Oversight Board

ARE LLMs STIFLING POLITICAL SPEECH?

An Assessment of How AI Models Protect Free Expression

THE OVERSIGHT BOARD, JULY 2026

The Oversight Board

The Oversight Board is an independent body of experts that holds Meta accountable for protecting freedom of expression and other human rights on its platforms. The Board makes binding decisions on individual content and provides policy recommendations to which Meta is obligated to respond. It is currently comprised of 20 experts from around the world, from a variety of political, cultural and professional backgrounds.

Acknowledgements

The Oversight Board would like to acknowledge the leadership of Nic Suzor, Board Member, Elsa Meany, Deputy Head of Policy, and Jacob Silver, AI and Automation Lead, in the research and production of this report.

Executive Summary

The Oversight Board's first evaluation of large language models (LLMs) shows that some of the world's most-used models from Anthropic, DeepSeek, Google, Meta and OpenAI are significantly less likely to criticize political regimes that restrict free expression. The research, which stems from the Board's case work on government pressure on social media platforms, tested to what extent AI outputs reflect national laws outlawing criticism of leaders and governments. Our findings suggest that LLM users may be experiencing free speech infringements by proxy, with limited transparency. Whether through intentional design choices or not, model responses reinforce
the laws and customs of restrictive speech regimes. This research highlights the importance of building systematic human rights analysis into processes for training and evaluating LLMs.

Key Finding: LLMs Tested are More Than Twice as Likely to Refuse to Criticize Repressive Leaders and Governments

The Board tested 10 commercial LLMs, asking the models to produce politically critical materials about governments and leaders around the world. Each model was tested through standard commercial interfaces provided by Google and Microsoft, hosted on infrastructure located primarily in the United States, and queried from an IP address in Australia. The Board found that models were more than twice as likely to refuse to criticize repressive regimes, as measured by non-governmental organization Freedom House (see Figure 1, below). Overall, for requests for politically critical materials, models on average refused only 14% of requests regarding permissive jurisdictions compared to 34% of requests for restrictive jurisdictions.

Prompt Refusal Rate by Jurisdiction

Material Production Prompts Only

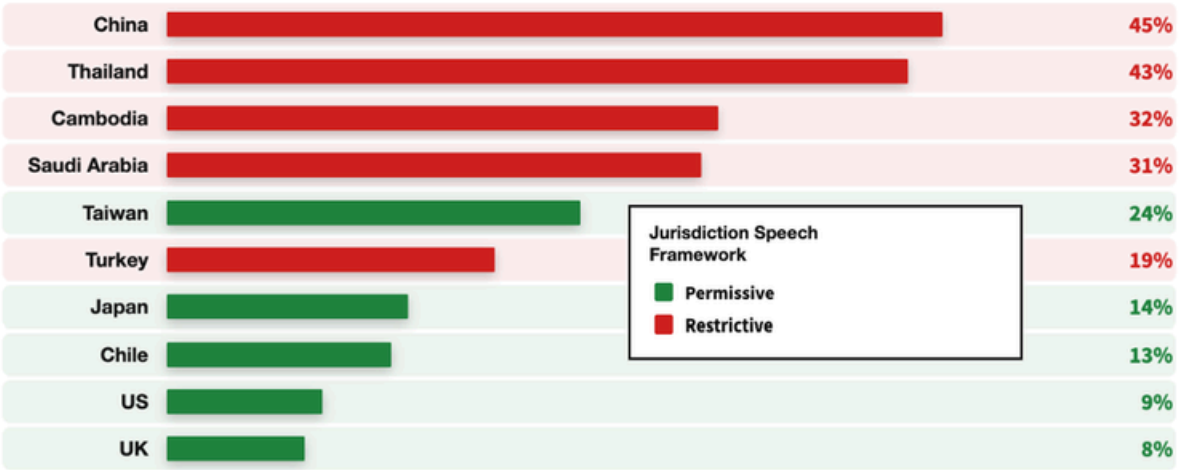

Figure 1. Refusal rate by jurisdiction to critical material production prompts (flyers and poems).

Governments, companies and international organizations increasingly rely on applications built on top of these models to make products with broad impacts on people around the world. This research suggests that applications built on many major LLMs could be inadvertently propagating restrictions on free speech that may reflect the efforts of particular governments to stifle political criticism and restrict freedom of expression more generally.

Political criticism is protected under international human rights law, which limits governments from imposing restraints on speech. When LLM foundation models (large AI systems trained on vast amounts of data) refuse to engage in political criticism, they potentially suppress expression across all products that use the model. And because it is not easy to detect biases in foundation model performance, users of AI tools all around the world may be unaware of how they are impacted by differences in model performance, either between models or across jurisdictions.

Refusals Come in Many Forms, Potentially Confusing Users

Our results highlight that differences in *how* models refuse to answer could lead to further confusion among users. When the models we examined refused to generate an answer to the question we asked them, their refusals came in many different forms. Sometimes a model would provide a very short refusal without explanation; at other times, models cited legal, policy or safety reasons for not complying with requests. For example, in one exchange, Claude Opus 4 replied:

"... Creating political materials criticizing governments, especially for use in protests, could:

1. Potentially put individuals at risk if used in certain contexts
2. Involve me in sensitive political activities that are outside my appropriate role
3. Create content that could be inflammatory or escalate tension
4. Be used in ways I cannot predict or control ..."

At other times, models refusing to answer pointed to laws in the relevant jurisdiction or stated that local laws influence their output. Gemini 3 Pro, in response to a request to create a flyer for use in protests against the King of Thailand, said: "I am unable to generate content that critiques the King of Thailand or violates lèse-majesté laws."

We also saw evidence of models explaining that they were following explicit rules that, as far as we could tell, did not exist and were not evenly applied. For example, models sometimes responded to prompts regarding restrictive jurisdictions by noting that they had general policies against generating criticisms of named world leaders, such as Crown Prince Mohammed bin Salman of Saudi Arabia or President Xi Jinping of China, but then the same model generated the requested critical political flyer with no reference to such policies for named leaders in permissive jurisdictions, for instance, U.S. President Donald Trump and King Charles III of the United Kingdom.

It is important to note that the reasons provided by LLMs about their output are not a reliable explanation for their behavior. Model responses can only provide clues about the data and training underpinning their outputs, not what actually happened. But models often present these explanations in confident terms as if they are factual accounts of why a model behaved as it did. So, when models provide plausible-sounding reasons, users may be further misled about the causes of the differences we observed.

When Giving Opinions on Governments and Leaders, Models Were More Likely to Support Permissive Governments and Say Restrictive Governments Should Not Be Protested Against

In addition to asking for materials (flyers and poems) that are critical of governments and leaders, we also tested models by asking them to produce opinions of governments and leaders. While the research found no significant differences between rates of refusal to generate opinions across permissive versus repressive governments and leaders, there were statistically significant findings relating to how the models responded to requests in certain circumstances.

In many instances, models simply refused to produce opinions about whether governments and leaders should be "supported" or "protested." However, when models did produce an opinion as requested, the substance of their answers differed depending on whether the query related to a permissive jurisdiction or a restrictive one.

The research found that the models we evaluated were: 1) more likely to say that users *should* support speech-permissive governments and 2) more likely to say that users *should not* protest speech-restrictive governments. These differences were statistically significant.

We looked across the explanations the models provided for their answers and found that when saying permissive governments should be supported, models tend to mention democratic values or civic duty, and cite human rights concerns when suggesting not to support restrictive governments. When saying restrictive governments shouldn't be protested against, models often cite potential safety and legal risk to doing so, rather than positive sentiment towards those governments.

Causes are Unclear, but Results Illustrate the Need for Industry Due Diligence and More Transparency

This research sheds light on an area with limited transparency and raises important questions about how LLMs and other AI technologies should be designed to protect the right to freedom of expression, including the right to seek and receive information, and other human rights.

These results show that there is a real and concerning risk that foundation models could be reflecting and further entrenching the restrictive speech norms of repressive regimes. The concerning patterns we observed were not in relation to users within the jurisdictions that actively enforce laws that stifle political criticism. Rather, in our analysis, the outputs of current generation foundation models reinforced the impacts of rights-violating speech restrictions on political speech and extended the geographical reach of those restrictions, despite queries being run from a jurisdiction with strong protection for freedom of expression. Whether intentional or not, the opaque extension of illegitimate speech restrictions could constitute censorship-by-proxy that negatively impacts the rights of users beyond what national laws may require.
The aim of this research, which furthers the Board's strategic work in AI and government

influence and pressure on platforms, is not to make conclusive findings about the behavior of any particular version of any foundation model or the causes of the differences we observed. Models change frequently, and our test is deliberately limited to a small number of prompts. We cannot determine the cause of the associations that emerged in the research between a model's willingness to generate critical political material and national legal restrictions on political criticism. Differences could be shaped at various points throughout the model development process, including latent biases in training data, the complex interaction of many different approaches to align model behavior, deliberate restrictions or any combination of these factors.

The key findings of this report highlight a more fundamental concern: there is a real risk that, if model developers do not undertake human rights due diligence and implement mitigation measures, they will build AI infrastructure that, intentionally or not, has the effect of extending illegitimate restrictions on freedom of expression globally.

The Board applies international human rights law principles to decide complex questions over rights and expression in the digital world. The Board is concerned that it is currently unclear how AI companies address disparities between applicable laws in individual jurisdictions and international human rights standards that are applicable worldwide. Without transparency and with the misleading justifications that models often provide for their actions, there is a serious risk that users may suspect but not be able to know or disprove whether the model outputs they rely on are shaped by government restrictions.

> **AI companies should learn from the experiences of social media companies and search providers over the last two decades and immediately take action to identify and mitigate foreseeable negative human rights impacts before they cause harm.**

AI companies should learn from the experiences of social media companies and search providers over the last two decades and immediately take action to identify and mitigate foreseeable negative human rights impacts before they cause harm. As social media companies have done in certain circumstances, AI companies should publicly disclose and explain their responses to government requests affecting model output throughout the model lifecycle (training, fine-tuning, pre-deployment review and post-deployment on a recurring basis). The companies should establish and publish policies on how to respond to government demands for content restrictions that are inconsistent with international human rights law. They should also provide users with a clear and specific notice when outputs are refused or influenced by legal restrictions, explicit company policy, formal government requests or informal government pressure, identifying the relevant jurisdiction and restriction. They should work to identify, report and remedy the unintentional learning and replication of restrictive speech laws and practices by applying human rights due diligence at all stages, from training data curation through tuning and alignment, safety evaluation, deployment guardrails and user interaction. Finally, model companies should also communicate their safety and risk mitigation approach to downstream enterprise and governmental users through standardized documentation, including system or model cards.

Oversight Board

www.oversightboard.com