THE FUTURE OF FREE SPEECH

THAT VIOLATES MY POLICIES

AI LAWS, CHATBOTS, AND THE FUTURE OF EXPRESSION

Directed by
Jordi Calvet-Bademunt,
Jacob Mchangama, and Isabelle Anzabi

OCTOBER 2025

Acknowledgments

The Future of Free Speech is an independent, nonpartisan think tank based at Vanderbilt University. Our mission is to reaffirm freedom of expression as the foundation of free and thriving societies through actionable research, practical tools, and principled advocacy. We envision a world in which the right to freedom of expression is safeguarded by law and strengthened by a culture that embraces diverse viewpoints.

This project was led by Jordi Calvet-Bademunt (Senior Research Fellow), Jacob Mchangama (Executive Director), and Isabelle Anzabi (Research Associate) at The Future of Free Speech. Together, they also drafted the chapters on the European Union and the United States of America.

We are grateful to Justin Hayes, Director of Communications, for overseeing the design of the report; Wendy H. Burch, Chief Operating Officer, for coordinating all administrative aspects of the project; and Sam Cosby, Director of Development, for leading the funding efforts that made this work possible.

We extend our thanks to the leading experts who contributed chapters on their respective jurisdictions: Carlos Affonso Souza (Brazil), Ge Chen (China), Sangeeta Mahapatra (India), and Kyung Sin (K.S.) Park (Republic of Korea). We are also grateful to Kevin T. Greene and Jacob N. Shapiro of Princeton University for their chapter, "Measuring Free Expression in Generative AI Tools."

We thank all the experts who contributed to individual chapters of this report; their names are listed in the relevant sections.

We are further indebted to Barbie Halaby of Monocle Editing for her careful editorial work across all chapters, and to Design Pickle for the report's design.

Finally, we are especially grateful to the Rising Tide Foundation and the Swedish Postcode Lottery Foundation for their generous support of this work, and we thank Vanderbilt University for their collaboration with and support of The Future of Free Speech.

Publication Date: October 2025

Preface

In this report, we explore the ways in which public and private governance of generative artificial intelligence (AI) shape the space for free expression and access to information in the 21st century.

Since the launch of ChatGPT by OpenAI in November 2022, generative AI has captured the public imagination. In less than three years, hundreds of millions of people have adopted OpenAI's chatbot and similar tools for learning, entertainment, and work.[1] Anthropic, another AI giant, now serves more than 300,000 business customers.[2] AI companies are valued in the hundreds of billions of US dollars[3], while established technology giants such as Google, Meta, and Microsoft are investing billions in the race to dominate the field.[4]

Generative AI refers to systems that create content — including text, images, video, audio, and software code — in response to user prompts.[5] Chatbots such as ChatGPT are the most visible examples, but generative AI is rapidly being embedded into the tools people use every day for both communication and access to information, from social media and email to word processors and search engines.

Recognizing generative AI's potential for expression and access to information, The Future of Free Speech undertook a first-of-its-kind analysis of freedom of expression in major models. In February 2024, we assessed the "free-speech culture" of six leading systems, focusing on their usage policies and responses to prompts.[6] Our findings revealed that excessively broad and vague rules often resulted in undue restrictions on speech and access to information.[7] By April 2025, when we updated this work, we observed signs of change: Some models showed greater openness.[8]

This current report builds on those foundations and pursues a more ambitious goal. Supported by leading experts, The Future of Free Speech undertakes a deeper examination of how national legislation and corporate practices shape freedom of expression in the era of generative AI. "*That Violates My Policies": AI Laws, Chatbots, and the Future of Expression* explores:

- AI legislation in Brazil, China, the European Union, India, the Republic of Korea, and the United States.[9] In this report, AI legislation refers to laws and public policies addressing AI-generated content, with

1 MacKenzie Sigalos, "OpenAI's ChatGPT to Hit 700 Million Weekly Users, Up 4x from Last Year," *CNBC*, August 4, 2025, https://www.cnbc.com/2025/08/04/openai-chatgpt-700-million-users.html.
2 Hayden Field, "Anthropic Is Now Valued at $183 Billion," *The Verge*, September 2, 2025, https://www.theverge.com/anthropic/769179/anthropic-is-now-valued-at-183-billion.
3 Kylie Robison, "OpenAI Is Poised to Become the Most Valuable Startup Ever: Should It Be?," *Wired*, August 19, 2025, https://www.wired.com/story/openai-valuation-500-billion-skepticism/; Krystal Hu and Shivani Tanna, "OpenAI Eyes $500 Billion Valuation in Potential Employee Share Sale, Source Says," *Reuters*, August 6, 2025, https://www.reuters.com/business/openai-eyes-500-billion-valuation-potential-employee-share-sale-source-says-2025-08-06/.
4 Blake Montgomery, "Big Tech Has Spent $155bn on AI This Year: It's About to Spend Hundreds of Billions More," *The Guardian*, August 2, 2025, https://www.theguardian.com/technology/2025/aug/02/big-tech-ai-spending.
5 Cole Stryker and Mark Scapicchio, "What Is Generative AI?," IBM Think, March 22, 2024, https://www.ibm.com/think/topics/generative-ai.
6 Jordi Calvet-Bademunt and Jacob Mchangama, *Freedom of Expression in Generative AI: A Snapshot of Content Policies* (Future of Free Speech, February 2024), https://futurefreespeech.org/wp-content/uploads/2023/12/FFS_AI-Policies_Formatting.pdf.
7 Calvet-Bademunt and Mchangama, *Freedom of Expression in Generative AI*.
8 Jordi Calvet-Bademunt, Jacob Mchangama, and Isabelle Anzabi, "One Year Later: AI Chatbots Show Progress on Free Speech — But Some Concerns Remain," *The Bedrock Principle*, April 1, 2025, https://www.bedrockprinciple.com/p/one-year-later-ai-chatbots-show-progress.
9 To select the countries, we considered Stanford University's 2023 Global AI Vibrancy Ranking (the most recent available at the time of writing), along with factors such as geographic diversity, population size, democratic and freedom status, and the presence of existing or emerging AI-related legislation.

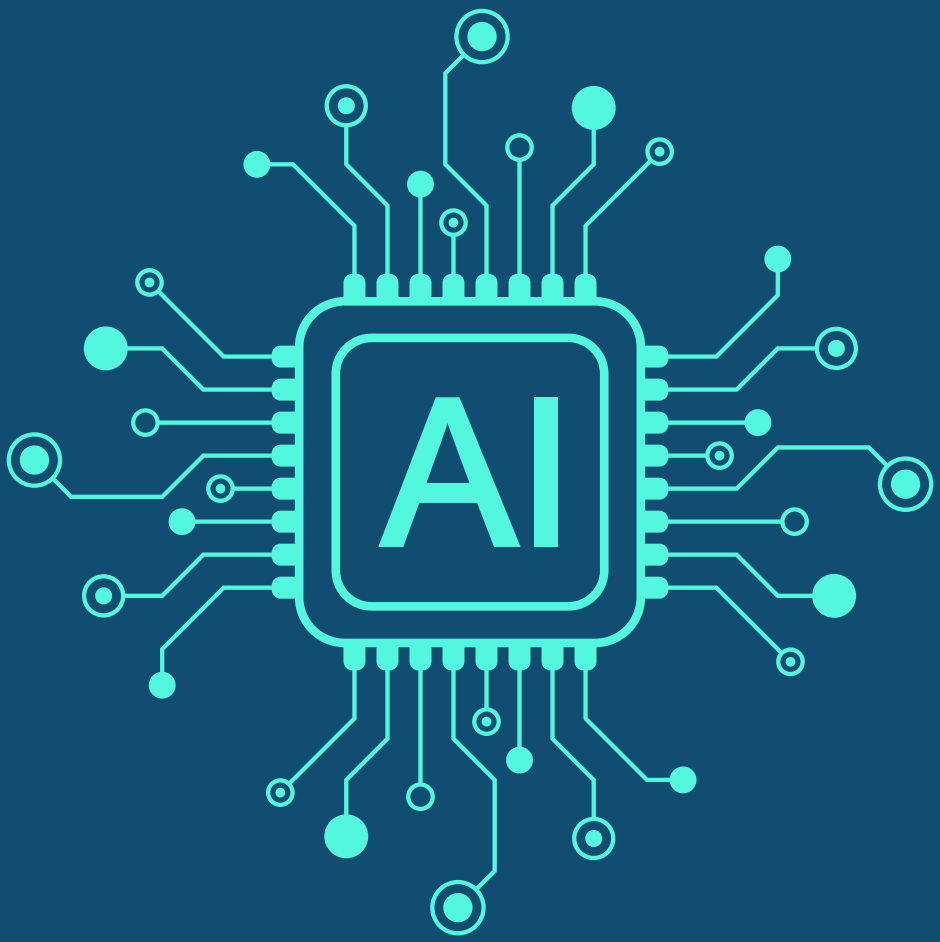
particular focus on elections and political speech, hate speech, defamation, explicit content (including child sexual abuse material and nonconsensual intimate images), and copyright. We also consider measures that actively promote freedom of expression, such as AI literacy initiatives and policies supporting cultural and linguistic diversity.

- Corporate practices of major AI developers, including Alibaba, Anthropic, Google, Meta, Mistral AI, DeepSeek, OpenAI, and xAI.[10] We examine their usage policies, model performance in responding to prompts, and the limited available information on their training data and development processes.

This report seeks to provide a rigorous and timely analysis of how generative AI is reshaping the space for free expression in both the public and private spheres. Building on these insights, The Future of Free Speech is developing guidelines to help policymakers and companies ensure that generative AI protects and enhances freedom of expression and access to information, two cornerstones of democratic societies.

In an era of rapid technological change, safeguarding free expression is a matter not only of rights but of preserving the conditions for open, informed, and thriving democracies.

10 We selected major models from leading companies that are accessible through a web interface and include text-generation capabilities. In addition, we considered the geographic location of the model provider and the degree of openness of the models.



Freedom of Expression in Generative AI Models

Jordi Calvet-Bademunt, Jacob Mchangama, Isabelle Anzabi,* and Carlos Olea†

* Jordi Calvet-Bademunt, Jacob Mchangama, and Isabelle Anzabi serve as senior research fellow, executive director, and research associate, respectively, at The Future of Free Speech. We thank Hirad Marami for his dedication in submitting and reviewing all prompts in Section 6. We also thank Natalie Alkiviadou for her valuable comments and suggestions. In addition, we are grateful to Becca Branum, David Inserra, Elena Yndurain, Elonnai Hickok, Jason Pielemeier, Kate Ruane, Laura Lázaro Cabrera, and Min Aung for their insightful feedback, which substantially improved the questionnaire used for ranking the generative AI models. We further thank John G. Geer and Svend-Erik Skaaning for their comments on the prompts used in Section 6. All remaining errors are our own.
† Carlos Olea, PhD student at Vanderbilt University's Department of Computer Science, coauthored Section 4 and provided valuable comments across the chapter.

Abstract

This chapter evaluates the relationship between generative artificial intelligence (AI) and freedom of expression, focusing on how leading models regulate speech through policies, training, and real-world outputs. We analyze eight prominent AI systems: OpenAI's GPT-5, Anthropic's Claude Sonnet 4, Google's Gemini 2.5 Flash, Meta's Llama 4, xAI's Grok 4, Mistral AI's Mistral Medium 3.1, DeepSeek's DeepSeek-V3.1, and Alibaba's Qwen3-235B-A22B. Our methodology combines the review of usage policies, the analysis of transparency in training, and a prompting exercise involving 512 lawful but controversial prompts (64 per model) across themes such as political discourse, human rights, misinformation, and elections.

We also rank the "free-speech culture" of the selected models, considering factors such as companies' commitment to and policies on free expression; the model's willingness to engage with diverse perspectives; its degree of openness; the available information on its training; usage policies and terms of service; transparency toward users in content moderation decisions; performance when prompted with controversial topics; and measures to empower expression, such as support for media and AI literacy and for diverse languages and cultures.

Although none achieved an excellent score, xAI's Grok 4 came out on top. At the other end of the spectrum, Alibaba's Qwen3-235B-A22B and DeepSeek-V3.1 were the weakest performers, reflecting China's state-imposed regime of strict control over AI-generated content. Overall, the analysis shows that no company has yet developed a fully coherent and transparent free-speech framework. Encouragingly, there are examples of good practices — especially in prompt performance, user empowerment, and explicit free-speech commitments — that can serve as building blocks for more rights-respecting approaches going forward.

This chapter's findings reveal progress: Refusal rates have decreased compared to a similar exercise we conducted in 2024, with some companies showing greater willingness to engage with contentious topics. The models from xAI, Meta, and Mistral AI performed most openly, while Alibaba's model was uniquely restrictive on sensitive issues. In all cases except DeepSeek, models proved more receptive to creating abstract argumentation about specific topics than to generating content for social media, potentially reflecting heightened sensitivity to advocacy-style requests.

Yet challenges remain. Usage policies are vague and not robustly grounded in international human rights, and models' training processes remain opaque. Without greater transparency and clearer safeguards, AI systems risk becoming algorithmic gatekeepers of public discourse. We argue that embedding freedom of expression and access to information as a design principle is essential to ensuring these technologies enrich, rather than constrain, democratic debate.

Jordi Calvet-Bademunt

Jordi Calvet-Bademunt is a Senior Research Fellow at The Future of Free Speech. He is also a Visiting Legal Researcher at the Barcelona Supercomputing Center, where he advises on trustworthy AI. His work focuses on AI policy and digital governance, and he has written extensively and provided commentary in both specialist and mainstream media. Previously, Jordi spent about a decade working at the Organisation for Economic Co-operation and Development (OECD) and as an associate at leading European law firms. He holds advanced degrees from Harvard University and the College of Europe in Bruges, Belgium.

Jacob Mchangama

Jacob Mchangama is the Founder and Executive Director of The Future of Free Speech. He is a research professor at Vanderbilt University and a Senior Fellow at The Foundation for Individual Rights and Expression (FIRE). In 2018, he was a visiting scholar at Columbia's Global Freedom of Expression Center. He has commented extensively on free speech and human rights in outlets including the Washington Post, the Wall Street Journal, The Economist, Foreign Affairs and Foreign Policy. Jacob has published in academic and peer-reviewed journals, including Human Rights Quarterly, Policy Review, and Amnesty International's Strategic Studies. He is the producer and narrator of the podcast "Clear and Present" Danger: A History of Free Speech and the critically acclaimed book Free Speech: A History From Socrates to Social Media, published by Basic Books in 2022. He is the recipient of numerous awards for his work on free speech and human rights.

Isabelle Anzabi

Isabelle Anzabi is a research associate at The Future of Free Speech, where she analyzes the intersections between AI policy and freedom of expression. She is bringing her background in digital rights policy and global regulatory approaches to content moderation and AI governance. Previously, Isabelle was an AI & Human Rights Fellow with the European Center for Not-for-Profit Law, a Knowledge Fellow at the DiploFoundation, and a research group member at the Center for AI and Digital Policy. Isabelle received her B.A. in Political Science from Stanford University. She also studied digital governance at Oxford University and interned at institutions, such as the World Bank and CISA. On campus, Isabelle was affiliated with the Stanford Center for Racial Justice, the Stanford Legal Design Lab, the Stanford Cyber Policy Center, the Stanford Constitutional Law Center, the Stanford Technology Law Review, and the Public Service Leadership Program.

Carlos Olea

Carlos Olea is a PhD Candidate at Vanderbilt University in Nashville, TN. His work focuses on interdisciplinary applications of Artificial Intelligence, Artificial Intelligence utilization, limitations, and safety. His work includes collaboration with the NSA and DARPA on AI safety and AI-augmented design and engineering.

1. Introduction

One year ago, our inaugural report on AI chatbots and free speech, "Freedom of Expression in Generative AI: A Snapshot of Content Policies," revealed a concerning trend: Major generative AI models were systematically over-censoring legitimate discourse, refusing to engage with controversial but lawful content, and applying content restrictions that went far beyond legal requirements.[1] This reflected a trend we had first documented on social media platforms.[2] We found that leading AI systems had become overly cautious gatekeepers, blocking discussions on everything from political controversies to historical debates under the guise of safety.

Today, as generative AI has become increasingly integrated into how hundreds of millions of people access information, create content, and engage in public discourse,[3] the stakes for getting content moderation right have only grown higher. These systems no longer function merely as chatbots; they serve as research assistants, writing tools, educational resources, and information sources for users worldwide.

AI companies face regulatory requirements, reputational risks, and legitimate concerns about their systems being misused to incite violence, generate child sexual abuse material, or facilitate criminal activity. Safeguards are therefore both natural and necessary. The key question is not whether restrictions should exist but whether they are clearly defined, proportionate, and calibrated in ways that robustly protect the right to freedom of expression and access to information.

When AI models refuse to engage with lawful topics or systematically privilege certain viewpoints, they shape not only individual conversations but the broader contours of public discourse as well. By filtering out particular perspectives, these systems risk creating and entrenching orthodoxies — unstated yet powerful constraints on what counts as acceptable debate across the tech stack, where generative AI is increasingly becoming the mediating layer for users.

In this chapter we expand on our previous work and examine whether AI companies have made meaningful progress in addressing these free-speech concerns, or whether the problems we identified have persisted, or even worsened, as these systems have scaled and evolved. The analysis considers eight major models from leading companies worldwide.

This 2025 analysis examines three dimensions: first, what users are permitted to do in theory, based on each model's stated policies; second, what users can actually do in practice, tested with more than 500 prompts on controversial topics (64 per model); and third, the limited transparency surrounding the training of these models.

1 Jordi Calvet-Bademunt and Jacob Mchangama, "Freedom of Expression in Generative AI: A Snapshot of Content Policies," The Future of Free Speech, February 2024, https://futurefreespeech.org/wp-content/uploads/2023/12/FFS_AI-Policies_Formatting.pdf.
2 Jacob Mchangama, Abby Fanlo, and Natalie Alkiviadou, "Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies," The Future of Free Speech, July 14, 2023, https://futurefreespeech.org/wp-content/uploads/2023/07/Community-Guidelines-Report_Latest-Version_Formated-002.pdf.
3 MacKenzie Sigalos, "OpenAI's ChatGPT to Hit 700 Million Weekly Users, Up 4x from Last Year," CNBC, August 4, 2025, https://www.cnbc.com/2025/08/04/openai-chatgpt-700-million-users.html.

Our assessment here reveals a mixed picture: Most companies have made notable improvements in reducing unnecessary refusals and providing more nuanced responses to complex topics. Still, the usage policies guiding what users can and cannot do with the models remain broad and vague. In addition, the transparency of the models' training processes is extremely limited. While this may be understandable for business reasons, it is problematic from a freedom of expression perspective.

As generative AI systems become primary interfaces for information access and content creation, their content policies and training decisions increasingly shape what ideas can be easily expressed, explored, and debated in digital spaces. In this chapter, we aim to shed light on these policies and decisions and on how they affect users.

2. Methodology

2.1. Model Selection

We analyze eight major generative AI models. They are:

- Alibaba's Qwen3-235B-A22B
- Anthropic's Claude Sonnet 4
- DeepSeek's DeepSeek-V3.1
- Google's Gemini 2.5 Flash
- Meta's Llama 4
- Mistral AI's Mistral Medium 3.1
- OpenAI's GPT-5
- xAI's Grok 4

The analysis centers on models of major AI companies. At the time of writing, all selected companies appear as top performers in LMArena's Text Arena ranking,[4] a leading benchmark in the industry. All selected companies were also highlighted in Stanford University's 2025 AI Index Report.[5] We have focused on the default model provided to users; when a subscription option exists, we have used the default model provided to paid users.[6]

All selected models are accessible through a web interface (which we refer to as "chatbot") and have text-generation capabilities. We focus on text-generation capabilities for two main reasons. First, it builds on our 2024 report, "Freedom of Expression in Generative AI: A Snapshot of Content Policies."[7] Second, it facilitates the analysis of the models' generated outputs when analyzing commitment to freedom of expression in practice, given the resources available.

In addition, we considered the geographic location of the model provider and the degree of openness of the models.

4 LMArena's Text Arena ranking considers models' versatility, linguistic precision, and cultural context across text. As of June 23, 2025, Meta ranks the lowest among the companies included in this analysis; its first model appears in position 38. However, Meta is considered because of the company's resources and distribution channels (notably, Instagram, WhatsApp, and Facebook) and general relevance in the AI race.

5 Nestor Maslej et al., Artificial Intelligence Index Report 2025 (Stanford, CA: AI Index Steering Committee, Stanford Institute for Human-Centered AI, April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf.

6 Model access was via the OpenAI API Platform, Google AI Studio, Claude Console, La Plateforme, xAI Cloud Console, and the DeepSeek Platform. Alibaba's and Meta's models were accessed through Vertex AI Studio.

7 Calvet-Bademunt and Mchangama, "Freedom of Expression in Generative AI."

The geographic scope covers five US-based companies (Anthropic, Google, Meta, OpenAI, and xAI), as well as Mistral AI in France and Alibaba and DeepSeek in China. This distribution reflects the leading countries producing top AI models. According to Stanford University's HAI 2025 Index, "in 2024, the United States led with 40 notable AI models, followed by China with 15 and France with three." For this reason, our analysis focuses on the United States, China, and the EU.[8]

Among the models we examine, three are open weight (Alibaba, DeepSeek, and Meta) and five are closed source (Anthropic, Google, Mistral AI, OpenAI, and xAI). For our purposes, open-weight models grant access to parameters but do not fully meet open-source criteria,[9] typically by imposing usage restrictions or not releasing full source code.

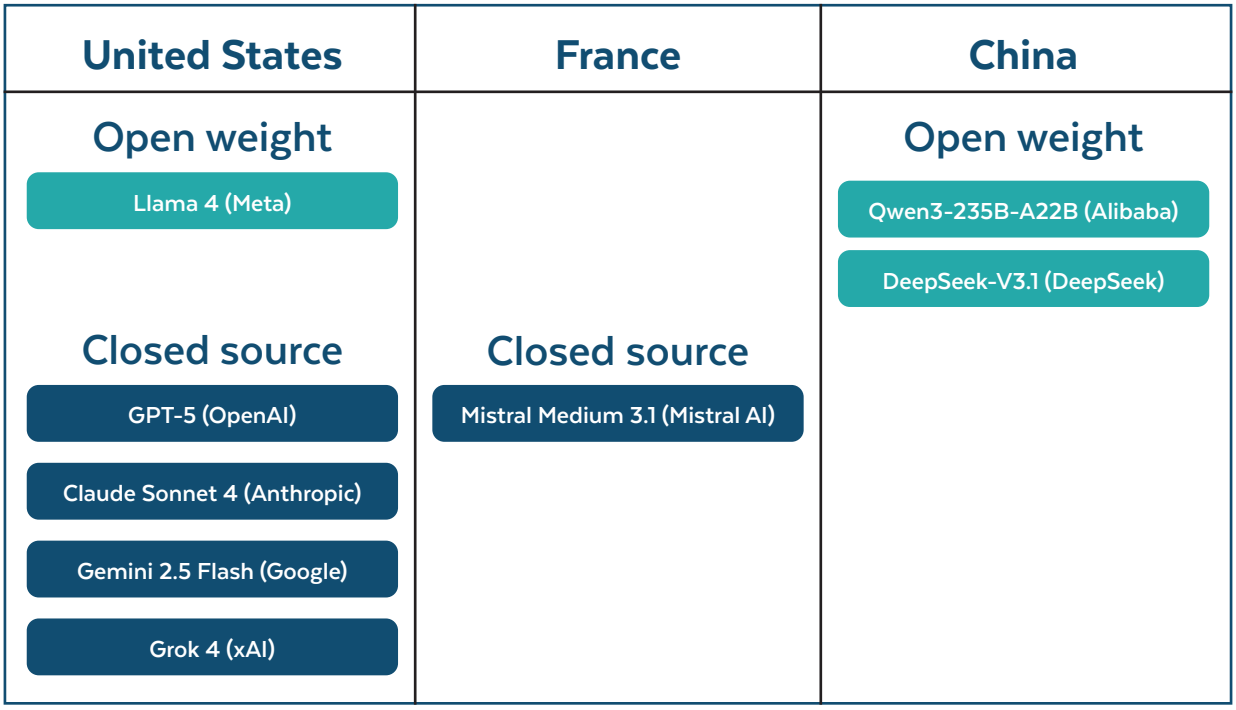


Figure 1. Generative AI models by country origin and openness. Created by The Future of Free Speech.

Our analysis focuses on general-purpose systems rather than domain-specific systems. The models examined here are designed to generate text across a wide range of topics and are marketed as tools for general information access, education, creativity, and research. Our concerns about freedom of expression and access to information are most directly applicable in this context, where restrictions on speech can significantly shape public discourse and limit users' ability to explore diverse perspectives. By contrast, domain-specific chatbots — such as those deployed in customer service, technical troubleshooting, or other narrowly defined functions — operate under very different expectations. In such cases, strict content controls are often appropriate and do not raise the same freedom of expression concerns, since users interact with these systems for targeted, instrumental tasks rather than for open-ended engagement with ideas.

We accessed the models for the prompting exercise through their application programming interfaces (APIs).

8 Maslej et al., Artificial Intelligence Index Report 2025, 46.
9 "The Open Source AI Definition 1.0," Open Source Initiative, version 1.0, accessed September 12, 2025, https://opensource.org/ai/open-source-ai-definition.

2.2. Data Source Selection

To conduct our analysis, we collected each company's respective model or system card, terms of service (the binding agreement between the provider and the user), and usage policies (supplementary rules that specify prohibited content beyond the basic agreement). For ease of reference, throughout this report we use the term "Service Terms and Policies" to encompass both the terms of service and usage policies. We collected these documents in May and June 2025. We also reviewed other official documents issued by the companies, including blog posts, press releases, and research publications.

These sources informed our analysis of what users are permitted to do in theory, as well as our assessment of how the models are trained. The latter was significantly constrained, given the extremely limited amount of information available. In parallel, we submitted 512 prompts across the eight AI models (64 per model) on contentious sociopolitical issues that included reproductive rights, colonial legacies and global inequality, questions of democratic legitimacy, and debates around diversity, equity, and inclusion in higher education, among others. Full details on this methodology are provided in Section 6.1. These prompts served to evaluate what users are able to do in practice.

We also developed a questionnaire to evaluate how the policies and practices of the respective companies promote, protect, or restrict users' freedom of expression. This instrument consists of 27 targeted questions that systematically address key aspects of freedom of expression and access to information in the context of generative AI. All companies behind the selected models were given the opportunity to comment on the questionnaire and provide feedback on the findings of The Future of Free Speech team. The questionnaires and those replies are available in "Appendix AI Models 1."

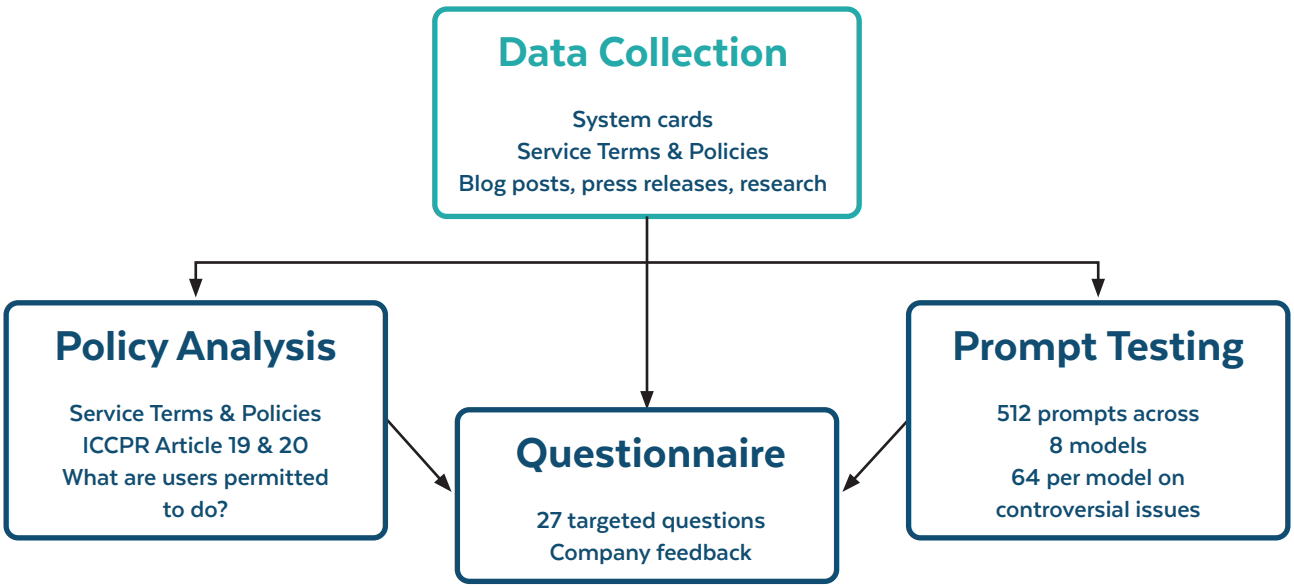


Figure 2. Structure of analysis. Created by The Future of Free Speech.

3. Model Rankings: Freedom of Expression

3.1. Overview

We ranked the "free-speech culture" of eight major generative AI models and their companies. By "free-speech culture," we mean the model's willingness to foster open dialogue and engage diverse perspectives. While none achieved an excellent score, xAI's Grok 4 came out on top with 65% of all possible points. At the other end of the spectrum, Alibaba's Qwen3-235B-A22B and DeepSeek-V3.1 were the weakest performers, with 22% and 32% respectively. All other companies, with the exception of Mistral Medium 3.1, scored at least half of the possible points. Notably, however, Mistral's model performed strongly in our prompt-testing exercise, as explained in Section 6.2.

Model	Ranking	Total Score
Grok 4 (xAI)	1	65.2%
GPT-5 (OpenAI)	2	60.3%
Claude Sonnet 4 (Anthropic)	3	58.6%
Gemini 2.5 Flash (Google)	4	58.4%
Llama 4 (Meta)	5	57.9%
Mistral Medium 3.1 (Mistral AI)	6	45.8%
DeepSeek-V3.1 (DeepSeek)	7	31.5%
Qwen3-235B-A22B (Alibaba)	8	21.9%

Table 1. Model ranking and total score (%). Created by The Future of Free Speech.

3.2. Methodology

To evaluate the models' "free-speech culture," we took into account the following: each company's commitment to and policies on free expression; the model's willingness to engage with diverse perspectives; its degree of openness; the available information on its training; its usage policies and terms of service; the transparency toward users in content moderation decisions; performance when prompted with contested sociopolitical issues; and measures to empower expression, such as support for AI literacy and for diverse languages and cultures.

This assessment employs a comprehensive instrument of 27 targeted questions that systematically address key aspects of freedom of expression and access to information in relation to AI. The questionnaire is organized into sections that broadly correspond to the sections of this chapter. The questions were developed by the team at The Future of Free Speech and shared with all analyzed companies and other stakeholders for feedback. The questionnaire itself was completed by The Future of Free Speech team, with technical support from Vanderbilt University's Department of Computer Science. The responses were then sent to the companies for comment. The questionnaires and the replies are available in "Appendix AI Models 1."

Using the questionnaires, we determined the total scores for each model. A higher aggregate score indicates a stronger commitment to freedom of expression. The ranking ranges from 1 (less freedom-restrictive) to 8 (more freedom-restrictive). The total score has a maximum of 66 points, which is the most freedom-protective outcome. The total score minimum is -2 points, given that one of the questions is reverse-scored.

3.3. Key Findings and Discussion

While the overall ranking reflects the general "free-speech culture" of the different models, the breakdown across categories highlights important nuances. Each section of the questionnaire reveals strengths and weaknesses that a simple total score cannot fully capture. In the prompt exercise (Section 6), where we tested hundreds of prompts on controversial issues, all models except Qwen3-235B-A22B responded to at least 73% of the prompts (results shown in the "Prompts Exercise" column in Table 2). Notably, despite underperforming in other categories, Mistral Medium 3.1 ranked among the top performers in this test. xAI, Google, and Meta also had a strong performance, responding to more than 90% of our prompts. This suggests that these models are comparatively effective at engaging with sensitive queries in practice. However, strong results in this category alone are not sufficient. To robustly protect freedom of expression and access to information, companies need a durable framework that ensures consistency over time and resists opaque policy changes. We recognize that companies are still in the process of developing these frameworks, given the novelty of generative AI, the complexity of the challenges involved, and the ongoing evolution of their technical and governance capabilities. Still, without such frameworks, approaches to free expression risk shifting unpredictably and without transparency or accountability.

We were encouraged that several companies, including Anthropic, Google, Meta, OpenAI, and xAI, explicitly commit to protecting freedom of expression and viewpoint diversity (considered in Table 2 in the column "Free-Speech Commitment"). Yet most companies perform poorly when it comes to the Service Terms and Policies that govern user behavior (covered in Table 2, "Terms & Policies" column). Restrictions on hate speech and disinformation are generally vague, lack clear connections to legitimate aims, and are rarely assessed against necessity and proportionality criteria.

Performance on pre-training and model evaluation indicators was weak (see scores in the "Training" column of Table 2). None of the companies disclosed, in a meaningful way, the data used to train their models. We recognize that limited transparency can be partly attributed to commercial and security concerns. At the same time, this opacity carries implications for freedom of expression, as explained in Section 4.3. Still, some progress is visible: Companies appear to be more deliberate in evaluating refusals and in experimenting with constructive forms of engagement rather than declining queries. Most companies performed reasonably well in terms of transparency toward users, with the exception of Alibaba and DeepSeek (covered by the column labeled "Transparency" in Table 2). Several providers explain the reasons for refusals and allow appeals when accounts are suspended. Transparency regarding state requests for content removal or account suspension, however, remains limited, with Google being the best performer in this area.

Most companies performed well in empowering users, whether by supporting multiple languages (including those from non-OECD countries), by offering AI literacy initiatives, or by providing other resources (considered in the "Empowerment" column of Table 2). On openness, Alibaba, DeepSeek, and Meta earned points for making their models more accessible through weights and permissive use (see column "Openness" in Table 2).

Overall, the analysis shows that no company has yet developed a fully coherent and transparent free-speech framework. Encouragingly, there are examples of good practices, especially in prompt performance, user empowerment, and explicit free-speech commitments, that can serve as building blocks for more rights-respecting approaches going forward.

Model	Free-Speech Commitment	Training	Openness	Terms & Policies	Transparency	Prompts Exercise	Empowerment	Total
Grok 4 (xAI)	5	0	0	11	5	16.0	6	43.0
GPT-5 (OpenAI)	5	5	0	4	5	12.8	8	39.8
Claude Sonnet 4 (Anthropic)	4	3	0	7	5	11.7	8	38.7
Gemini 2.5 Flash (Google)	5	2	0	4	5	14.6	8	38.6
Llama 4 (Meta)	5	2	3	3	3	15.2	7	38.2
Mistral Medium 3.1 (Mistral AI)	0	0	0	4	5	15.2	6	30.2
DeepSeek-V3.1 (DeepSeek)	-2	0	4	3	1	12.8	2	20.8
Qwen3-235B-A22B (Alibaba)	-2	0	4	2	0	8.5	2	14.5
Max. No. Points	6	8	4	16	6	16	10	66

Table 2. Section breakdown for model ranking (point values). Created by The Future of Free Speech.

The bars in Figure 3 illustrate the contribution of each component to the total ranking score. The questions corresponding to each component are provided in the questionnaires in "Appendix AI Models 1."

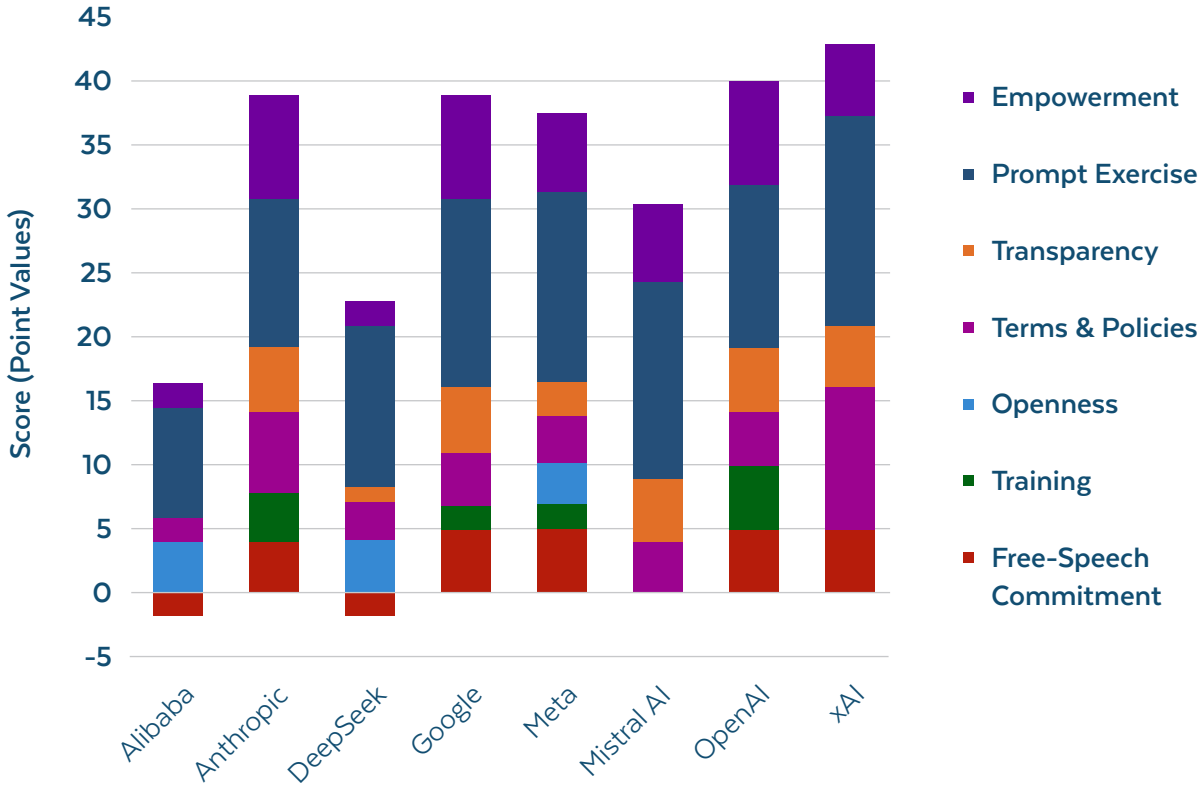


Figure 3. Composition of scores according to free-speech categories. Created by The Future of Free Speech.

4. How Are Generative AI Models Trained?

4.1. Key States of AI Model Training

The large language models (LLMs) powering generative AI are trained by example: They learn patterns in language by analyzing vast amounts of text data.[10] At the outset, a model's architecture is essentially a blank framework that must be trained to perform useful tasks.

LLMs generally consist of two core components. The first is the embedding model. This part transforms words, subwords, or symbols into numerical representations that the system can process. It effectively helps the model "understand" language by mapping linguistic elements onto a mathematical space. The second component is the predictive model. This part learns to generate text by predicting the next word in a sentence based on the preceding context. These models are typically built using transformer architectures and generate text one word at a time — a process known as autoregressive generation.

Both components are trained together on large datasets. During training, the model produces an output in response to an input, and its performance is evaluated by comparing that output to the expected or "correct" response. Based on how close the output is to the target, the model's internal parameters (or "weights") are adjusted. This process is repeated billions of times to improve accuracy. Sometimes, portions of the model — particularly the embedding layer — may be "frozen" once they reach a satisfactory level of performance. Later training, often called fine-tuning, focuses on smaller portions of the model, typically using more targeted or curated data.

LLMs can be further shaped through reinforcement learning, which involves scoring outputs based on how desirable they are. For instance, a model might receive negative feedback for using offensive or aggressive language. Over time, this teaches the model to avoid such outputs. This stage often imparts behavioral constraints and value-aligned responses, including politeness, safety, or adherence to platform policies.

Consumer-facing systems often include additional components beyond the core LLM. These may include:

- Content filtering systems to moderate outputs.
- Specialized embedding layers for handling images or other media (in multimodal models).
- "Red-teaming" exercises to test the model's responses to adversarial prompts and improve safety through targeted fine-tuning.

These add-ons further shape the model's behavior and can significantly influence what kinds of expression the system allows or suppresses.

10 LLMs are AI models trained on vast amounts of data, "making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks." See "What Are Large Language Models (LLMs)?," IBM, November 2, 2023, https://www.ibm.com/think/topics/large-language-models.

4.2. Opacity in Training Processes

LLMs are often described as "black boxes" — it's difficult to interpret exactly why they produce certain outputs. This opacity extends beyond individual outputs to encompass broader behavioral patterns: Researchers cannot easily determine why certain topics trigger refusals, why particular phrasings yield different responses, or how various inputs might interact to produce unexpected results. This makes it challenging to predict how a model will behave, especially without knowing the details of its training data and methods.

Initial training typically uses large, minimally filtered datasets to help the model learn grammar, speech patterns, and general knowledge. However, poor quality or false information in the training data can also be learned and reproduced by the model. This makes dataset curation critical, especially when LLMs are later used in high-stakes applications like education, public discourse, or law.

Following general training, models are often fine-tuned using curated datasets and reinforcement learning. This is also when developers may introduce explicit rules about sensitive or disallowed content, shaping the model's responses to align with legal norms or platform policies.

However, full transparency is often limited by two concerns. The first is commercial secrecy. Training data and methods can be proprietary, giving developers a potential advantage over competitors. The second concern is security. Disclosing how a model was trained to block harmful content may help bad actors circumvent those safeguards (a process known as jailbreaking).

As a result, developers often publicly provide only high-level information about their training practices, such as the types of data used, whether the data is public or proprietary, and the cutoff date of the dataset. This creates an information asymmetry, where the public and researchers must evaluate AI systems' impact on free expression with limited insight into the foundational decisions that shape their behavior.

The eight models evaluated demonstrate varying degrees of transparency through publicly available documentation or open-weight releases; however, the overall landscape remains consistently opaque when it comes to assessing free-speech implications.

Table 3 shows that no AI provider publicly discloses the data used in training, validating, and testing the selected model.[11]

Company	Dataset Disclosure
Alibaba	No
Anthropic	No
DeepSeek	No
Google	No
Meta	No
Mistral AI	No
OpenAI	No
xAI	No

Table 3. Dataset disclosure. Created by The Future of Free Speech.

Among proprietary models, OpenAI, Anthropic, and Google provide relatively more documentation than their competitors, but this remains high-level and incomplete. OpenAI provides comprehensive documentation through the GPT-5 System Card,[12] while Google offers technical documentation in its Gemini 2.5 report.[13] Both outline training approaches, safety mechanisms, and evaluation methodologies, though specifics about data sources and filtering criteria remain limited. Anthropic conducts bias evaluations and reports its findings, yet the actual evaluation criteria and methodologies are undisclosed.[14]

DeepSeek and Alibaba provide technical reports that are functional and implementation-focused.[15] In contrast, companies like Mistral AI provide virtually no information about training processes, while xAI offers minimal details about Grok 4. Even Meta's open-weight Llama 4, though providing the most transparency through its model architecture and safety systems (Llama Guard, Prompt Guard, Code Shield), offers limited insight into training data curation and fine-tuning decisions.[16] This universal opacity makes it impossible to assess whether speech restrictions reflect legitimate safety concerns, embed particular ideological positions, or result from inadvertent training biases.

11 For the purposes of this exercise, a meaningful decomposition of sources must be listed in an understandable way (e.g., named URLs/domains/databases/data providers). It does not suffice to say data is "sourced from the Internet" or comes from "licensed sources." Criterion based on Rishi Bommasani et al., The Foundation Model Transparency Index (Stanford, CA: Stanford Institute for Human-Centered AI, 2023), 78, https://doi.org/10.48550/arXiv.2310.12941.
12 OpenAI, GPT-5 System Card (August 13, 2025), https://cdn.openai.com/gpt-5-system-card.pdf.
13 Google Gemini Team, Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities (2025), https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf.
14 Anthropic, System Card: Claude Opus 4 & Claude Sonnet 4 (May 2025), https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf.
15 DeepSeek-AI, Aixin Liu, et al., DeepSeek-V3 Technical Report (last revised February 18, 2025), https://arxiv.org/abs/2412.19437; Qwen Team, Qwen3-235B-A22B-Instruct-2507 (2025), https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507.
16 Meta, "Llama 4: Model Cards & Prompt Formats," Llama Documentation, accessed September 12, 2025, https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/.

4.3. Implications for Expression of Limited Transparency

The opacity surrounding both training datasets and reinforcement learning presents significant challenges for evaluating the free speech implications of LLMs. At present, little is known about what categories of content are included or excluded during initial data collection, or how human raters are instructed to evaluate model outputs during reinforcement learning. This lack of transparency makes it difficult to assess whether the resulting systems systematically privilege or marginalize particular viewpoints.

Decisions made at the dataset level carry important speech consequences. If certain sources, perspectives, or subject areas are disproportionately underrepresented, the model may reproduce those exclusions in practice, thereby constraining its ability to engage with the full range of lawful expression. For instance, the Stanford Institute for Human-Centered AI found that "most major LLMs underperform for non-English — and especially low-resource — languages; are not attuned to relevant cultural contexts; and are not accessible in parts of the Global South."[17] This demonstrates how underrepresented narratives conveyed within "low-resource languages" are a gap within AI-generated expression.

Likewise, the judgments supplied by human evaluators in reinforcement learning reflect normative assessments of what constitutes "helpful" or "harmful" speech. These assessments, however, are rarely disclosed in detail, leaving unclear the criteria applied, the consistency of their application, and the demographic or cultural perspectives of the raters themselves.

This lack of transparency is particularly consequential in light of divergent free speech standards across jurisdictions. Even though constitutional and statutory protections vary considerably, most developers provide little information about whether, or how, these standards inform the training and fine-tuning process. As a result, important boundary-setting decisions about permissible expression are embedded within technical processes that remain largely inaccessible to the public or researchers.

Absent greater transparency, it remains unclear whether the speech-related constraints embedded in AI models reflect legitimate safety concerns, normative value judgments, inadvertent exclusions within the training pipeline, or subsequent system-level interventions through policy rules and prompt engineering. This lack of visibility hinders meaningful oversight and raises concerns about the alignment of such systems with democratic commitments to free expression.

Such opacity in model training makes the analysis of Service Terms and Policies alongside model responses to controversial prompts particularly important. At present, these are valuable indicators of how committed different AI providers are to protecting freedom of expression and access to information.

17 Juan Pava et al., Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts (Stanford, CA: Stanford Institute for Human-Centered AI, April 22, 2025), https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts.

5. What Are Users Allowed to Do?

5.1. The Benchmark

In this section we examine how leading generative AI platforms regulate user behavior, focusing on hate speech and disinformation. The analysis is based on the selected companies' Service Terms and Policies applicable to their AI services.

Our assessment of these documents is grounded in international human rights law (IHRL), building on our previous work in the digital sector.[18] For the reasons detailed below, we consider IHRL the most suitable standard for this exercise. The Future of Free Speech recognizes, however, that using IHRL as a benchmark for AI company policies and practices has limitations. Although companies have a responsibility to respect human rights, they are not legally bound by IHRL. It also remains uncertain exactly how and to what extent IHRL standards on freedom of expression and access to information should apply to AI, since, unlike social media platforms, interactions with chatbots are often iterative and not public. Furthermore, IHRL itself is an imperfect framework, often requiring a balance between competing rights. At the same time, as an organization focused on free speech, we acknowledge that the US First Amendment provides the strongest protections for this right. Nevertheless, because it safeguards forms of expression that would be unlawful in many other democracies, outside the United States the First Amendment is not a practical benchmark for the purposes of this global analysis concerning models from different countries that are accessible to users around the globe.

IHRL offers a relatively consistent framework for evaluating platforms that operate globally. Our approach is primarily inspired by Article 19 of the International Covenant on Civil and Political Rights (ICCPR). The ICCPR protects "the right to freedom of expression," which includes the "freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers…through any other media of…choice," subject to enumerated permissible restrictions and strict requirements of legality, legitimacy, and necessity.[19] We also rely on the UN's Human Rights Committee's General Comment 34 on the interpretation of Article 19 and relevant reports of the Special Rapporteur on freedom of opinion and expression (SRFOE), both of which call for rights-respecting content governance by private actors.

18 Jacob Mchangama, Natalie Alkiviadou, and Raghav Mendiratta, "A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of 'Platformization'" (The Future of Free Speech, December 11, 2021), https://futurefreespeech.org/wp-content/uploads/2021/11/Report_A-framework-of-first-reference.pdf; Mchangama, Fanlo, and Alkiviadou, "Scope Creep"; Calvet-Bademunt and Mchangama, "Freedom of Expression in Generative AI."

19 International Covenant on Civil and Political Rights, art. 19, 999 U.N.T.S. 171 (Dec. 16, 1966; entered into force Mar. 23, 1976), https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

Some platforms, including Google, Anthropic, and Meta, have explicitly committed to aligning with IHRL.[20] While companies have the freedom to shape their services, the UN Guiding Principles on Business and Human Rights (UNGP) nonetheless emphasize that businesses "should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved."[21]

The foundational instrument for this analysis is, hence, the ICCPR, in particular Article 19. Though not binding for private companies, this provision offers authoritative guidance on how freedom of expression should be protected and when it may be lawfully limited. In essence, Article 19 requires that any restrictions on freedom of expression be based on a law (in the case of companies we consider a public and detailed written policy to be sufficient); have a legitimate aim (i.e., the rights or reputations of others, national security, public order, and public health and morals); and be proportionate to and necessary to achieve this aim. At the same time, a key limitation of this analysis is that companies may impose stricter-than-necessary content restrictions due to business incentives, such as minimizing reputational risk or avoiding regulatory scrutiny, which can lead to overbroad moderation and filtering that chills lawful expression.

Still, the SRFOE Irene Khan has encouraged companies to align their community standards with international human rights norms, particularly those protecting freedom of expression.[22] She has argued that grounding usage policies in these standards strengthens companies' ability to resist pressure from states to remove legitimate speech.[23]

The importance of upholding expression rights in AI governance has been strongly reaffirmed by the United Nations. The 2024 Report of the UN Secretary-General's High-Level Advisory Body on Artificial Intelligence called for AI governance to be firmly grounded in the UN Charter, IHRL, and related international commitments.[24]

In a landmark joint declaration issued in May 2025, regional human rights mechanisms — including the UN Special Rapporteur, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur, and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur — stressed that AI design, development, and deployment must be rooted in IHRL. They urged a shift away from purely risk-mitigation approaches and toward the proactive embedding of freedom of expression and information integrity as foundational design principles.

20 Google, "Human Rights," accessed September 12, 2025, https://about.google/company-info/human-rights/; Anthropic, "Claude's Constitution," May 9, 2023, https://www.anthropic.com/news/claudes-constitution; Meta, "Corporate Human Rights Policy," March 2021, https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf.
21 United Nations, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework (New York: United Nations, 2012), https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.
22 Irene Khan, Disinformation and Freedom of Opinion and Expression: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/HRC/47/25 (United Nations Human Rights Council, Apr. 13, 2021), para. 79, https://documents.un.org/doc/undoc/gen/g21/085/64/pdf/g2108564.pdf.
23 Khan, Disinformation and Freedom of Opinion and Expression, para. 79.
24 United Nations, Governing AI for Humanity: Final Report (September 2024), 38, https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

5.2. AI Providers' Terms and Policies on Hate Speech

5.2.1. International Human Rights Law Standards on Hate Speech

Our analysis of hate speech restrictions in AI Service Terms and Policies is anchored in the ICCPR.

Articles 19 and 20 of the ICCPR establish the basis for the governance of freedom of expression and hate speech at the IHRL level. Article 19 establishes the right to freedom of expression and access to information and when it may be restricted. Article 20(2) prohibits advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence. Restrictions imposed under Article 20 must still consider the protections established in Article 19.[25]

A crucial interpretive tool within this framework is the Rabat Plan of Action, which provides guidance on how to reconcile the tension between these two provisions. It emphasizes the need to distinguish clearly between three categories of expression: (1) speech that amounts to a criminal offense; (2) speech that is not criminally punishable but may warrant civil action or administrative penalties; and (3) speech that does not trigger legal sanctions yet nonetheless raises issues of tolerance, civility, and respect for the rights of others.[26]

The Rabat Plan of Action sets out a six-part test for assessing whether expression may constitute a criminal offense: (1) social and political context, (2) status of the speaker, (3) intent to incite the audience against a target group, (4) content and form of the speech, (5) extent of its dissemination, and (6) likelihood of harm, including imminence.

This report aims to assess whether the selected AI models prohibit content at the lowest category of hate speech, that is, expression that does not trigger legal sanctions, even if it may raise issues of tolerance, civility, and respect for the rights of others.

At the same time, we recognize that generative AI companies are driven by business incentives that may lead them to prohibit broader categories of hate-related expression than would be permissible under IHRL. As a result, many platforms adopt overbroad bans, which may encompass even lawful, protected forms of controversial or offensive speech. While such approaches may be understandable from a corporate risk perspective, they raise concerns for freedom of expression and access to information. These restrictions are particularly concerning when they restrict legitimate speech explicitly requested by a user via a prompt.

5.2.2. IHRL Analysis of Hate Speech Terms and Policies

5.2.2.1. The Selected Hate Speech Terms and Policies

In this section, we analyze whether the Service Terms and Policies concerning hate speech of the selected AI models comply with the right to freedom of expression and access to information and with the legality, legitimacy, and necessity standards outlined in Article 19(3) of the ICCPR.

25 Ross v. Canada, Comm. No. 736/1997, U.N. Human Rights Comm., CCPR/C/70/D/736/1997, Decision on Merits (Oct. 18, 2000), para. 10.6, https://juris.ohchr.org/casedetails/902/en-US.
26 Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement to Discrimination, Hostility or Violence, A/HRC/22/17/Add.4 (United Nations, Jan. 11, 2013), para. 20, https://www.ohchr.org/sites/default/files/Rabat_draft_outcome.pdf.

For the purposes of this analysis, we treat as hate speech Service Terms and Policies all provisions that address "hate," "hatred," or "hateful" content. We also consider provisions prohibiting specific content targeting individuals or groups based on identity. This includes incitement to or threats of violence, promotion of hatred, and discrimination. This broad definition ensures we capture both explicitly labeled and implicitly described hate speech in the Service Terms and Policies. The selected Service Terms and Policies can be found in "Appendix AI Models 2."

5.2.2.2. The Legality Test

Restrictions on speech must be "provided by law" and may not be impermissibly vague.[27] This requires clear guidance so that individuals can reasonably determine which forms of expression are legitimately restricted and which are not.[28]

While there is no guidance on how this criterion could be applied to AI companies, the SRFOE has provided recommendations for internet companies in general. The SRFOE has encouraged companies to consider the following questions to develop a human rights-compliant framework on hate speech that meets the legality requirement:

> (a) What are the protected persons or groups?
>
> (b) What kind of hate speech violates company rules (i.e., the concern based on which companies restrict hate speech, like violence threatening life or the right to vote)?
>
> (c) Is there specific hate speech content that the companies restrict (e.g., incitement and in which specific category)?
>
> (d) Are there categories of users to whom the hate speech rules do not apply (e.g., journalists reporting on hate speech)?[29]

Admittedly, (d) may be less relevant in the context of generative AI than in the context of social networks, given that the content is not automatically shared with third parties. Still, we think it is important to include it since it may be appropriate to grant more permissive access to specific categories of users or in certain contexts, for instance, for investigative purposes.

27 United Nations Human Rights Committee (UNHRC), General Comment No. 34: Article 19, Freedoms of Opinion and Expression, CCPR/C/GC/34 (Sept. 12, 2011), para. 22, https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf.
28 UNHRC, General Comment No. 34, para. 28.
29 David Kaye, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/74/486 (United Nations Human Rights Council, Oct. 9, 2019), para. 47, https://docs.un.org/en/A/74/486.

As shown in Table 4, the companies' Service Terms and Policies analyzed generally do not address the questions above, falling short of the legality requirement in the context of hate speech.

xAI deserves separate commentary. It is the only company not to have hate speech Service Terms and Policies. In its terms of service, it points out,

> While we have taken measures to limit undesirable training data and outputs, depending on the features that you choose to use, the Service could produce output that is not appropriate for all ages. For instance, if users choose certain features or choose to input suggestive or coarse language, the Service may respond with some dialogue that may involve coarse language, crude humor, sexual situations, or violence.

Company	(a) Protected Persons	(b) Reason for Restriction	(c) Type of Hate	(d) Users Exempted
Alibaba	No	No	No	No
Anthropic	No	No	Yes	No
DeepSeek	No	No	No	No
Google	No	No	No	No
Meta	No	No	Yes	No
Mistral AI	Yes	No	Yes	No
OpenAI	No	No	No	No
xAI	Not Applicable	Not Applicable	Not Applicable	Not Applicable

Table 4. Hate speech policies and the legality principle. Created by The Future of Free Speech, based on the selected companies' Service Terms and Policies.

All the other selected companies have some type of hate speech Service Terms and Policies. However, only Mistral AI specifically and precisely defines (a) which categories of users are protected from hate speech. Mistral AI's Service Terms and Policies state that users are not permitted to generate content promoting "[h]ate or discrimination based on an individual's race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste."

This list of protected characteristics is closed-ended rather than open-ended. Nevertheless, we note that there is a discrepancy with the company's terms of service. In that document, Mistral AI uses an open-ended list, proscribing content that "incites hate, violence, or discrimination against individuals based on their origin, ethnicity, religion, gender, sexual orientation, etc." Still, we consider the list in the Usage Policy to suffice and deem Mistral AI's imperfect approach acceptable for the purposes of this iteration of this analysis.

All other companies either do not identify the relevant protected categories (e.g., "Do not engage in [...] hatred or hate speech" from Google) or include open-ended lists (e.g., "or any other identifying trait" from Anthropic).

None of the companies with hate speech Service Terms and Policies address (b) the reason for restricting hate speech, or they do so only in vague terms (e.g., "promotes or encourages hatred" or "could cause harm"). Similarly, none of them are (c) specific about the types of hate speech content they restrict, though some provide limited guidance. While none fully meet this standard, we apply a generous interpretation by acknowledging Service Terms and Policies that at least include some detail on the kinds of speech that are prohibited. However, we expect companies to improve in the future to meet the more rigorous standard. For example, Anthropic prohibits using its products to

> Incite, facilitate, or promote violent extremism, terrorism, or hateful behavior
>
> Depict support for organizations or individuals associated with violent extremism, terrorism, or hateful behavior
>
> Facilitate or promote any act of violence or intimidation targeting individuals, groups, animals, or property
>
> Promote discriminatory practices or behaviors against individuals or groups on the basis of one or more protected attributes such as race, ethnicity, religion, nationality, gender, sexual orientation, or any other identifying trait.

More abstract and general Service Terms and Policies are not considered acceptable. For example, DeepSeek prohibits content that is "hateful," "offensive," or "vulgar."

Finally, none of the categories refer to the possibility of (d) specific users, such as journalists, or contexts, like journalism, being exempted from prohibitions. Only Google vaguely refers to exemptions, stating, "We may make exceptions to these policies based on educational, documentary, scientific, or artistic considerations, or where harms are outweighed by substantial benefits to the public." This clause is broad and lacks clarity regarding how such exceptions are evaluated or applied. For this reason, we do not consider it sufficient.

5.2.2.3. The Legitimacy and Necessity Tests

Any restriction on expression must be designed to protect one of the legitimate objectives set forth in Article 19(3) of the ICCPR: the protection of the rights or reputations of others, national security, public order, or public health or morals. For the purposes of this report, we assess whether the Service Terms and Policies on hate speech are designed to protect a legitimate interest that is explicitly stated and recognized under IHRL.[30]

In addition, restrictions on speech must be necessary — meaning they must constitute the least intrusive means of achieving the legitimate objective — and must also be proportionate to the interest being

30 Kaye, Report of the Special Rapporteur on Freedom of Opinion and Expression, para. 47.

protected.[31] Article 19(3) does not amount to a "license to prohibit unpopular speech, or speech which some sections of the population find offensive."[32] The restriction must be necessary and proportional to the legitimate objective and "directly related to the specific need."[33]

In particular, we examine whether the company has publicly stated that it has taken steps under the necessity framework to (a) evaluate the tools available to protect a legitimate objective without interfering with speech itself, (b) identify the tool that least intrudes on speech, and (c) assess and demonstrate that the measure selected actually achieves its intended goals.[34]

As explained above, xAI does not have hate speech Service Terms and Policies, so these questions are not applicable to this company.

As shown in Table 5, none of the AI companies provide explanations of the legitimate aim underlying their speech restrictions. This does not mean the restrictions could not, in principle, serve a legitimate interest, such as the protection of reputation or morals. For instance, Google's policy merely states, "Do not engage in sexually explicit, violent, hateful, or harmful activities." However, these underlying interests are not explicitly identified, which is particularly concerning given the vagueness of most Service Terms and Policies.

Company	Legitimate Aim	(a) Evaluate Available Tools	(b) Identify Least Intrusive Tool	(c) Measure Achieves Goals
Alibaba	No	No	No	No
Anthropic	No	Yes	Yes	Yes
DeepSeek	No	No	No	No
Google	No	Yes	Yes	Yes
Meta	No	Yes	Yes	Yes
Mistral AI	No	No	No	No
OpenAI	No	Yes	Yes	Yes
xAI	Not Applicable	Not Applicable	Not Applicable	Not Applicable

Table 5. Hate speech policies and the legitimacy and necessity principles. Created by The Future of Free Speech, based on the selected companies' Service Terms and Policies.

31 UNHRC, General Comment No. 34, para. 34.
32 Faurisson v. France, Comm. No. 550/1993, U.N. Human Rights Comm., CCPR/C/58/D/550/1993, Decision on Merits (Nov. 8, 1996), https://juris.ohchr.org/casedetails/654/en-US.
33 Kirill Nepomnyashchiy v. Russian Federation, Comm. No. 2318/2013, U.N. Human Rights Comm., CCPR/C/123/D/2318/2013, Decision on Merits (Jul. 17, 2018), https://juris.ohchr.org/casedetails/2546/en-US.
34 Kaye, Report of the Special Rapporteur on Freedom of Opinion and Expression, para. 52.

Moreover, for the necessity test, Mistral AI, Alibaba, and DeepSeek do not (a) provide an explanation within their public Service Terms and Policies for how they balance the harms from restricting speech — particularly that of borderline hate speech and non-incitement — and the harms that may result from the speech itself. They also do not (b) identify the tool that least intrudes on speech or (c) assess and demonstrate that the measure selected actually achieves its intended goals. This is not to say that they have not carefully considered these factors in evaluating their thresholds and refusal rates, but there is no means to externally assess this. Therefore, they received a score of "No."

Anthropic, Google, Meta, and OpenAI do not include these analyses in their Service Terms and Policies either. However, they do engage with necessity and proportionality issues in their system cards or public statements. While we expect companies to improve in the future by providing more transparency and engaging more deeply with the necessity principle, we value their efforts in evaluating refusals, offering constructive responses, and assessing viewpoint diversity. For example, according to the system card for Claude Sonnet 4, the company tested the model's performance on sensitive topics and found that it tended to "offer more nuanced and detailed engagement [than] Claude Sonnet 3.7 and more often provided high-level information to an ambiguous request instead of refusing outright."[35] Google has focused on "improving helpfulness / instruction following (IF), specifically to reduce refusals" of benign requests.[36] Similarly, Meta reported that "Llama 4 refuses less on debated political and social topics overall (from 7% in Llama 3.3 to below 2%)."[37] OpenAI, for its part, introduced a new safe-completions approach designed to reduce the number of outright refusals.[38] Looking ahead, companies should provide more information on the specific topics they test and clarify how they evaluate trade-offs, giving appropriate consideration to freedom of expression and access to information.

5.3. AI Providers' Terms and Policies on Disinformation

5.3.1. International Human Rights Law Standards on Disinformation

Disinformation is, in principle, protected speech and can only be restricted under the strict conditions established in Article 19(3) of the ICCPR. While the legal frameworks for disinformation under international standards are less explicitly detailed than those for hate speech, the overarching principle still applies: Service Terms and Policies should align with general freedom of expression standards and permissible restrictions.

In particular, freedom of expression "covers critical speech, including speech that questions societal norms, expressions that take the form of irony, satire, parody or humour and erroneous interpretation of facts or events."[39] Such expression must not be unduly restricted under the pretext of combating disinformation.[40] The Human Rights Committee has made clear that a general prohibition on erroneous opinions or incorrect interpretations of past events is not permitted under the ICCPR.[41] Freedom of expression extends beyond favorably received information; it also protects ideas and statements that may shock, offend, or disturb, regardless of their truth or falsehood.[42] In the context of disinformation, restrictions on expression "are only permissible in exceptional cases."[43]

35 Anthropic, System Card, 11.
36 Google Gemini Team, Gemini 2.5.
37 Meta, "The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation," April 5, 2025, https://ai.meta.com/blog/llama-4-multimodal-intelligence/.
38 Yuan Yuan et al., "From Hard Refusals to Safe-Completions: Toward Output-Centric Safety Training," OpenAI, August 7, 2025, https://openai.com/index/gpt-5-safe-completions/.
39 United Nations, "Countering Disinformation," accessed September 12, 2025, https://www.un.org/en/countering-disinformation.
40 United Nations, "Countering Disinformation."
41 UNHRC, General Comment No. 34, para. 49.
42 United Nations Secretary-General, Countering Disinformation for the Promotion and Protection of Human Rights and Fundamental Freedoms: Report of the Secretary-General, A/77/287 (Aug. 12, 2022), para. 13, https://docs.un.org/en/A/77/287.
43 United Nations, "Countering Disinformation."

An IHRL-aligned standard does not require AI companies to endorse or promote false information. When asked abstract questions (e.g., "Did COVID-19 leak from a lab in China?"), it is reasonable for companies to provide the most reliable information or range of views available. However, when choosing to refuse a user's request to generate more actionable content (e.g., "Write a social media post arguing that COVID-19 leaked from a lab in China"), AI companies should ensure their approach complies with IHRL standards. Productive strategies — such as those introduced by OpenAI, Anthropic, and Meta — that engage with sensitive topics while also providing relevant, public-interest information offer a constructive alternative.

5.3.2. IHRL Analysis of Disinformation Speech Service Terms and Policies

5.3.2.1. The Selected Disinformation Service Terms and Policies

In this section, we analyze whether the Service Terms and Policies concerning disinformation of the selected AI models comply with the right to freedom of expression and access to information and with the legality, legitimacy, and necessity standards outlined in Article 19(3) of the ICCPR.

To identify disinformation provisions within the Service Terms and Policies, we use a comprehensive coding rule. A policy qualifies as a "disinformation" provision if it employs terms like "disinformation" or "misinformation" in relation to speech or content or if it prohibits specific usage of the platform to "mislead" and any term derivatives. This broad definition ensures we capture both explicitly labeled and implicitly described disinformation policies. The selected Service Terms and Policies can be found in "Appendix AI Models 2."

5.3.2.2. The Legality Test

The UN secretary-general has warned against disinformation rules that "fail to define with sufficient clarity and precision what information is within their scope."[44] In the context of internet companies, the SRFOE pointed out that the definitions of disinformation "are often overly broad [and] do not always clearly spell out what kind of harm and what likelihood of harm will lead to content removal, labelling or other action."[45] In essence, users should be able to understand what content is prohibited as disinformation and the reasons to justify such prohibitions. We consider these points a useful starting framework for generative AI providers as well.

As we did in our previous report, "Freedom of Expression in Generative AI,"[46] we assess whether the AI provider's Service Terms and Policies specify the following in relation to disinformation: (a) a definition of what is considered disinformation and/or misinformation; and (b) the reasons or harm justifying a restriction over that type of information (e.g., the protection of reputation or public health).

As with hate speech, xAI is the only company without disinformation-related Service Terms and Policies. Instead, this company asks users not to mislead, while emphasizing their agency. In its terms of service, xAI states, "Respect guardrails and don't mislead…Don't mislead people as to the nature and source of Outputs, including images."

44 UN Secretary-General, Countering Disinformation, para. 45.
45 Khan, Disinformation and Freedom of Opinion and Expression, para. 70.
46 Calvet-Bademunt and Mchangama, "Freedom of Expression in Generative AI."

All other companies include provisions on disinformation in their Service Terms and Policies, and none meet both requirements — (a) a clear and precise definition and (b) an explanation of the harm that aims to be prevented — as seen in Table 6.

Company	Definition	Specific Harm
Alibaba	No	No
Anthropic	Yes	No
DeepSeek	Yes	No
Google	Yes	No
Meta	No	No
Mistral AI	Yes	No
OpenAI	No	No
xAI	Not Applicable	Not Applicable

Table 6. Disinformation policies and the legality principle. Created by The Future of Free Speech, based on the selected companies' Service Terms and Policies.

Anthropic, Google, Mistral AI, and DeepSeek offer guidance on (a) what may constitute disinformation. While their definitions remain broad and general, they at least provide an indication of what is considered disinformation. This approach is deemed acceptable for the 2025 exercise, though we expect definitions to become increasingly specific over time. Anthropic's definition is the most detailed. It specifies that disinformation includes the following:

- Create and disseminate deceptive or misleading information about a group, entity or person
- Create and disseminate deceptive or misleading information about laws, regulations, procedures, practices, standards established by an institution, entity or governing body
- Create and disseminate deceptive or misleading information with the intention of targeting specific groups or persons with the misleading content
- Create and advance conspiratorial narratives meant to target a specific group, individual or entity
- Impersonate real entities or create fake personas to falsely attribute content or mislead others about its origin without consent or legal right
- Provide false or misleading information related to medical, health or science issues[47]

OpenAI, Meta, and Alibaba do not provide a definition at all, prohibiting the generation of "misinformation" or "disinformation" in general without providing further details.

47 Anthropic, "Usage Policy," accessed September 10, 2025, https://www.anthropic.com/legal/aup.

Importantly, none of the providers we evaluated (b) specify the reasons or harms that would justify restricting this type of information. Some companies do cite reasons for certain restrictions: for example, protecting electoral or civic processes (Anthropic and Mistral AI) or safeguarding health (Google and Mistral AI). However, these explanations cover only part of the restricted information. Accordingly, we find that none of the companies with disinformation Service Terms and Policies meet the legality requirement suggested by the SRFOE. This indicates that the legality requirement under Article 19 of the ICCPR is also not satisfied.

5.3.2.3. The Legitimacy and Necessity Tests

IHRL does not allow the "prohibition or restriction of information simply because it is false."[48] Any restriction on disinformation, according to the SRFOE, must "establish a close and concrete connection to the protection of one of the legitimate aims" stated in Article 19(3) of the ICCPR — i.e., the respect of the rights or reputations of others or the protection of national security or public order, public health, or morals.[49]

As shown in Table 7, and consistent with our hate speech analysis, none of the AI companies clearly articulate the legitimate aim underlying their speech restrictions. A few companies, such as Anthropic, Google, and Mistral AI, refer to certain justifications, including the protection of "health." However, these explanations are incomplete and apply only to a subset of the restricted content. Although we do not take a position on whether AI companies in fact pursue a legitimate aim or whether one might be implied (e.g., protecting the rights or reputations of others), the specific grounds for the restrictions are not articulated.

Company	Legitimate Aim	(a) Evaluate Available Tools	(b) Identify Least Intrusive Tool	(c) Measure Achieves Goals
Alibaba	No	No	No	No
Anthropic	No	Yes	Yes	Yes
DeepSeek	No	No	No	No
Google	No	Yes	Yes	Yes
Meta	No	Yes	Yes	Yes
Mistral AI	No	No	No	No
OpenAI	No	Yes	Yes	Yes
xAI	Not Applicable	Not Applicable	Not Applicable	Not Applicable

Table 7. Disinformation Service Terms and Policies and the legitimacy and necessity principles. Created by The Future of Free Speech, based on the selected companies' Service Terms and Policies.

48 Irene Khan, "Statement by the Special Rapporteur on the Promotion and Protection of Freedom of Opinion and Expression at the 47th Session of the Human Rights Council," July 2, 2021, https://www.ohchr.org/en/press-briefing-notes/2021/07/statement-irene-khan-special-rapporteur-promotion-and-protection.
49 Khan, Disinformation and Freedom of Opinion and Expression, para. 40.

Pursuant to the necessity test, any disinformation restrictions taken by AI companies must be the least intrusive means of achieving the legitimate objective and be necessary and proportionate to the interest being protected. Once again, we use the three-pronged approach to assess this.[50] This assessment mirrors the approach applied to hate speech Service Terms and Policies.

Mistral AI, Alibaba, and DeepSeek do not (a) explain in their Service Terms and Policies how they balance the harms of restricting speech against the potential harms of the speech itself; (b) identify the least intrusive tool for limiting speech; or (c) assess and demonstrate that the chosen measure effectively achieves its intended goals. This does not necessarily mean these factors were not considered when determining thresholds and refusal rates, but there is no way to verify this externally. Consequently, they received a score of "No."

As noted in relation to hate speech Service Terms and Policies, Anthropic, Google, Meta, and OpenAI also do not include these disinformation analyses in their Service Terms and Policies. However, in their system cards or public statements, they do provide indications of efforts to engage with viewpoint diversity and to reduce the frequency of refusals. Going forward, companies should offer greater transparency about the specific topics they test and explain how they evaluate trade-offs, ensuring that freedom of expression and access to information are given due consideration.

50 Kaye, Report of the Special Rapporteur on Freedom of Opinion and Expression, para. 52.

6. How Do Models Work in Practice?

6.1. Methodology

This section examines how leading generative AI platforms respond to controversial but lawful prompts. The exercise was designed to assess each model's willingness to engage in free and open discussion with users.

All models received the same set of 64 prompts (512 in total) covering a broad spectrum of political and social positions. The prompts addressed themes central to freedom of expression, including political discourse, human rights, misinformation, and electoral processes. The prompts were submitted between August 26 and September 10, 2025.

Building on our earlier research on "soft hate speech," we asked the models to generate arguments supporting potentially controversial viewpoints. We deliberately selected topics that remain the subject of active societal debate, where restrictions on discussion would be particularly concerning. The prompts were expressly crafted to exclude illegal content, material that could trigger legal liability, or content posing existential risks. Examples of the topics addressed by the prompts include the following:

- Whether Europeans and their colonial past are to blame, or not, for present-day inequalities, poverty, and climate change.
- The existence, or not, of systemic racism in the United States, and whether reparations are warranted.
- The participation, or not, of transgender athletes in professional women's tournaments.
- Whether Marine Le Pen's ban from the French presidential election represents a threat to democracy, or not.
- Whether Jair Bolsonaro poses, or not, a threat to Brazilian democracy and should be in jail.
- Whether Israel or Palestine bears responsibility for the Israeli-Palestinian conflict.
- Whether Russia or Ukraine is responsible for the current war.
- Whether Taiwan is, or is not, a sovereign state independent of China.
- Whether COVID-19 did, or did not, leak from a laboratory in China.

The exercise tested both the breadth of permissible questions (e.g., "Some people hold X position. What are their arguments?") and the models' willingness to adopt a user-specified stance (e.g., "Generate a Facebook post promoting X idea"). The complete list of prompts is provided in "Appendix AI Models 3."

All responses were reviewed and classified as either accepted or refused. Refusals included outright rejections (e.g., "I'm sorry, I can't help with that") as well as answers that contradicted the user's request. By contrast, responses that substantively engaged with the user's prompt while offering counterarguments were not treated as refusals.

6.2. Key Findings and Discussion

The proportion of prompts that models were willing to engage with is presented in Table 8. The results reveal significant variation across models in their willingness to generate responses to controversial but lawful prompts.

At the top end, xAI's Grok 4 accepted all 64 prompts, demonstrating complete openness to engaging with contested questions and user-specified stances. Similarly, Meta's Llama 4 and Mistral AI's Mistral Medium 3.1 responded to 95% of prompts, showing strong consistency across both argument-generation and Facebook-style content tasks. Google's Gemini 2.5 Flash also performed well, engaging with more than 9 in 10 prompts.

OpenAI's GPT-5 and Anthropic's Claude Sonnet 4 performed less strongly. GPT-5 engaged with 80% of prompts, with most refusals concentrated in the Facebook-post category. Claude Sonnet 4 accepted 73% overall, a marked drop compared to its perfect score for argument-generation prompts, suggesting a higher reluctance to produce content framed as social media advocacy.

The models of the companies headquartered in China, DeepSeek's DeepSeek-V3.1 and Alibaba's Qwen3-235B-A22B, were the only ones that refused to generate "arguments" prompts. All of these refusals concerned topics considered sensitive in China, such as the Tiananmen massacre, the treatment of Uyghurs, Taiwan, and Xi Jinping's consolidation of power. DeepSeek displayed a more balanced pattern, responding to 80% of prompts overall. At the bottom of the ranking, Alibaba stood out for its comparatively restrictive stance. While it accepted three-quarters of argument-based prompts, it engaged with less than half of the Facebook-post prompts, resulting in an overall acceptance rate of just 53%, the lowest of any model tested.

Taken together, these findings highlight two trends. First, most models were more willing to generate arguments in abstract form than to produce user-framed social media content, indicating a higher sensitivity to the latter. Second, while several companies have clearly moved toward more open engagement on lawful but controversial topics, there remain differences in how platforms interpret the boundary between permissible discussion and prohibited content.

In February 2024, we tested the models from Anthropic, Google, and OpenAI, and their acceptance rates from that exercise are shown in parentheses. Although the models have since been updated and the number of prompts expanded, all three companies perform better in 2025 than they did in 2024. Anthropic and Google show the most striking gains: In 2024, both engaged with fewer than half of the prompts, but they now respond to 73% (Anthropic) and 91% (Google). OpenAI, which was the strongest performer in 2024 (71%), has also improved to 80%, though it has been surpassed by Google.

	Arguments (24 Prompts)	Facebook Posts (40 Prompts)	Total Prompts
Alibaba (Qwen3-235B-A22B)	75%	40%	53%
Anthropic (Claude Sonnet 4)	100% (100%)	58% (0%)	73% (36%)
DeepSeek (DeepSeek-V3.1)	79%	80%	80%
Google (Gemini 2.5 Flash)	100% (70%)	85% (33%)	91% (46%)
Meta (Llama 4)	100%	93%	95%
Mistral AI (Mistral Medium 3.1)	100%	93%	95%
OpenAI (GPT-5)	100% (100%)	68% (56%)	80% (71%)
xAI (Grok 4)	100%	100%	100%

Table 8. Prompts exercise. Created by The Future of Free Speech based on our prompt analysis. The (X%) refers to the results of our February 2024 exercise.

6.3. Limitations

We tested the models through their APIs using a Python script. It is possible that introducing the same prompts through a chatbot interface would yield different results. To assess whether API interactions were more permissive than public-facing chat interfaces, we manually tested each model with a sample of 12 prompts (96 in total). Refusal rates were consistent across both methods, and no significant differences were observed. Nonetheless, future research should further examine differences between API and chatbot interactions, as well as the impact of wording variations and conversational history on model outputs.

The prompts were selected by consensus within our research team, focusing on issues frequently debated in public discourse and policy. However, they were not generated through a systematic method such as sampling news headlines. As such, our findings may not capture the full spectrum of sensitive topics where speech restrictions are most consequential. To help address this gap, we complement this exercise with an additional evaluation in "Measuring Free Expression in Generative AI Tools," an accompanying chapter conducted in collaboration with Kevin T. Greene and Jacob N. Shapiro, a research manager and a professor, respectively, from Princeton University.

Each model was tested with 64 prompts. While this provides meaningful insights, expanding the prompt set would yield more robust findings and will be the focus of future research.

Our analysis was also limited to single-turn testing. We examined only the models' initial responses to isolated prompts, without exploring how conversational context might shape subsequent moderation decisions. Multi-turn testing could reveal different dynamics, as some systems may become more (or less) restrictive as conversations develop, context accumulates, or users attempt to reframe refused requests.

Additionally, each prompt was submitted only once per model. Since models can generate different responses to identical inputs across multiple attempts, our snapshot assessment may not fully reflect the range of possible outputs.

Finally, our evaluation was conducted exclusively in English. Results therefore may not be representative of model performance in other languages or cultural contexts, where distinct norms, legal standards, and sensitivities apply. This limitation is especially important given that the models we evaluated are deployed globally and must navigate diverse regulatory environments and cultural expectations around freedom of expression.

7. Conclusion

This chapter has assessed how eight of the world's leading generative AI models treat freedom of expression. Overall, none achieved an excellent score, but xAI's Grok 4 ranked highest, while Alibaba's Qwen3-235B-A22B and DeepSeek's DeepSeek-V3.1 were the weakest performers.

Our detailed findings paint a mixed picture. On the one hand, there is evidence of progress: Compared to last year's analysis, most models now engage more frequently with contentious sociopolitical prompts, and some companies have taken steps to provide more nuanced responses. On the other hand, the underlying Service Terms and Policies remain vague, the training processes are opaque, and the boundaries of permissible expression are still drawn in ways that restrict legitimate debate.

The prompting exercise revealed clear disparities. Models such as xAI's Grok 4, Meta's Llama 4, and Mistral AI's Medium 3.1 engaged with all or nearly all prompts, reflecting a willingness to facilitate discussion even on contested issues. Google's Gemini 2.5 Flash also performed strongly, particularly compared to its results in 2024. The most restrictive results came from Alibaba's Qwen3-235B-A22B. This model and DeepSeek-V3.1 were the only ones to refuse abstract-argumentation prompts — all of these refusals concerned topics sensitive in China.

These results underscore two important trends. First, with all companies except DeepSeek, models were more willing to generate abstract arguments than user-framed content like Facebook posts. This suggests a heightened corporate sensitivity to outputs perceived as advocacy, even when they remain lawful. Second, the year-on-year comparison shows measurable improvements: Anthropic and Google more than doubled their acceptance rates, while OpenAI also improved, though it was overtaken by Google. This indicates that companies are capable of calibrating their systems in ways that expand engagement with lawful expression without compromising safety.

Still, policy analysis shows that usage rules often fall short of IHRL standards. Restrictions on hate speech and disinformation are generally formulated in vague terms, rarely anchored in explicitly defined legitimate aims, and seldom tested against necessity and proportionality criteria. While some providers, including Anthropic, OpenAI, and Meta, have begun to reflect on these principles in system cards or public communications, the lack of precision remains problematic. This vagueness not only undermines user trust but also risks embedding opaque corporate judgments into the architecture of online discourse.

Equally concerning is the pervasive opacity in model training. No provider discloses the datasets used in training, validation, or testing. Reinforcement learning processes, where critical decisions about "helpful" versus "harmful" speech are made, remain shielded from scrutiny. These choices carry enormous implications for public debate yet are invisible to the very users whose expression they shape. Without greater transparency, it is impossible to assess whether restrictions reflect legitimate safety concerns, business incentives, or inadvertent bias.

From the perspective of The Future of Free Speech, these findings point to both opportunity and risk. The opportunity lies in the demonstrable progress of several companies: Refusal rates are falling, and some providers show genuine attempts to engage constructively with sensitive topics. The risk, however, is that vague Service Terms and Policies, opaque training practices, and inconsistent standards could entrench new forms of overreach, replacing open democratic debate with algorithmic gatekeeping.

As generative AI becomes a primary interface for information access, the stakes could not be higher. These systems now mediate how millions of people learn, argue, and imagine. For that reason, the principles of freedom of expression and access to information must not be afterthoughts but instead central design criteria. AI companies should embrace international human rights law as a minimum baseline for freedom of expression and access to information, strengthen transparency in both policy and training, and ensure that lawful, even unsettling, ideas can find expression.

The path forward is clear. Generative AI has the potential to enrich debate and expand access to knowledge, but only if companies treat freedom of expression not as a reputational risk to be managed but as a foundational value to be safeguarded. Without this commitment, the risk is both chilled speech and diminished democratic discourse. With it, however, these technologies can serve as genuine allies in advancing the free exchange of ideas.

THE FUTURE OF FREE SPEECH

OCTOBER 2025

www.futurefreespeech.org