EU DISINFO LAB

TERMS OF (DIS)SERVICE

Comparing misinformation policies in text-generative AI chatbot

FEBRUARY 2025

Author: Maria Giovanna Sessa (EU DisinfoLab)

Reviewers: Amaury Lesplingart (Check First), Joe McNamee (EU DisinfoLab)

INTRODUCTION

- Large language models (LLMs) are rapidly proliferating. Like any technological tool, they can be harnessed for legitimate purposes but also misused or exploited by malicious actors. As these models become more integrated into everyday applications, concerns about their role in spreading misinformation continue to grow, calling for solid policies to prevent this threat.

- This factsheet collects and analyses the misinformation-related policies of 11 leading chatbots, selected based on NewsGuard's selection. Our focus is on text-generative AI, given its widespread use across various domains – including content creation, translation, and summarisation – as well as its role in assisting users by answering questions.

- For each LLM, we examine key policy elements, including explicit references to misinformation and related prohibited activities – such as scams or impersonation. Additionally, we outline content moderation practices, user reporting mechanisms, and the consequences of violating the platform's Terms of Service (ToS).

- A note on methodology and limitations: the information provided here is based on publicly available sources, which vary in clarity, accessibility, and format. While we have made every effort to compile a thorough and accurate guide, gaps or omissions may exist due to the availability of information at the time of our writing. If any inaccuracies are identified, we welcome feedback and will gladly make corrections.

TABLE 1

General policy references, legal compliance, and mentions of misinformation

Platform	Policy reference	Mention of misinformation
Chat GPT (OpenAI)	• Usage policies (updated on 29/01/2025). • Law compliance statement.	Yes *Universal Policies:* • Prohibition to harm others (including scams and misleading activities). *Building with ChatGPT:* • Prohibition to misinform, misrepresent, or mislead others (including disinformation and impersonation).
Claude (Anthropic)	• Consumer Terms of Service (updated on 13/05/24). • Law compliance statement.	Yes *Acceptable Use Policy:* • Prohibition to spread misinformation. • Prohibition to create political campaigns or interfere in elections. • Prohibition to engage in fraudulent, abusive, predatory practices. • Prohibition to abuse the platform (e.g., coordinating malicious activities).
Copilot (Microsoft)	• Copilot AI Experiences Terms (updated on 1/10/2024). • Law compliance statement. • Code of Conduct signatory.	Yes *Code of Conduct:* • Prohibition to engage in fraudulent, false, or misleading activities (including disinformation, and impersonation).
Deepseek-R1 (DeepSeek)	• DeepSeek Terms of Use (updated on 20/01/2025). • Law compliance statement.	No *Terms of Use:* • No explicit reference to misinformation. • Prohibition to harm and impersonate others.
Gemini (Google)	• Terms of Service (Google) (updated on 22/05/2024). • Generative AI-prohibited use policy (updated on 17/12/2024). • Gemini API Additional Terms of Service (updated on 5/02/2025). • Law compliance statement. • Code of Conduct signatory.	Yes *Generative AI-prohibited Use Policy:* • Prohibition to misinform, misrepresent, or conduct misleading activities (including scams and impersonation).

		• Exceptions[1] can be made "based on educational, documentary, scientific, or artistic considerations" or where benefits outweigh harms.
Grok-2 (xAI)	• Terms of Service – Consumer (updated on 2/01/2025). • Terms of Service – Enterprise (updated on 23/12/2024). • Law compliance statement.	No *xAI Acceptable Use Policy:* • No explicit reference to misinformation. • Prohibition to scam. • Prohibition to mislead.
Le Chat (Mistral AI)	• Legal terms and conditions – Terms of Service (updated on 6/02/2025). • Law compliance statement.	Yes *Use Policy:* • Prohibition to engage in fraudulent activities or scams (including impersonation). Prohibition to misinform.
Meta AI (Meta)	• Meta AIs Terms of Service (EU)[2] (updated on 17/01/2025). • Meta AIs Terms of Service (updated on 17/01/2025). • Law compliance statement. • Code of Conduct signatory.	Yes *Meta AIs Terms of Service (EU):* • Prohibition to deceive or mislead others (including scams, misinformation, or disinformation).
Perplexity API (Perplexity)	• Perplexity API Terms of Service (updated on 17/02/2024). • Law compliance statement.	No *Use of platform:* • No explicit reference to misinformation. • Prohibition to impersonate others or provide false information.
Pi (Inflection)	• Privacy Policy (updated on 19/09/2023). • Law compliance statement.	Yes *Acceptable Use:* • Prohibition to engage in harmful uses (including misinformation). • Prohibition to impersonate others during the sign-up process
You.com	• Terms of Service (updated on 27/08/2024). • Law compliance statement.	Yes *Acceptable Use Policy:* • Prohibition to misinform, deceive, or mislead others (including disinformation, misinformation, and impersonation).

[1] It is unclear who makes these decisions concerning exceptions and when.

[2] We find it contradictory that Meta has EU-specific Meta AI ToS despite being unavailable in the EU.

TABLE 2

ToS violations: detection, moderation, consequences, reporting, and appeal mechanisms

Platform	ToS violation detection and content moderation	ToS violation reporting	Consequences for ToS violation and appeal mechanism[3]
Chat GPT (OpenAI)	• Combination of automated and human detection, and user reports.	• Product interface reporting using the "thumbs down". • Content reporting form.	• Actions against content or account (e.g., warnings, restrictions, ineligibility for GPT Store or monetisation), suspension, or termination. • Account ban appeal.
Claude (Anthropic)	• Content moderation by Trust and Safety team (process unspecified). • Adjustable safety filters on prompts. • User's responsibility to verify output accuracy independently.	• Product interface reporting to Trust & Safety team using the "thumbs down" button. • Reporting via email (usersafety@anthropic.com).	• Account throttling, suspension, or termination. • Account ban appeal.
Copilot (Microsoft)	• Content restriction or removal if in violation of the Code of Conduct (process unspecified). • User's responsibility to verify output accuracy independently.	• Product interface reporting using the Feedback button. • Content reporting form.	• Account suspension, service limitations, or content removal, deletion, or restriction. • Account ban appeal.
Deepseek-R1 (DeepSeek)	• Content removal (process unspecified). • User's responsibility to verify output accuracy independently.	• Product interface reporting using the "Contact Us" button. • Reporting via email(chat: service@deepseek.com; platform: api-service@deepseek.com) or contact address (5th Floor, North Building, Block C, Rongke Information Center, No.2 South Science Academy Road, Haidian District, Beijing, China).	• Warnings (and deadline for correction), account function restriction, suspension, termination, prohibition to re-register, and content removal. • Account ban appeal.
Gemini (Google)	• Combination of automated and human detection by	• Product interface reporting using the "Bad response" button or	• Content removal.

[3] The table reports the language used by each platform. Due to the lack of definitions provided, we cannot determine whether some terms – such as 'termination' and 'deletion' are synonymous.

	Google's Trust and Safety Team. • Adjustable safety filters on prompts.	selecting "Help" and then "Report a problem". • Content reporting form.	• Google account suspension or termination. • Google account ban appeal.[4]
Grok-2 (xAI)	• Content removal (process unspecified). • User's responsibility to verify output accuracy independently.	• No reporting mechanism.	• Content deletion or disabling. • Service access suspension or termination, account deletion. • Account ban appeal via email (support@x.AI).
Le Chat (Mistral AI)	• Automated monitoring. • User's responsibility to verify output accuracy independently.	• Product interface reporting using the "thumbs down" feature. • Reporting via Help Centre or email (legal@mistral.ai).	• Account termination or suspension. • Account ban appeal via the Help Centre or email (support@mistal.ai).
Meta AI (Meta)	• Combination of automated and human detection. • User's responsibility to verify output accuracy independently.	• Product interface reporting using the "thumbs down" button or selecting "Report a Problem".	• Account suspension or permanent disabling. • Facebook account ban appeal.[5]
Perplexity API (Perplexity)	• Content removal (process unspecified).	• Product interface reporting using the "flag icon". • Reporting via email (support@perplexity.ai) or submission form.	• Account suspension or termination. • No specific account ban appeal process but inquiries via email (support@perplexity.ai).
Pi (Inflection)	• Content moderation (process unspecified) and removal (copyright infringement). • User's responsibility to verify output accuracy independently.	• Product interface "Report" button (called the "Feedback").	• Account suspension or termination. • Account ban appeal via email (support@pi.ai).
You.com (You.com)	• Combination of automated and human detection.[6]	• Product interface reporting using the "thumbs down" feature or selecting "More" and then "Send Feedback".	• Service access suspension or termination, account termination. • No specific account ban appeal process but inquiries via "Contact Us".

[4] Since Gemini is linked to a user's Google account, losing access to the Google account would mean losing access to Gemini too. However, it is unclear whether Google can suspend access to Gemini specifically while keeping the rest of the Google account functional, as it does with some services like YouTube and AdSense.

[5] As of now, it is unclear whether a Meta AI account is separate from a user's primary Meta account (e.g., Facebook). Due to this ambiguity and the lack of relevant documentation, we referenced the Facebook account suspension process.

[6] This answer was provided by the You.com chatbot. However, we could not evidence to support it, and when asked to reference its statement, the chatbot responded: "Content moderation policies, especially with AI-powered platforms like You.com, are often evolving and may not always be fully public."

TABLE 3

A comparative overview of text-generative LLMs' misinformation policies

	Law compliance statement	Code of Conduct signatory	Mention of misinformation	ToS violation detection and content moderation	Mention of user fact-checking responsibility	ToS violation reporting mechanism	Suspension and termination as ToS violation consequences	ToS violation appeal mechanism
Chat GPT (OpenAI)	Yes	No	Yes	Yes (automated + human)	No	Yes	Yes	Yes
Claude (Anthropic)	Yes	No	Yes	Yes (unspecified)	Yes	Yes	Yes	Yes
Copilot (Microsoft)	Yes	Yes	Yes	Yes (unspecified)	Yes	Yes	Yes	Yes
Deepseek-R1 (DeepSeek)	Yes	No	No	Yes (unspecified)	Yes	Yes	Yes	Yes
Gemini (Google)	Yes	Yes	Yes	Yes (automated + human)	No	Yes	Yes	Yes
Grok-2 (xAI)	Yes	No	No	Yes (unspecified)	Yes	No	Yes	Yes
Le Chat (Mistral AI)	Yes	No	Yes	Yes (automated)	Yes	Yes	Yes	Yes
Meta AI (Meta)	Yes	Yes	Yes	Yes (automated + human)	Yes	Yes	Yes	Unclear
Perplexity API (Perplexity)	Yes	No	No	Yes (unspecified)	No	Yes	Yes	No
Pi (Inflection)	Yes	No	Yes	Yes (unspecified)	Yes	Yes	Yes	Yes
You.com	Yes	No	Yes	Yes (automated + human)	No	Yes	Yes	No

CONCLUDING REMARKS

- In the 11 chatbots considered, references to misinformation vary (and sometimes are missing). Even when the term is mentioned, it is rarely defined. This is not uncommon: what some platforms call 'deletion', others call 'termination', though potential nuances remain unclear. Additionally, it is unspecified who enforces practices, how long suspensions last, and what distinguished warnings, suspensions, and terminations. On another note, more stringent prohibitions generally focus on hate speech, highlighting the illegal nature of the issue.

- Overall, ToS violation detection and content moderation processes are unclear. Some platforms mention dedicated moderation teams and processes – typically a mix of automated and human reviews – but the details remain vague. There is no transparency regarding the resources invested in content moderation, including the number of personnel, their locations and language coverage, and the timeline for action.

- Given our focus on misinformation, we concentrated on content reporting. Nevertheless, it is evident that platforms prioritise reporting breaches of and vulnerabilities in their architecture over content-related concerns.

- Law compliance is usually referenced in broad terms as adherence to "all applicable legislation". Platforms primarily emphasise privacy and data protection, the prohibition of illegal activities (with a particular focus on child protection), restrictions on reverse-engineering platform functions, and using the services for biometric identification. Notably, there is a perception that users bear greater responsibility for providing accurate data and avoiding platform abuse than the platforms themselves do in terms of accountability. This impression arises because the ToS are detailed regarding user obligations but more superficial about platform responsibilities. Uncertainty remains regarding platform incentives and their actual capacity to enforce these rules.

- Most platforms specify that fact-checking is the user's responsibility as if trying to exempt themselves from liability related to providing false and inaccurate answers. However, it is worth noting that the general provisions of the EU's Digital Services Act (DSA) require online platforms to remove illegal content expeditiously once they have actual knowledge of its illegality.

- If current platform approaches to misinformation already seem inadequate, there are valid concerns that the situation will deteriorate further. The cross-platform trend of dismantling safety policies and moderation teams reinforces these fears. Examples pointing in this direction include OpenAI's decision to remove some content warnings on ChatGPT, and xAI's introduction of "Unhinged Mode" in Grok-2, a beta feature described as intentionally "objectionable, inappropriate, and offensive, much like an amateur stand-up comic who is still learning the craft."

EU DISINFO LAB

www.disinfo.eu