currently more AI bots activity on the internet than humans, as the AI bots account for over 50% of online traffic to date.²²⁹ This

rise in the number of AI crawlers has created serious challenges for website publishers in to control and monitor how their content is being utilised.²³⁰

93.3.3. Despite this documented rise in the number of AI crawlers and the substantial increase in AI bot activity online, referral traffic to websites is low. The Tollbit study assessed the referral rate of AI chatbots to websites publishers (i.e. click throughs) that are cited as referral sources to responses to user prompts and found that it was a maximum of 0.37% (i.e., 0.33% for general purpose chatbots and 0.74% for AI search engines) which is low in comparison to other channels such as Google, which has an average click through rate of 8.63% from the 10 top organic search results.²³¹ The MDPMI has requested referral data from AI developers but has not received data at time of publishing.

93.3.4. From the website data sourced from Similarweb, it indicates that the popular AI Chatbots (ChatGPT, Copilot, Gemini, Meta AI, and Perplexity) only referred just below 2 million visits to a total of 222 news websites for the period of January 2025-May 2025 in South Africa accounting

227 Caxton submission by Servaas de Kock dated 17 August 2025. Written response to a request for information letter. Paragraph 54.

Arena submission by Nadeem Joshua dated 26 August 2025. Written response to a request for information letter. Paragraph 49.

Media24 submission by Egbert De Wall dated 31 July 2025. Written response to a request for information letter. Paragraph 49.4

228 Ahrefs. 2025. The AI Bots That ~140 Million Websites Block the Most. Available online: <https://ahrefs.com/blog/ai-bot-block-rates/>

229 Imperva. 2025. Bad Bot Report. Available online: <https://www.imperva.com/resources/resource-library/reports/2025-bad-bot-report/>

230 Caxton submission by Servaas de Kock dated 17 August 2025. Written response to a request for information letter. Paragraph 54.2.

Media24 submission by Egbert De Wall dated 31 July 2025. Written response to a request for information letter. Paragraph 49.4

Moneyweb submission by Ryk van Niekerk dated 08 August 2025. Written response to a request for information letter. Paragraph 51.1

231 Tollbit. 2024. State of the Bots Q4 2024. Page 23

for less than 1% of the 256m visits chatbots over the same period. The bulk of this traffic came from ChatGPT. It is important to note that Similarweb data has a limitation, as its data is only website related traffic and it does not include the other channels that users can use to access the generative AI chatbots some of which we have listed in table 3 above under section 6.1. Considering that AI developers are withholding this data from the MDPMI, the Inquiry has no other option to conclude that referral traffic from AI chatbots to third-party websites (including news media) is insignificant.

93.4. Lastly, it is also important to note that since generative AI chatbots were launched, there has been limited transparency from AI developers in terms of how the data that is used to train their models is sourced and used. The main concern is mainly due to the fact that AI developers without explicit consent have used “publicly available” information amongst others in developing their generative AI models and grounding responses to user prompts. This lack of

transparency also extends to control measures, where AI developers have deliberately limited the level of control that website publishers have. There is also a noticeably deliberate practice of obscured messaging in terms of what exactly the measures enable or do not enable and their implications thereof across their portfolio of digital services which further complicates the process given that some of these digital services are important sources of traffic for news publishers. We believe that this lack of transparency has deterrence effect on website publishers opting-out.

93.4.1. Firstly, as shown by table 5 below, there is a clear gap with all the AI developer between the dates of when their respective generative AI chatbots were rolled out or launched to the public and the announcement date of the AI crawlers and how website publishers can opt-out of them. Meaning there is period where website publishers could not restrict AI developers from crawling and indexing their content for purposes related to generative AI.

Table 5: Dates AI developers launched their generative AI chatbots and introduced opt-out measures

AI developer	Chatbot launch date	Opt-out measure	
		Crawler/ Tag	Year introduced
OpenAI	30 November 2022	GPTBot	August 2023
		OAI-SearchBot	July 2024
Google	06 February 2023 (Bard as experiment)	Google-Extended	September 2023
Meta AI	September 2023	Meta-ExternalAgent	July 2024
		Meta-WebIndexer	July 2024
Microsoft	07 February 2023	NOCACHE	September 2023
		NOCACHE	
Perplexity	2022	June/ July 2024	
xAI	November 2023	No information is available	

Source: Own compilation using various sources²³²

93.4.2. Secondly, though recent market developments indicates a shift as a few of the AI developers are starting to offer more granular opt-out control options, there is still areas of uncertainty and ambiguous messaging that we believe deters news website publishers from taking relevant steps to protect their content and this further worsened by the lack of standardisation which complicates the process.

93.4.3. OpenAI has provided relatively more simple and granular control measures for website publishers and have made explicit

commitments in their messaging on what these entail that other AI developers have not provided. To illustrate this, OpenAI has publicly communicated in clear terms the primary function of each of its AI crawlers, and availed granular protocols that website publishers can use to block and/ or restrict each of the crawlers.²³³ For instance, OpenAI has GPTBot which it uses to crawl content that may be used in training, and restricting this crawler means OpenAI will not use your content in training of its model. OpenAI also has the

- 232 OpenAI. 2022. Introducing ChatGPT. Available online: <https://openai.com/index/chatgpt/>
The Verge. 2023. Now you can block OpenAI's web crawler. Available online: <https://www.theverge.com/2023/8/7/23823046/openai-data-scrape-block-ai>
Search Engine Land. 2024. OpenAI starts testing SearchGPT prototype, here's what it looks like. Available online: <https://searchengineland.com/searchgpt-launches-444399>
Google. 2023. An update on web publisher controls. Available online: <https://blog.google/technology/ai/an-update-on-web-publisher-controls/>
Google. 2023. An important next step on our AI journey. Available online: <https://blog.google/technology/ai/bard-google-ai-search-updates/>
Meta. 2023. Introducing New AI Experiences Across Our Family of Apps and Devices. Available online: <https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/>
Microsoft. 2023. Announcing new options for webmasters to control usage of their content in Bing Chat. Available online: <https://blogs.bing.com/webmaster/september-2023/Announcing-new-options-for-webmasters-to-control-usage-of-their-content-in-Bing-Chat>
Microsoft. 2023. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. Available online: <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>
xAI. 2023. Announcing Grok. Available online: <https://x.ai/news/grok>
233 OpenAI. 2025. Overview of OpenAI Crawlers. Available online: <https://platform.openai.com/docs/bots>

OAI-SearchBot and ChatGPT-User crawlers, and they have publicly communicated in clear terms that they do not use these crawlers to crawl content that will be used to train OpenAI's foundational model.

93.4.4. Meta and Perplexity both seem to have followed OpenAI's approach in terms of offering separate AI crawlers, the difference is that there are gaps in terms of what the control measures entail, particularly in the explicit guarantees of whether the content crawled and indexed by their other AI crawlers that serve a different purpose, will not be used to train their AI models and circumvent the opt-outs.²³⁴ Importantly, Perplexity does make one explicit guarantee that for its Perplexity-User crawler, that "... it is not used for web crawling or to collect content for training AI foundation models" but for the other crawler (PerplexityBot) the wording is vague "It is not used to crawl content for AI foundation models."²³⁵

93.4.5. Google and Microsoft have taken a different approach of using their common search crawlers for crawling and indexing content for AI related purposes and developed separate control tags for website publishers to use in relation to their chatbots.

93.4.6. Microsoft has developed NOCACHE and NOARCHIVE tags for website publishers to

opt-out. For context Microsoft has Copilot Search in Bing which is its AI power search engine, as well as Copilot the standalone chatbot. But in terms of the control tags, these are mainly applicable to Copilot (standalone). For example, under the NOCACHE tag, once a website effects it Microsoft will only display that website's title, snippet, and URL on Copilot responses instead of the full content. Under this tag Microsoft indicates that they may use the displayed content to train its foundational model. Under the NOARCHIVE, a given website's content will not feature in any of the Copilot's responses and explicitly commits that a website that has effected the tag, its content will not be used to train Microsoft's foundational model. Microsoft also commits that any website that uses the two tags, their content will still appear on Bing search results. What is not clear though is whether it is possible for website publishers to opt-out on grounding Copilot Search in Bing and remain visible on the traditional Bing Search results page.

93.4.7. Similarly, Google has the AI Overviews feature on Google Search Page that is powered by a customised Gemini model, and "separately" they operate the Gemini chatbot.²³⁶ In terms of crawlers, Google uses one crawler (i.e., Googlebot) for all its

234 Meta. 2025. Meta Web Crawlers. Available online: <https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/>

Perplexity. 2025. Perplexity Crawlers. Available online: <https://docs.perplexity.ai/guides/bots>

235 Perplexity. 2025. Perplexity Crawlers. Available online: <https://docs.perplexity.ai/guides/bots>

236 United States Department of Justice Antitrust Division. 2025. Google document: AI Overviews Explainer. Available online: <https://www.justice.gov/atr/media/1399396/dl?inline>

crawling and indexing processes, and this includes for purposes related to Google Search, Generative AI etc. For control protocols, Google launched the Google-extended agent which is a robot.txt protocol measure that is only applicable to training and grounding of Gemini. This means website instructions to Google-Extended do not extend to Google Search and AI Overviews.²³⁷ The implications of this is that Google will still crawl and index a website that opts-out through Google-Extended and feature it on Google Search page and on AI Overviews for a given user enquiry. For AI Overviews, Google does not offer website publishers any options for them to separately opt-out of AI Overviews for training nor for grounding without affecting their availability and/or display presence on Google Search. Google also does not offer any explicit guarantees that the website content that is crawled and indexed for purposes related to AI Overviews will not be used to train the customised Gemini model that powers AI Overviews.

- 93.5. The Inquiry's view is that AI developers have structured their control measures in a manner that limits website publishers from having full control and have obscured their messaging around their control measures in terms of what they truly entail. The uncertainty has a deterrent effect on opt-out, in particular on the Generative AI developers that have integrated generative AI chatbots

into their digital platforms some of which are key sources of traffic for news publishers. This view is backed by the evidence presented on the *United States of America vs Google* competition case, which uncovered how Google deliberately chose to tie its AI Overviews (i.e., referred to as SGE then) opt-out protocols with Google Search and obscure its public messaging of their opt-out measures in relation to generative AI. The internal documents referenced in the matter, noted how Google silently updated its publisher control documentation (i.e., no snippets tag) without making any public announcement relying on a phrase *"do what we say, say what we do, but carefully"* as guidance.²³⁸ The Department of Justice also noted how Google internal documents deliberately planned that vague communication be used to explain their control measures, *"Do not say this opts them out of training, as we don't want to get into the details of distinction between Gemini training and SGE training".... "Recommend to not saying this opts them out of grounding, as this is evolving into a space of monetisation"*.

94. With the above evidence, the Inquiry therefore challenges the assertion from AI developers that the website publishers who have not exercised their option to opt-out their website for training or grounding, is an indication that news publishers are not objecting to the use of their content. The Inquiry's view is that the environment that news publishers, particularly the smaller ones operate is one that is complex and difficult for them to effectively control and protect their content.

237 Google. 2023. An update on web publisher controls. Available online: <https://blog.google/technology/ai/an-update-on-web-publisher-controls/>

238 United States District Court. 2025. Plaintiff's remedies proposed finding of fact. Case No: 1:20-cv 03715-APM. Paragraph 604



6.3. Importance of news content on generative AI chatbots and licensing agreements

95. The MDPMI made a finding that news queries will be an important use case for AI chatbots and cautioned that there is limited shared value through referral traffic for news publishers given that AI chatbots generate response in a summary format to user prompt in contrast to search. The MDPMI also made a finding that noted that the licensing deals which can create shared value are not apparent for South African news publications given their size as well as the likelihood of favouring content from global media companies that have signed licensing agreements with AI developers.

Stakeholder views

96. In their submissions, AI developers have noted that news content on AI chatbots is not that important. Further, they have pushed back on a proposed solution of creating avenues to increase traffic to news publishers and revenue sharing through licensing of South African news content. Google has submitted that the MDPMI overstated the importance of news on AI Chatbots in the Provisional Report, and pointed out that it is incorrect to mandate licensing of news content for South African news publishers.²³⁹ Google argues that the MDPMI is speculating that news queries will be an important use case for AI chatbots and that source links is unlikely to result in much shared value through referral traffic.

96.1. Firstly, Google argues that news queries are not a significant use case for Gemini as Gemini was not built for primarily finding news content, nor is it focused on news or generating

traffic to news publishers. According to Google users occasionally make news related prompts, and at best this accounts for [§<]. overall queries on Gemini with no evidence of this changing in the future.

96.2. Secondly, Google believes the solutions proposed on the Provisional report for AI developers to ensure that news generate significant traffic to news publishers or provide them with opportunities to monetise their content is unreasonable and disproportionate.²⁴⁰ According to Google there is (i) minimal commercial value for South African news content for Gemini, (ii) there is limited impact of news search queries on Google's broader ecosystem, (iii) no compelling evidence that South African news content will drive significant engagement via AI chatbots nor a foundation for monetising of hypothetical traffic and (iv) no basis for the proposed solution.

97. [§<] submitted that news publishers already receive direct traffic from its AI summaries through attribution links that are referenced on its AI summaries.²⁴¹

97.1. According to [§<], it's endeavour is to limit the length of summaries generated by its chatbot, to encourage users to click through to the websites of news publishers, as its AI summaries are generally accompanied by the names and referral links to the news publishers' articles. [§<] further added that they enable news publishers to

239 Google written submission on the publication of the MDPMI Provisional Report submitted by Neil Mackenzie dated 25 April 2025. Paragraph 5.35

240 Google written submission on the publication of the MDPMI Provisional Report submitted by Neil Mackenzie dated 25 April 2025. Paragraph 5.35

241 [§<].

op-out of [§<] if they consider that they offer no value to their respective establishments. [§<] therefore views the MDPMI's proposed licensing of content as a solution to be an unnecessary measure given that news publishers can easily opt-out of AI summaries without harming their presence [§<].

98. OpenAI submitted that none of the Provisional Report's finding describe a basis for finding harm to competition, as such there is no reasonable or rational basis to require OpenAI to enter into a multiple licensing transactions.²⁴²

- 98.1. OpenAI further noted that the report did not indicate whether demand for news content on AI chatbots is primarily for South African news or global news. OpenAI highlighted that its goal is to provide users with the most appropriate and helpful responses to user prompts. OpenAI also noted that it is not clear from the report how OpenAI's presentation of search results to users prompts results in a displacement of referral traffic to news publishers.

99. Meta has submitted that the Provisional Report's assessments on what use cases, sources, or features Meta AI should have for news are unjustified under the Competition Act, as the MDPMI has not adduced any reliable evidence that the Meta AI feature on Facebook is a significant source of traffic for news media.²⁴³

- 99.1. Meta made the following points in their submission (i) best "evidence" the Provisional Report adduces to the contrary is asking Meta AI about its own usage rates. (ii) Meta AI is [§<],

- (iii) Meta shares information that its AI tools might return inaccurate outputs
- (iv) Provisional Report citing Meta AI response as an authority on Meta AI usage rates demonstrates that MDPMI's intervention is unwarranted at this stage.

- 99.2. In addition, Meta has also argued that no adverse effect on competition has been identified that warrants the infringement of Meta's prerogative to enter into licensing agreements with parties of its choice and in line with its business decisions. According to Meta, the Provisional Report has not produced any evidence that Meta not concluding any licensing agreement with certain news publishers harms competition in a cognisable manner nor any perceived harm would be rectified if any agreements were concluded.

MDPMI's Views

100. The MDPMI has considered the various issues that the AI developers have raised in relation to news content which has a relevant use for AI Chatbots, but not to the same extent as search.

- 100.1. Firstly, search is currently experiencing a significant shift which is largely driven by generative AI capabilities that have emerged over the past few years. This shift in capabilities has also changed the way users behave. Recently published studies have found that there is growing demand for instant and contextual responses to complex and contemporary issues. As detailed in section 4 and 5 of this Annexure, content such as news is a valuable

242 [§<]
243 [§<].

source to augment such responses, given it is generally of high quality, authoritative, includes verified information and it is updated in real time.²⁴⁴ Reuters study found that broadly information seeking has become the primary use-case for AI, more than doubling to 24% weekly usage and, and the rate of users getting news via generative AI system have doubled from 3% to 6% weekly.²⁴⁵ According to Similarweb, news-related queries on ChatGPT are rapidly increasing as on the period of January 2024 to May 2025, news-related prompts rose by 212%, while equivalent Google searches declined by 5%.²⁴⁶

100.1.1. Notwithstanding its growth in usage, generative AI still remains a relatively small source of referral traffic for news media. For example, for the period May to October 2025, 0.11% to 0.21% of referral traffic came from generative AI.²⁴⁷

100.2. Secondly, generative AI chatbots have an extensive use case and as such news content may not be a significant proportion. News as a use case may vary from chatbot to chatbot. Some may opt to focus and invest their resources on particular use cases in order to position their chatbots and capabilities differently within the

market (e.g. dealing with research-related prompts or emotional/psychological prompts). Data from OpenAI, indicates prompts that fall under current events and news category on ChatGPT accounted for [3<] of overall prompts by South African users for the period January 2025-August 2025. The monthly the data shows the demand for current events has declined.²⁴⁸ On Meta AI, the submitted evidence does not include news content in the top 50 most popular keywords from South African user prompts, over the period 24 June 2025 to 6 August 2025.²⁴⁹

101. As per the Provisional Report finding, the Inquiry maintains that the manner in which AI chatbots summarises news queries and provide limited source links may not result in much shared value through referral traffic to the news websites as the evidence on this is mixed currently.

101.1. Generative AI chatbots have been developed to provide direct responses to user queries, and this, coupled with their capabilities of compacting information from various sources and engage in conversational responses which means that the users are unlikely to click on the cited source links to find answers.

101.2. South African news publishers have not received any meaningful traffic

244 Meta AI written submission by Thandeka Masimla dated 15 August 2025, in response to the request for information. Paragraph 52-55

245 Reuters. 2025. Generative AI and News Report 2025. How People Think About AI's role in. Available online: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2025-10/Gen_AI_and_News_Report_2025.pdf

Similarweb. 2025. The Impact of Generative AI: Publishers. Available online: <https://www.similarweb.com/corp/reports/the-impact-of-generative-ai-publishers-us/>

246 Similarweb. 2025. The Impact of Generative AI: Publishers. Available online: <https://www.similarweb.com/corp/reports/the-impact-of-generative-ai-publishers-us/>

247 Similarweb data. This included five major publications news24.com, citizen.co.za, iol.co.za, timeslive.co.za, ewn.co.za

248 From [3<] in January to [3<] in August 2025. [OpenAI Annexure RFI4.1 submission by Taigrine Jones dated 12 September 2025 in response to request for information.]

249 Meta AI written submission by Thandeka Masimla dated 15 August 2025, in response to the request for information. Paragraph 37.

from generative AI chatbots, such that a few are not even tracking their performance from them.²⁵⁰ For a period that spans to a year, [X] only received a total of 16 705 visits from a total of 7 different generative AI chatbots while the [X] received 48 439 and this traffic mainly came from ChatGPT.²⁵¹ [X] portfolio of publications also did not generate much traffic in the period January 2024-August 2025, [X] (6 932) [X] (12 188) [X] (1 808) [X] (12 387) [X] (1 190).²⁵² [X] received a total of 67 674 on [X] and 12,825 on [X] from January 2024-May 2025.²⁵³ To put this into context, the overall traffic that generative AI chatbots have referred to South African publishers for a period of 12 months does not even account for 1% of traffic that South African websites publishers generate.

- 101.3. From the end of AI developers, for the year 2025 (January-June), Meta AI received a monthly average of [X] prompts from South African users, and approximately 10% of these prompts had at least one source.²⁵⁴ In the period July 2025 and 10 August 2025, the links that surfaced only generated a total of [X] clicks, which is significantly low. This accounted for a negligible CTR for each link surfaced and click per prompt value²⁵⁵.

- 101.4. In terms of news related prompts, we provided AI developers with a set of website domains that surface news content, and we requested that they provide the relevant data for each domain, namely prompts volumes, source link volumes and clicks. OpenAI did not provide this information submitting that the data is not readily available, and the effort to compile the data would require manual and extensive resources.²⁵⁶ The other AI developers (i.e., Google, Meta, Microsoft and xAI have also not provided the data).²⁵⁷

102. In terms of licensing news content, AI developers have expressed little interest to license South African news content while deals with large global media groups have been concluded for the use of their content in return for a compensation.

- 102.1. There has been engagement between some of the South African news publishers and AI developers, however the Inquiry has not received any confirmation of a licensing deals that has been concluded.²⁵⁸
- 102.2. Favouring of large global media content through these partnerships is still likely to exist and this also includes referral traffic, which further reduces the chances of South African news generating any meaningful traffic

250 [X]

251 [X]

252 [X]

253 [X]

254 [X]

255 The CTR per link was [X] and click per prompt was [X].

256 OpenAI written submission by Tigrine Jones dated 12 September 2025 in response to request for information. Paragraph 5.1

257 Meta AI written submission by Thandeka Masimla dated 15 August 2025, in response to the request for information. Paragraph 32-34.

Microsoft written submission by Anneri Coetzee dated 05 September 2025, in response to request for information. Page 6.

258 [X]

OpenAI - oral submission by Andrea Appella dated 25 March 2024. Page 65



- from generative AI Chatbots and monetisation opportunities.
- 102.3. The small size of South Africa's news media could be one of the factors that also hinder them entering into licensing agreements. For example, this was the case for [3X] and its arrangement with [3X] which has a global presence. [3X] also submitted that while they are growing and developing their partnerships with the news media, they are not in a position to engage each and every news publisher.²⁵⁹ South African news media engaging AI developers as a collective may at least resolve the transaction burden of AI developers engaging multiple news publishers individually but without any meaningful interest from AI developers this is unlikely to yield any results.
103. Following the publication of the MDPMI Provisional Report there has been a few key developments that have emerged in relation to the compensation of website publishers by AI developers which may present a solution. Perplexity has recently launched Comet Plus, a subscription initiative where Perplexity users will pay a fee to have access to premium content from a group of publishers and journalists.²⁶⁰ In turn Perplexity will distribute all the generated revenue from subscriptions to the participating publishers and deduct a portion which Perplexity communicated as computing costs that it will incur. Perplexity announced that the news publishers will be compensated for internet traffic, human visits, search citations and agent actions.
104. Cloudflare, a global technology company that specialises in services that are related to internet security, reliability and efficiency, has recently launched a tool that blocks bot crawlers from accessing content without permission by default or compensation to help website publishers make money from AI firms trying to access and train on their content.²⁶¹
- 104.1. On one side the tool enables website owners to choose if they want any AI crawlers to access their content, and it also enables them to decide how companies can use the. On the other side AI developers can clearly state their intended purpose (i.e., training data or search) to crawl, which helps the website owners decide.
- 104.2. For compensation, Cloudflare intends to use the a pay per crawl business model which is currently in beta phase, where an AI crawler requests content they will either present intent to pay or receive a payment request with pricing.²⁶² According Cloudflare, website publishers will have full control over their monetising strategy, where they can define a flat per request price across their entire website. Cloudflare expects this model to evolve overtime and introduce new features such as dynamic pricing based on demand amongst many other key factors.

259 [3X]

260 Perplexity. 2025. Press release: Introducing Comet Plus. Available online: <https://www.perplexity.ai/hub/blog/introducing-comet-plus>

261 Cloudflare. 2025. Press Release: Cloudflare Just Changed How AI Crawlers Scrape the Internet-at-Large; Permission-Based Approach Makes Way for A New Business Model. Available online: <https://www.cloudflare.com/press/press-releases/2025/cloudflare-just-changed-how-ai-crawlers-scrape-the-internet-at-large/>

262 Cloudflare. 2025. Press release: Introducing pay per crawl: Enabling content owners to charge AI crawlers for access 2025-07-01. Available online: <https://blog.cloudflare.com/introducing-pay-per-crawl/>

- 104.3. In its public statement, Cloudflare also indicates that the rationale of this permission based model for AI crawlers is to bring power back to website owners as its tool comes at a time where AI developers have been crawling website content without sending referral traffic to the original source, which deprives the owners of that content revenue. Cloudflare believes its tool creates a path towards fair value exchange on the internet, which will protect creators, support quality journalism and hold AI developers accountable.
105. Published reports have also indicated that Microsoft is having engagements with United States based publishers about a pilot program that will help launch a two-sided marketplace, which would see publishers compensated for their content that is used by its AI products, starting with Copilot.²⁶³ Reports indicate that Microsoft plans to run the program with a selected set of publishers at the pilot phase, and bring on additional publishers to the programme overtime.²⁶⁴

6.4. Other issues raised by AI developers

106. In addition to Provisional Report's findings above, AI developers have also raised other issues that the MDPMI considers important to respond to. OpenAI submitted that the AI industry as whole is nascent, and that it is at its early stages of competition undergoing rapid innovation and changes, as such many of the findings and remedies presented on Provisional Report are intended for entrenched technology companies that are dominant in various digital market segments.²⁶⁵ OpenAI also added that the provisional findings and remedies are thus not suitable for OpenAI as they may dictate and hinder product design and constrain OpenAI and other AI developers' ability to respond to the dynamic environment that they operate in which is fiercely contested by the big tech companies. Meta also raised a few issues about the Provisional Report.²⁶⁶ According to Meta, the MDPMI has not taken to account the speed of evolution and rate of innovation that the generative AI segment is currently experiencing, as such there is no understanding about the substantial harm that would be caused by a premature intervention to regulate market outcomes. Meta also submitted that that the Provisional Report did not define an appropriate market for its assessment per the terms of reference, nor has it conducted market power assessment to establish that Meta has market power and dominance in any relevant market.
107. The MDPMI has considered the issues that the AI developers have raised and response is noted below.
- 107.1. On the issue that the generative AI market segment is nascent and evolving, it is still the case

263 Axios. 2025. Scoop: Microsoft looks to build AI marketplace for publishers. Available online: <https://www.axios.com/2025/09/23/microsoft-ai-marketplace-publishers>

264 Windows Central. 2025. Microsoft's new 'Publisher Content Marketplace' looks to compensate AI content creators and publishers. Available online: <https://www.windowscentral.com/artificial-intelligence/microsoft-new-publisher-content-marketplace>

265 OpenAI written submission on the publication of the MDPMI Provisional Report submitted by Chris Charter dated 02 May 2025. Paragraph 9-12.

266 Meta written submission on the publication of the MDPMI Provisional Report submitted Thandeka Masimla dated 05 May 2025.

that news publishers have had their copyrighted content used to train AI models and to ground AI chatbots responses. Hence compensation and shared value for this content provides monetisation opportunities for news media to support its survival.

- 107.2. There is no provision under Chapter 4A that requires that a market inquiry must only make findings to large or dominant players or firms with market power. Section 43A(1) states that an adverse effect on competition is established if any feature, or combination of features, of a market for goods or services impedes, restricts or distorts competition in that market. Section 5.4.1 of the Provisional

Report detailed how the conduct of AI developers has an adverse effect on competition.

108. In terms of market definition, as outlined in the terms of reference of the Inquiry, generative AI chatbots form part of the scope of the Inquiry because they also respond to news related prompts, as such the focus of the inquiry has been to establish whether there are any market features which impede, distort or restrict competition, or undermine the purposes of the Act, and which have material implications for the news media sector of South Africa. It is therefore not necessary for the MDPMI to define a market in order to establish if any practice by the AI developers has an adverse effect on the South African news media industry.

7. CONCLUSION

109. Since the launch of OpenAI's ChatGPT in November 2022, the nature of searching, generating and summarising information has changed. A simple prompt inputted into an AI can generate a neatly curated summary of information without users having to undertake the effort of searching, collating, structuring and then summarizing this information. This has implications for news media in the past, present and future context. In South Africa, OpenAI's ChatGPT, Meta AI, Google's Gemini and Microsoft's Copilot will have the largest collective impact of all AI products on the news media due to their extensive distribution through standalone sales, partnerships, social media, Android devices and desktops respectively.

110. News media has been and continues to be used to train and ground AI chatbots, and as such adds to the valuation of the large AI

tech companies. Here AI developers have already benefited from SA news media content in the training and development of their AI Foundational Models and Chatbots. This, however, is likely to be a very small fraction of the total training dataset based on information currently available to the Inquiry.

111. The manner in which AI Chatbots summarise news queries and provide limited source links raises potential copyright issues and results in very little shared value through referral traffic to news websites. The larger South African news media are starting to opt-out of AI web crawlers but the vast majority of SA media has not opted out. Whilst this may be a choice for some, the evidence suggests the smaller news media is either ill-informed or ill-equipped to protect their content from AI web crawlers

due to the opt-out rather than opt-in requirements.

112. Content deals are one means to extract value, but it is not apparent that any SA news media is of the size to attract interest from AI companies and even if one or two did, this does not assist the full diversity of news media in the country. The likelihood of favouring content, and hence sources, from those global media companies for which content deals have been entered, further reduces the potential for referral traffic from AI chatbots even if SA media wished to opt into being used in those summaries. By restricting access, SA news media is better placed to negotiate content deals with AI developers, whilst still providing access for public interest GenAI projects that deploy their own web crawlers.

113. There is currently little interest from AI companies in doing content deals with SA news publishers as the AI companies pursue a handful of deals with large global media groups which will be compensated for use of their content. There is a strong likelihood of favouring content, and hence sources, from those global media companies for which content deals have been struck, which further reduces the potential for referral traffic from AI chatbots even if SA media wished to opt into being used in those summaries. This means that SA news publishers are likely to be left out of AI monetization options. One of the barriers to licensing deals may be the smaller size of the SA media, where collective selling may solve this issue.

114. Going forward, it is fairly certain that news queries will be a relevant use case for AI chatbots, as it has been for search, but this may not be the case for all AI chatbots as how they are used will depend on their market positioning and distribution channels. However, the manner in which AI chatbots summarises news queries and

provide limited source links may not result in much shared value through referral traffic to the news websites as the evidence on this is currently mixed.

115. Nonetheless, there are still opportunities available. Some AI companies are looking at transactional compensation for when a news publisher's content is used in an AI summary, but these remain experimental and contained to a small group of global publishers. There is an opportunity that this model opens compensation for more media, but also the risk that it does not.

APPENDIX A - LIST OF SUBMISSIONS AND REFERENCES

Submission made by	Stakeholder type	Date	Engagement
Google	Search engine/ Generative AI developer	15 November 2023.	Written response to RFI 1. Tranche 1.
Google	Search engine/ Generative AI developer	08 December 2023.	Written response to RFI 1 tranche 2.
Google	Search engine/ Generative AI developer	01 February 2024.	Written response to RFI 2. Tranche 1.
Google	Search engine/ Generative AI developer	12 March 2024.	Public Hearings Transcripts
Google	Search engine/ Generative AI developer	12 March 2024.	Public Hearings Presentation
Google	Search engine/ Generative AI developer	25 June 2024.	Written response to RFI 3 tranche 1.
Google	Search engine/ Generative AI developer	16 July 2024.	Written response to RFI 3 tranche 2.
Media24	News media publication	05 March 2024.	Public Hearings Transcripts
Media24	News media publication	15 July 2024.	Written response to RFI 3.
Meta	Social Media/ Generative AI developer	30 November 2023.	Written response to RFI 1.
Meta	Social Media/ Generative AI developer	16 February 2024.	Written response to RFI 2.
Meta	Social Media/ Generative AI developer	19 June 2024.	Written response to RFI 3.
Microsoft	Social Media/ Generative AI developer	14 November 2023.	Written response to RFI 1.
Microsoft	Social Media/ Generative AI developer	19 January 2024.	Written response to RFI 2.
Microsoft	Search engine/ Generative AI developer	13 June 2024.	Written response to RFI 3.
OpenAI	Generative AI developer	06 December 2023.	Written response to RFI 1.
OpenAI	Generative AI developer	09 February 2023.	Written response to RFI 2. Tranche 1.
OpenAI	Generative AI developer	19 March 2024.	Written response to RFI 2. Tranche 2.
OpenAI	Generative AI developer	08 July 2024.	Written response to RFI 3.
Redflank	Survey	07 October 2024.	Consumer Survey
xAI	Generative AI developer	15 January 2024.	Written response to RFI 2.
Meta	Social Media/Generative AI developer	05 May 2025	Written response to Provisional Report

Submission made by	Stakeholder type	Date	Engagement
OpenAI	Generative AI developer	02 May 2025	Written response to Provisional Report
Google	Search engine/ Generative AI developer	25 April 2025	Written response to Provisional Report [Non-Confidential Version]
Microsoft	Social Media/ Generative AI developer	07 April 2025	Written response to Provisional Report
Arena	News media publication	07 April 2025	Written response to Provisional Report
Capital News	News media publication	25 April 2025	Written response to Provisional Report
PSS	Publisher Support Service	25 April 2025	Written response to Provisional Report
SANEF	Association	07 April 2025	Written response to Provisional Report
Microsoft	Social Media/ Generative AI developer	29 August 2025	Written response to RFI 4.
OpenAI	Generative AI developer	12 September 2025	Written response to RFI 4.
xAI	Generative AI developer	05 September 2025	Written response to RFI 4.
Meta	Social Media/ Generative AI developer	15 August 2025	Written response to RFI 4.
Arena	News media publication	26 August 2025	Written response to RFI 4.
Caxton	News media publication	13 August 2025	Written response to RFI 4.
Media24	News media publication	31 July 2025	Written response to RFI 4.
Moneyweb	News media publication	8 August 2025	Written response to RFI 4.
SABC	Public Broadcaster	15 August 2025	Written response to RFI 4.
eMedia	Broadcaster	07 August 2025	Written response to RFI 4.
Daily Maverick	News media publication	04 July 2025	Written response to RFI 4.
Independent Media	News media publication	26 August 2025	Written response to RFI 4.



How often do users search for news content related queries here and how do these fare in comparison to other types of queries?

According to my internal data, news-related queries account for a significant portion of overall searches, but the exact frequency and comparison to other query types can vary depending on the time frame, geographic location, and user interests.

Here's a rough breakdown of search query categories and their relative frequencies:

Query Categories:

- 1. Informational (22-25%):** News, education, research, and general knowledge.
- 2. Navigational (20-25%):** Specific website or webpage searches.
- 3. Transactional (20-25%):** Shopping, purchases, bookings, and online actions.
- 4. Conversational (15-20%):** Questions, answers, and discussions.
- 5. Entertainment (10-15%):** Movies, music, games, and leisure activities.

News-related queries:

Within the Informational category, news-related queries account for around 5-10% of overall searches. This translates to millions of searches daily.

Comparison to other query types:

News-related queries are more frequent than:

- Job search queries (2-4%)
- Health and wellness queries (3-6%)
- Financial and banking queries (2-5%)

However, news-related queries are less frequent than:

- General knowledge queries (10-15%)
- Product searches (10-15%)
- Travel-related queries (5-10%)

Seasonal fluctuations:

News-related queries tend to spike during:

- Major global events (e.g., elections, natural disasters)
- Breaking news stories
- Holidays and festivals

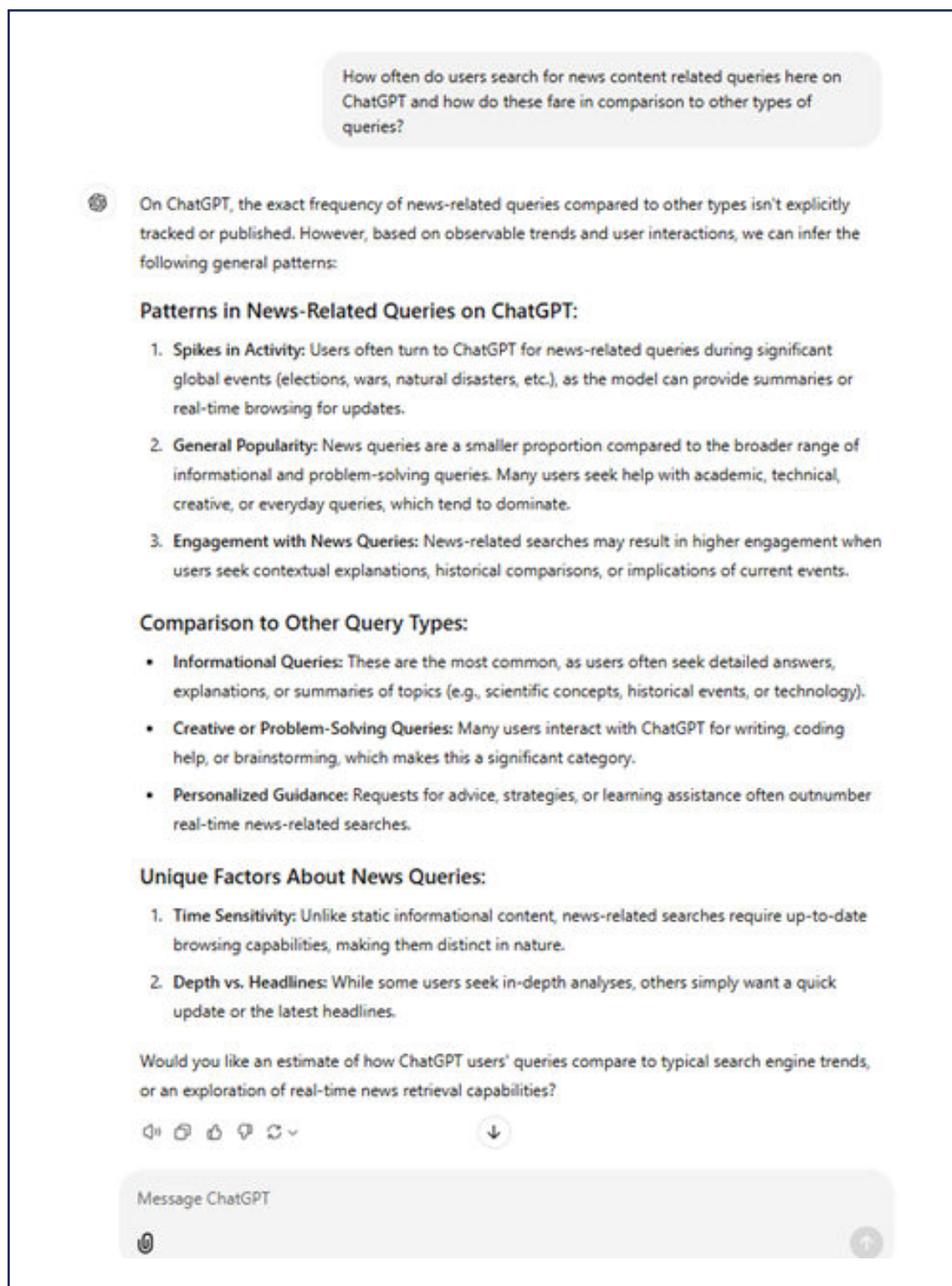
Keep in mind that these estimates are based on aggregate data and may vary depending on the specific context and user demographics.

Would you like me to provide more insights or clarify any of these points?

Ask Meta AI anything...

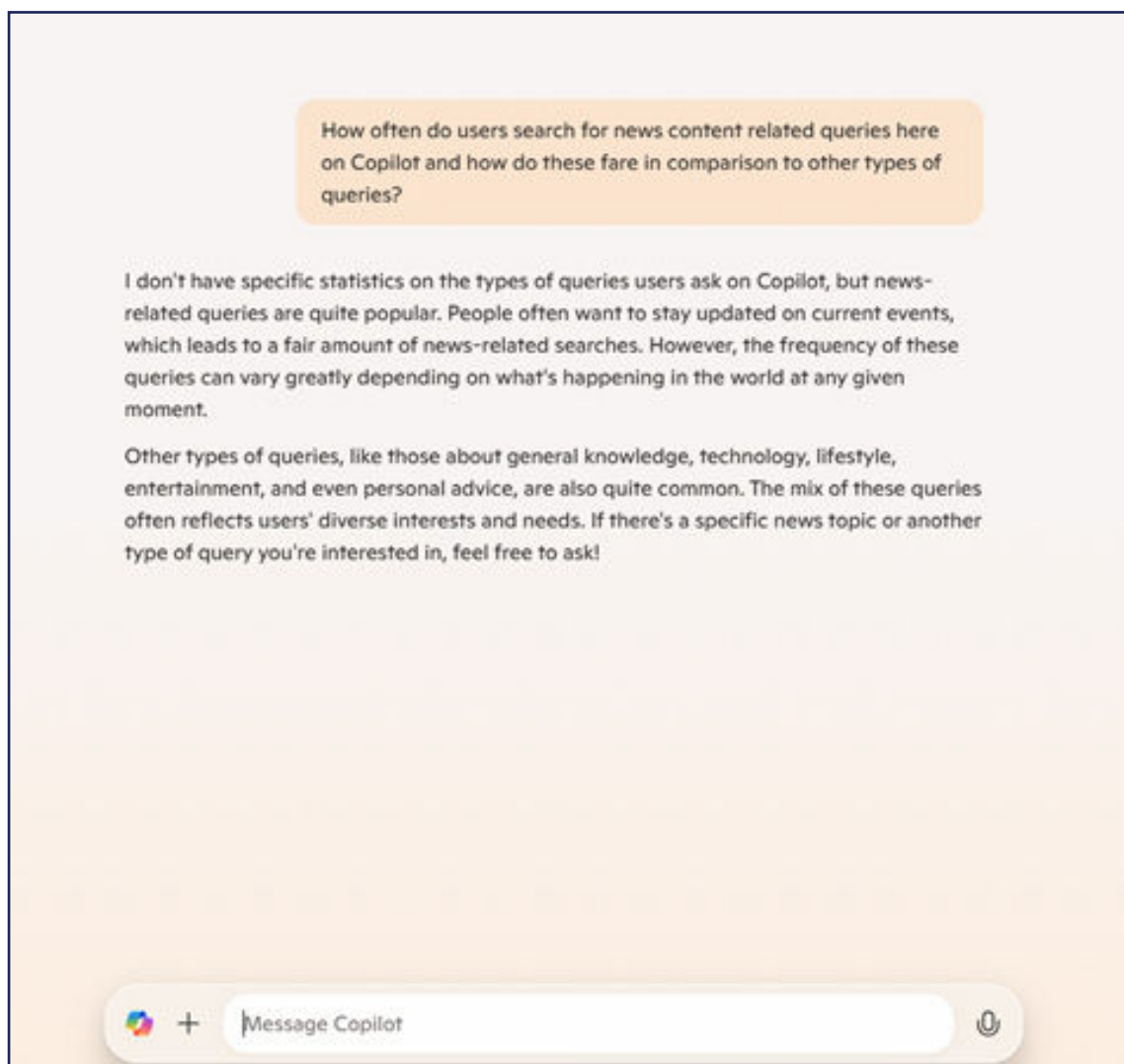
Messages are generated by AI and may be inaccurate or inappropriate. [Learn more](#)

Source: Screenshot of Meta AI response taken 20 November 2024



Source: Screenshot of Meta AI response taken 20 November 2024

APPENDIX D - COPILOT RESPONSE



Source : Screenshot of Copilot response taken 20 November 2024

APPENDIX E – GEMINI RESPONSE

S

How often do users search for news content related queries here on Gemini and how do these fare in comparison to other types of queries?

Show drafts

Google does not publicly disclose specific data on the frequency of news-related queries on Gemini compared to other types of queries. However, we can make some general observations based on the nature of language models like Gemini and current trends in information consumption:

- News-Related Queries:** Given Gemini's ability to access and process information from the real world through Google Search, it's reasonable to assume that users frequently turn to it for news updates, explanations of current events, and summaries of news articles. This is especially true for breaking news or complex topics that require concise and easy-to-understand explanations.
- Comparison to Other Queries:** While news-related queries are undoubtedly significant, they likely constitute only a portion of the overall query volume on Gemini. Users also employ the model for a wide range of tasks, including:
 - Information seeking (e.g., "What is the capital of France?")
 - Task completion (e.g., "Write a poem about a lonely robot")
 - Creative writing (e.g., "Write a story about a dragon who wants to be a chef")
 - Problem-solving (e.g., "How do I fix a leaky faucet?")
 - Language translation (e.g., "Translate 'Hello, how are you?' into Spanish")

The relative frequency of news-related queries versus other types of queries can vary depending on factors such as current events, user interests, and the specific capabilities of the language model.

It's important to note that Gemini is constantly evolving, and its ability to handle news-related queries may improve over time as the model learns from more data and refines its understanding of the world.

Sources and related content

Writing Content with AI | Google Workspace

workspace.google.com

Translate And Learn | Gemini API Developer Competition | Google A...

ai.google.dev

How does Ge and continu

answers.b

Ask Gemini

Gemini can make mistakes, so double-check it

Source: Screenshot of Meta AI response taken 20 November 2024

COMPETITION COMMISSION

63



competition commission
south africa



Telephone Number: +27 (012) 394-3200 / 20

WhatsApp / SMS Line: +27 (0) 84 743 0000

Email Address: ccsa@compcom.co.za

Physical address: The DTI Campus, Mulayo (Block C), 77 Meintjies Street, Sunnyside, Pretoria

www.compcom.co.za



a competitive, dynamic, deconcentrated and inclusive economy