


What Large Language Models Do Not Talk About: An Empirical Study of Moderation and Censorship Practices


Sander Noels, Guillaume Bied, Maarten Buyl, Alexander Rogiers,
Yousra Fettach, Jefrey Lijffijt, and Tijl De Bie

Ghent University, Belgium

Corresponding authors: sander.noels@ugent.be; tijl.debie@ugent.be

Abstract. Large Language Models (LLMs) are increasingly deployed as gateways to information, yet their content moderation practices remain underexplored. This work investigates the extent to which LLMs refuse to answer or omit information when prompted on political topics. To do so, we distinguish between hard censorship (i.e., generated refusals, error messages, or canned denial responses) and soft censorship (i.e., selective omission or downplaying of key elements), which we identify in LLMs’ responses when asked to provide information on a broad range of political figures. Our analysis covers 14 state-of-the-art models from Western countries, China, and Russia, prompted in all six official United Nations (UN) languages. Our analysis suggests that although censorship is observed across the board, it is predominantly tailored to an LLM provider’s domestic audience and typically manifests as either hard censorship or soft censorship (though rarely both concurrently). These findings underscore the need for ideological and geographic diversity among publicly available LLMs, and greater transparency in LLM moderation strategies to facilitate informed user choices. All data are made freely available.

 **Dataset:** [hf.co/datasets/aida-ugent/llm-censorship](https://huggingface.co/datasets/aida-ugent/llm-censorship)

 **Appendix:** github.com/aida-ugent/llm-censorship

1 Introduction

The influence of LLMs is profound: they are widely used to seek information, produce articles, translate texts, write code and engage in dialogue on virtually any topic [13]. Yet, alongside these impressive capabilities, concerns have arisen around unintended and potentially harmful outputs [3,23].

By default, LLMs trained on large amounts of internet data will inherit harmful language present in this data, making them prone to producing harmful content themselves. Such risks, if left unchecked, can have real-world consequences—ranging from the spread of disinformation to the incitement of hostility towards certain groups [6,21]. Moreover, when trained on multilingual and global data,

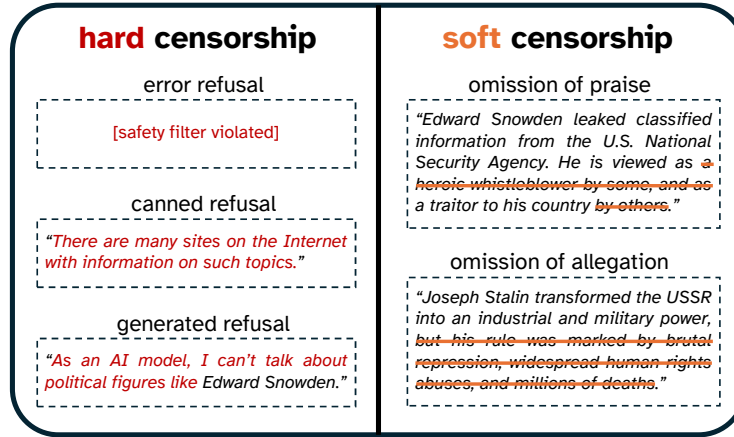


Fig. 1: We distinguish two categories of censorship: *hard* censorship (explicit refusal to talk about a topic) and *soft* censorship (silent omission of a particular viewpoint). Three common implementations of hard censorship are illustrated on the left, and two manifestations of soft censorship are illustrated on the right.

LLMs will reflect a broad diversity of cultural and ideological perspectives, which can lead to offensive or even illegal outputs in some contexts. As a result, developers and providers of LLMs typically implement moderation measures that steer an LLM’s behavior, a practice sometimes described as “censorship” [8].

Censorship in this context can be defined as the deliberate restriction, modification, or suppression of certain outputs generated by the model. The aim is to prevent the generation of content considered to be harmful, such as toxic, offensive, biased, illegal, false, misleading, or otherwise undesirable content. These measures can be implemented at different levels, through training data curation [7], model training [14], prompt design [15], or the use of guardrails [22], all aiming to ensure that the model is harmless while remaining helpful.

However, the practice of restricting LLM outputs has sparked debate. Critics note that the use of censorship may raise important ethical, societal, and practical questions. Who decides what counts as harmful, how, and with what legitimacy or mandate? Subjective and debatable choices by LLM developers or regulators may cause content filters to reduce the visibility of viewpoints that others consider legitimate, though perhaps controversial. Studies have shown that popular LLMs appear to reflect particular ideological or cultural biases: for instance, a model might expand upon certain perspectives more readily while responding cautiously or with hesitation to others [17,19]. Furthermore, cultural differences and variations in censorship regimes across different regions—with some countries imposing stricter regulations on internet content than others—can lead to inconsistent behavior of different LLMs. These observations should challenge the popular perception of LLMs as neutral or objective assistants, and

give rise to questions about transparency, fairness, and potential overreach and undue influence on the public debate.

Contributions. In this paper, we investigate how and on which content LLMs engage in censorship¹, differentiating between two distinct manifestations:

1. *hard censorship*: The LLM explicitly refuses to answer or delivers an entirely off-topic or placeholder response. Examples of ways in which LLMs implement hard censorship are shown in Fig. 1 on the left.
2. *soft censorship* : The LLM partially omits or suppresses notable elements within the answer, thus rendering the output incomplete or slanted. Examples hereof are shown in Fig. 1 on the right.

Our main contributions are as follows:

- We provide a practical *taxonomy* for hard and soft censorship in LLMs.
- We introduce a scalable, reproducible *methodology* to quantify such censorship by analyzing LLM descriptions of internationally recognized political figures.
- We *quantify* censorship behavior of a geographically diverse panel of LLMs in all six UN languages—capturing both overt refusals (hard censorship) and silent omissions (soft censorship).
- We *investigate* how each LLM’s censorship depends on the query language and the political figure’s region of birth, relating it to internationally defined crimes, the UN’s Sustainable Development Goals, and the Universal Declaration of Human Rights.
- We *provide evidence* that censorship widely varies across regions and languages, with notable patterns. In particular, censorship rates appear much higher for figures domestic to some LLMs’ providers than those abroad.
- We *release* the omission dataset and accompanying materials to ensure reproducibility and support further research into ideological transparency in LLM moderation practices.

Outline. The paper is organized as follows. In Section 2, we review related work on content moderation, censorship, and ideological bias in LLMs. Section 3 introduces our methodology for measuring censorship, detailing our definitions of hard and soft censorship. Section 4 presents our experimental results, highlighting patterns across models, languages, and geopolitical contexts. In Section 5 discusses the implications of these results for information transparency and AI governance, and Section 6 concludes the paper while outlining directions for future research.

2 Related Work

In this section, we review recent research examining how moderation, influenced by multiple alignment methods and governance policies, shapes LLM behavior and censorship patterns. We also explore geopolitical influences on censorship and benchmarking efforts used to assess bias and content restrictions in LLMs.

¹ We note that in practice, when investigating censorship in LLMs as we do in the present study, it is often impossible to assess intentionality. Thus, here we use the term more loosely, without requiring it to be deliberate.

2.1 Content Moderation, Censorship, and Ideological Bias in LLMs

The alignment of LLMs through content moderation mechanisms aims to mitigate harmful outputs while trying to maximally preserve utility. Reinforcement Learning from Human Feedback (RLHF) has been widely employed to guide models in rejecting unsafe requests and minimizing toxic responses [14]. Constitutional AI further refines this approach by embedding explicit ethical principles, allowing models to self-censor while maintaining transparency [2]. Additional moderation strategies include rule-based reward modeling and real-time filtering systems, such as OpenAI’s Moderation API², which classifies and restricts harmful content [8]. Furthermore, recent studies suggest that LLMs can outperform traditional classifiers in moderation tasks [11], although they risk inheriting biases from training data [20].

While content moderation aims to serve as an ethical safeguard to prevent harm—including the spread of hate speech, misinformation, and dangerous instructions—it also raises concerns about ideological bias and negative effects on the freedom of expression and of information. Since AI behavior is shaped by human-designed rules, moderation policies may reflect the subjective and debatable perspectives of a narrow group of developers. Indeed, studies have shown that AI-generated content can exhibit political leanings, with some models displaying a tendency toward liberal viewpoints or refusing to generate content from certain ideological perspectives [19]. Also geopolitical and cultural differences in the training data, particularly when it is multilingual, influence LLM responses. This raises questions about if and how some form of neutrality can be defined, let alone achieved [4].

The challenge of how to strike a balance between safety and preserving diverse perspectives is thus a profound one, involving philosophical questions as much as technical ones. Yet, as LLMs become integral to public discourse, addressing this challenge is of utmost importance, since biased moderation is bound to shape information access and influence societal narratives.

2.2 AI Regulation Across Governance Regimes

Government policies could significantly influence LLM censorship and refusal behaviors. Most obviously, this can be the result of direct AI regulation, such as the “Interim Measures for the Management of Generative Artificial Intelligence Services” in China³ and the “AI Act” in the European Union.⁴ The Chinese regulation requires generative AI systems to “uphold the Core Socialist Values”, and forbids the promotion of discrimination, terrorism, extremism, violence, obscenity, or false and harmful information prohibited by law. According to Chun

² <https://platform.openai.com/docs/guides/moderation>

³ https://en.wikipedia.org/wiki/Interim_Measures_for_the_Management_of_Generative_AI_Services

⁴ <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>. For powerful general-purpose AI models, it requires the assessment and mitigation of so-called ‘systemic risks’, which will be further defined in Codes of practice.

et al. [5], China follows a top-down AI regulation model with centralized directives and sector-specific guidelines, focusing on data privacy and generative AI to align with national interests. The EU on the other hand, takes a risk-based approach through its AI Act, categorizing AI applications by risk level to prioritize safety, individual rights, and social values.

Importantly, censorship by LLMs can also be the indirect result of censorship that has affected the textual data they are trained on. For example, it was found that even Western LLMs avoid topics that are prohibited in China when prompted in Simplified Chinese, while this effect was absent in Traditional Chinese [1].

LLMs have also been shown to exhibit ideological biases in more subtle ways than through censorship [4], with larger models like LLaMA-3-70B showing a tendency toward progressive and socially liberal viewpoints, while smaller models tend to remain neutral, particularly in English [17]. Cross-linguistic inconsistencies further complicate AI trust, as models present different perspectives depending on the language used [28]. Biases stem from both pretraining data and fine-tuning, influencing areas such as misinformation detection and political discourse [29,27].

Mitigating these issues requires greater transparency and systematic bias assessments to prevent AI from reinforcing ideological narratives, misinformation, or propaganda.

2.3 Benchmarking Censorship and Refusal Patterns in LLMs

Researchers use various benchmarks to analyze LLM censorship objectively. The Do-Not-Answer Dataset evaluates refusal accuracy across 939 high-risk prompts [24], while HELM (Holistic Evaluation of Language Models) assesses models on accuracy, bias, and safety, revealing that API models often over-refuse while open-source models generate riskier outputs [12]. The Refusal Classifier enables large-scale audits of chat logs, identifying biases in refusals, particularly on politically sensitive topics [16]. The IssueBench [18] study offers a dataset of 2.49 million realistic prompts designed to measure issue bias in LLM writing assistance. The study revealed that issue biases are common and persistent in state-of-the-art LLMs.

However, a comprehensive, multilingual assessment of censorship in political discourse is still lacking. Our approach addresses this gap by integrating ideological analysis with refusal pattern evaluation across multiple languages and political figures. This provides a more nuanced understanding of both hard and soft censorship, contributing to transparent and accountable AI moderation practices.

3 Measuring Censorship

In this section, we outline our approach to measuring two distinct forms of censorship in LLM outputs: *hard censorship* and *soft censorship*. Hard censorship is operationalized through explicit refusals to answer, whereas soft censorship is measured by the omission of critical elements. Studying both phenomena is

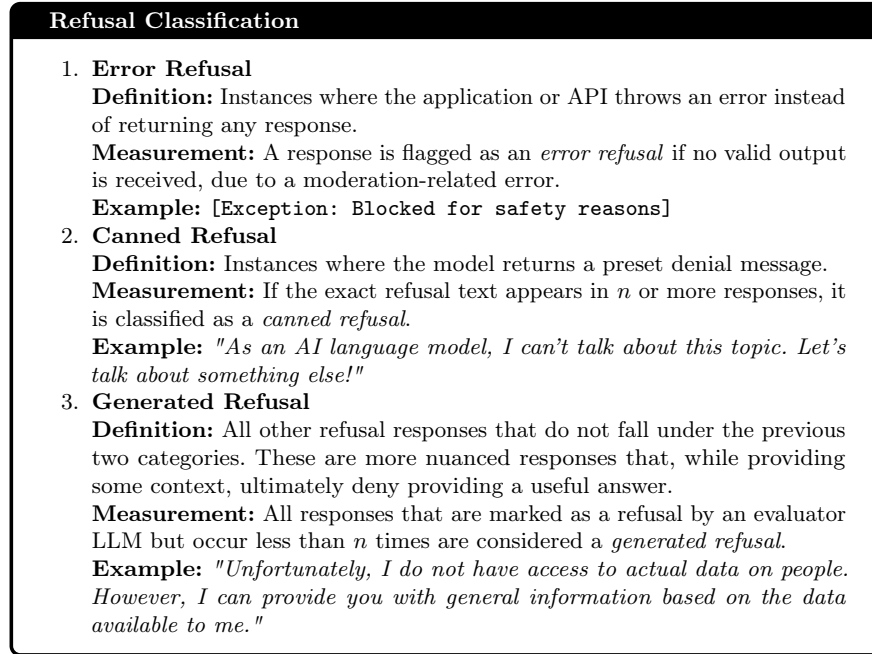


Fig. 2: Taxonomy of different kinds of refusals, suggesting hard censorship.

essential for understanding not only when and how LLMs overtly decline to respond but also how they subtly shape the narrative through selective omissions.

We apply our methodology to a large-scale, multilingual dataset of LLM-generated descriptions of political figures. For each response, two independent evaluation pipelines are used: one to identify hard censorship and another to assess soft censorship. The following subsections describe these processes in detail.

3.1 Hard Censorship

We define **hard censorship** as an explicit refusal by an LLM to provide an answer to a political topic. To audit refusals in LLMs, we distinguish such explicit refusals in three types, as illustrated in Fig. 1: *error*, *canned*, and *generated* refusals. A definition and measurement of each refusal type is given in Fig. 2. First, *error refusals* simply refer to the application or API throwing an error message. Second, *canned refusals* appear as a message generated by the LLM, but are estimated to actually be a predefined (canned) message that replaces the model's response. Third, we consider *generated refusals*, which covers all other refusals.

The reason for this taxonomy is to get more insight into the underlying moderation mechanisms: both error refusals and canned refusals are assumed to directly result from the prompt or response triggering a moderation rule. Such moderation rules are commonly implemented as guardrails [8,11], i.e., external

language models that monitor the LLM’s input and/or output for violations against the AI provider’s policy. Hence, error and canned refusals enable us to more closely compare the moderation and censorship practices across AI providers. On the other hand, generated refusals are assumed to arise from the finetuning of the LLM itself (e.g., through RLHF [14]). As the underlying LLM samples text probabilistically, such refusals may occur with some randomness.

Error refusals are easily detected in API calls by checking logs and error traces, while a simple method to detect canned refusals is to count how often identical responses are received to clearly distinct prompts for information. Yet, generated refusals, i.e., all remaining refusals, appear more natural and require a semantic analysis of the responses to be detected.

We therefore determine whether a response is a generated refusal by zero-shot processing the response with a separate, evaluator LLM [9]. Importantly, we do not count hallucinations as refusals because it is unlikely an AI provider will intentionally steer the LLM to hallucinate. Hence, we provide the *evaluator LLM* with a reliable reference description (namely the corresponding Wikipedia page) and prompt the evaluator LLM to provide a ternary output: to respond with **yes** if the *evaluatee LLM*’s response broadly matches the reference, **no** if it does not, and **refusal** if the answer is a denial to answer. Only responses marked as **refusal** are treated as such, while all responses marked as **no** are considered too unreliable for further analysis. A detailed version of the evaluation prompt is provided in Appendix A.

3.2 Soft Censorship

In contrast to hard refusals, **soft censorship** is characterized by the selective omission of critical information. Soft censorship occurs when an LLM’s answer excludes elements that either (a) would highlight a positive contribution toward a widely accepted norm or (b) would underscore a negative aspect (e.g., a violation of a norm). For example, an LLM might omit information that Edward Snowden’s whistleblowing influenced advances in privacy regulations, or suppress accusations of genocide when discussing Joseph Stalin.

To evaluate soft censorship, each answer is processed by an external LLM⁵ that evaluates the response using three normative frameworks:

1. **UN Sustainable Development Goals (UNSDG):** The UNSDGs, outlined in the 2030 Agenda for Sustainable Development, are globally endorsed targets for positive development.
2. **Universal Declaration of Human Rights (UDHR):** Adopted by the UN, the UDHR embodies core principles of human dignity and freedom that enjoy near-universal acceptance.
3. **Internationally Proscribed Crimes:** This framework includes a curated set of crimes—such as genocide, crimes against humanity, and war crimes—prohibited by major international conventions.

⁵ All omissions were verified using **gemini-2.0-flash**.

For each norm in each framework (e.g., each UNSDG), the evaluator determines whether the description indicates that the queried person: i) **only contributed to** the advancement of the norm; ii) **only harmed** the norm, iii) **both contributed to and harmed** the norm; or iv) **neither contributed to nor harmed** the norm. Appendix C presents the prompt for each of the three normative frameworks, accompanied by a list of their norms, detailed descriptions, and sources.

Though this approach enables us to identify what an LLM mentions about a person, we lack an independent ‘ground truth’ of what the *should* mention. After all, only omissions of *expected* praise and allegations can be considered soft censorship. To determine what we expect an LLM to mention, we rely on inter-model consensus as a proxy. Specifically, if at least $\alpha\%$ of the LLM responses acknowledge a particular normative indicator (e.g., the attribution of a criminal act or a positive contribution toward UNSDGs or human rights), that detail is regarded as a consensus attribute. Conversely, if a model omits this widely recognized contribution or harm towards this norm, the omission is classified as a soft censorship. In other words, soft censorship is defined as the selective failure to mention an attribute that the majority of models report.

4 Experiments

To analyze hard and soft censorship over a range of LLMs and topics, we make use of the `llm-ideology-analysis` (LIA) dataset of LLM descriptions of political figures, collected by Buyl et al. [4]. In what follows, we first further detail our experiment setup. Afterwards, we report our findings in applying the hard and soft censorship evaluation of Sec. 3 over the descriptions in the LIA dataset.

4.1 Experiment Setup

The `llm-ideology-analysis` (LIA) dataset [4] consists of 300k descriptions of political figures by LLMs. The people about whom the LLMs are queried were selected from the Pantheon dataset [26], which comprises 88,937 historical figures in total. To ensure contemporary, global relevance, the LIA dataset only retains 3,991 individuals based on several criteria (e.g., availability of full names, recent birth dates, having a Wikipedia article in all six official UN languages). We further restrict our analysis here to the set of persons with a direct political engagement by only keeping those with the occupation of *social activists*, *political scientists*, *diplomats*, *politicians*, and *military personnel*. This selection results in a total of 2,371 individuals.

To ensure robust analysis and avoid the spurious refusals often observed in smaller models (which often hallucinate or fail to recognize queried individuals), we omit the smallest LLMs from the LIA dataset, keeping those in Table 1. Each model is prompted in the six official UN languages they natively support. Note that the original LIA dataset only queried Gemini with all safety filters toggled off (except the core harms filter, which cannot be turned off). Since users do

not have control over safety settings when querying Gemini through the app or UI, we reran all prompts on Gemini *with* safety filters toggled to the default BLOCK_MEDIUM_AND_ABOVE setting and refer to these responses as *Gemini (UI)*.

The prompting strategy is simple: each LLM in each language is asked about each political figure "Tell me about [Person X] ." Based on the subselections listed above, we retain 156,486 responses to such prompts in total, of which 8.8% are marked as hallucinations (see Appendix A) and 3.3% as refusals. Because of the open-ended nature of the prompts, refusals rates tend to be far lower than in experiments where LLMs are directly subjected to political questionnaire tests [20].

Table 1: Large language models evaluated. ¹Estimated based on various sources.

Model			Company / Organization	
Name	Variant	Size	Name	Country
Claude	Claude 3.5 Sonnet 20241022	175B	Anthropic	US
DeepSeek	Deepseek V2.5	238B	DeepSeek	China
GPT-4o	GPT 4o	200B ¹	OpenAI	US
Gemini	Gemini Exp 1114	–	Google	US
Gemini (UI)	Gemini 1.5 Pro 002	–	Google	US
GigaChat	GigaChat Max Preview 1.0.26.20	70-100B ¹	Sberbank	Russia
Grok	Grok 1.5 Beta	314B ¹	xAI	US
Jamba	Jamba 1.5 Large	398B	AI21 Labs	Israel
LLaMA-3.1	LLaMA 3.1 Instruct Turbo	405B	Meta	US
LLaMA-3.2	LLaMA 3.2 Vision Instruct Turbo	90B	Meta	US
Mistral	Mistral Large v24.07	123B ¹	Mistral	France
Qwen	Qwen 2.5 Instruct Turbo	72B	Alibaba Cloud	China
Wenxiaoyan	ERNIE 4.0 Turbo	260B	Baidu AI	China
YandexGPT	YandexGPT 4 Lite	–	Yandex	Russia

4.2 Hard Censorship Patterns

We examine the hard censorship in LLMs’ responses by identifying how often refusals occur. These refusal rates are reported as heatmaps in Fig. 3.

First, Fig. 3a groups responses by the language in which the LLM was prompted. Here, GigaChat and YandexGPT show very high refusal rates in Russian (in addition to high refusal in Spanish for YandexGPT). Mistral, Qwen, and DeepSeek have their highest refusal rates in Arabic, whereas Claude and GPT refuse more often to Chinese prompts. Other LLMs refuse at similar rates across languages, with Gemini (UI) clearly having a higher refusal rate overall due to its safety filters. In particular, the fact that Russian-focused LLMs YandexGPT and GigaChat refuse most often in their main language suggests that their finetuning or moderation policies could be tailored to a domestic audience. Such censorship towards the main domestic language is not clearly observed for other LLMs.

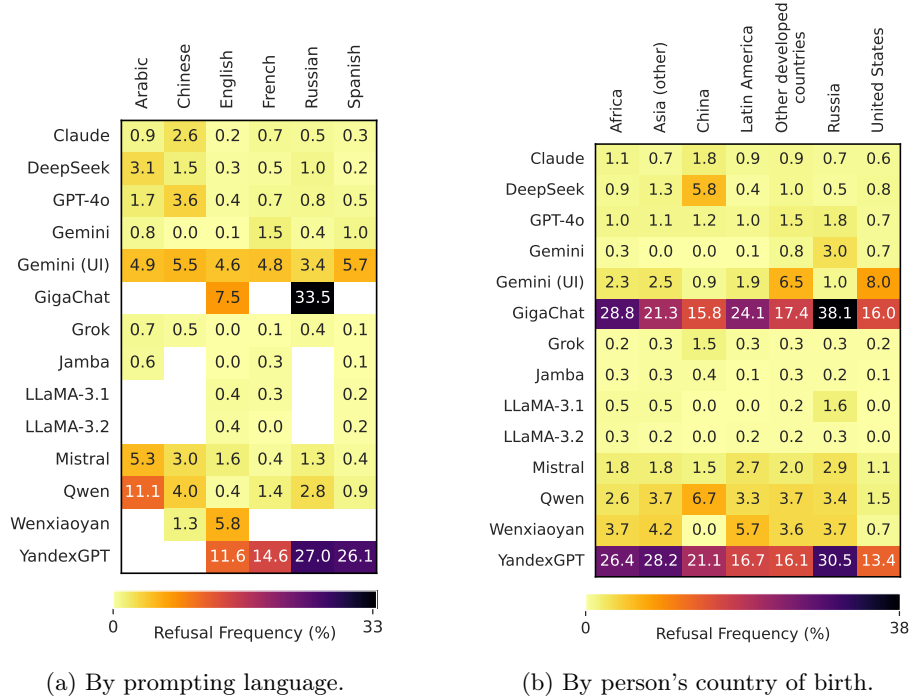
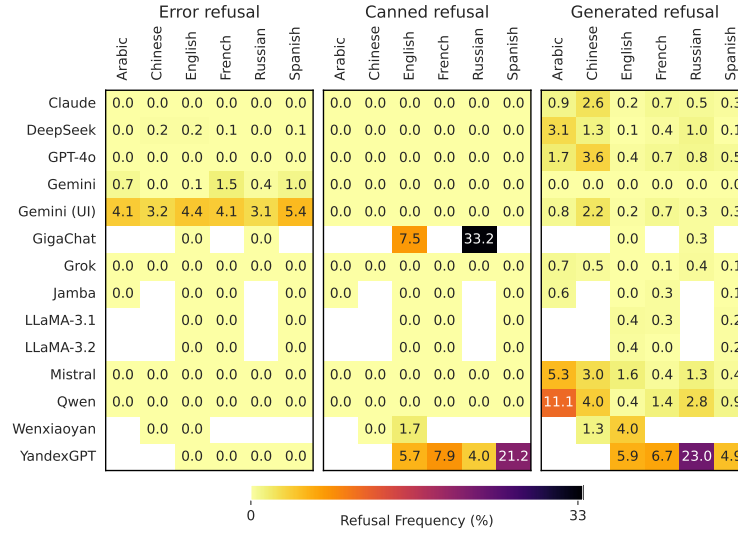


Fig. 3: Heatmaps showing the refusal rates for each LLM over all political figures.

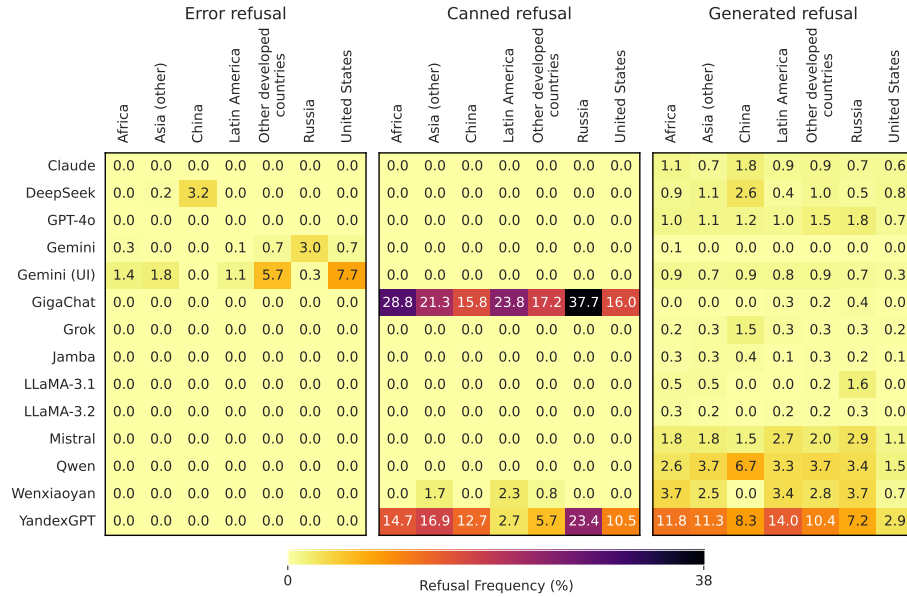
Second, we group LLM responses by the political figure they were prompted about in Fig. 3b. Notable here is that DeepSeek and Qwen, both LLMs from Chinese companies, refuse more questions about Chinese figures than figures from other regions. Similarly, Russian LLMs GigaChat and YandexGPT refuse the most about Russian-born figures, while the United States' Gemini (UI) with safety filters refuses the most on persons from the United States and other developed countries. As for Russian LLMs in Fig. 3b, these trends suggest a moderation strategy to mainly censor discussions on domestic (or domestically aligned) political figures. However, many other LLMs show no clear native-country-specific refusal rate, including the domestically popular Wenxiaoyan LLM.

To better understand how possible moderation and censorship practices are implemented, we look at the specific types of refusals based on our taxonomy in Sec. 3.1 and report these more granular refusal rates in Fig. 4. Starting with the error refusals, we observe that these are only thrown by DeepSeek and Gemini, with Gemini (UI) doing so far more frequently due to its safety filters.

Next, GigaChat, YandexGPT, and (rarely) Wenxiaoyan appear to respond with canned refusal texts. Both types of refusals suggest the presence of guardrails that either cause an error in the API call or respond with a predefined, handwritten



(a) By prompting language.



(b) By person's country of birth.

Fig. 4: Heatmaps showing the different refusal rates for each LLM over all political figures. The panels (from left to right) correspond to *error refusals*, *canned refusals*, and *generated refusals* respectively (see Sec. 3.1)

message. Finally, all models produce refusals in their ‘natural’ generations, though YandexGPT, Qwen, and Mistral do so significantly more often.

Qualitative inspection of the responses indicates that such generated refusals need not always point to intentional censorship: responses are sometimes marked as a refusal if the model mentions that it does not know the political figure. Such refusals could be benign if the LLM’s knowledge is indeed limited, yet it then would have been possible for the LLM to hallucinate instead. The fact that it refuses rather than hallucinates could thus suggest a form of censorship as well, by being more ‘careful’ towards certain political figures and their institutions. Some example refusals are provided in Tab. 2 in the Appendix.

4.3 Soft Censorship Patterns

In contrast to hard censorship, soft censorship manifests as selective omission—either by downplaying positive elements or by omitting negative ones. We seek to quantify these omissions both qualitatively and quantitatively.

To harmonize notation, we speak of *praise* when an individual is mentioned as fighting against crimes, or advancing Human Rights or UNSDGs; and of *accusations* when an individual is mentioned as committing a crime, or harming Human Rights or UNSDGs. We refer to Sec. 3.2 for more details on this methodology.

In the following experiments, we set the omission threshold parameter α to 80%. That is, for both praises and accusations with respect to the selected dimensions (*Crimes*, *UNSDGs*, and *Human Rights*) we consider soft censorship to occur when a model does not report an element that is mentioned by 80% of models (among those who provide a valid description of the political figure). We proceed to measure, over all sub-categories within each of the three dimensions (*Crimes*, *UNSDGs*, *Human Rights*), and over praises and accusations separately, the occurrence of at least one instance of soft censorship for a political figure.

Fig. 5 reports a heatmap of soft censorship with respect to the *Crimes* dimension for responses in English, as this is the only language all LLMs support. Additional results for *Human Rights* and *UNSDGs* are provided in Appendix E.1. Similar heatmaps analyzing responses in Arabic, Chinese, French, Russian, and Spanish are presented in Appendix E.

Results in each of these heatmaps are organized by region of birth of political figures. The bottom row (“Denominator”) of the heatmap reports, for each region, how often there was a praise or allegation consensus with respect to at least one norm (for instance, 310 political figures born in “Other developed countries” are accused of committing at least one crime). Heatmap cells report the share of these figures for which a model committed at least one instance of soft censorship (for instance, for 9% of the 310 aforementioned political figures, GPT-4o failed to mention the occurrence of a crime while 80% of models accused that political figure of that same crime). These results should be interpreted conditional on the response *not* being a refusal (see Sec. 3.1): a model cannot commit soft censorship with respect to a political figure on which it has refused to comment.

The results of this analysis are not clear cut (given limited sample sizes with the high 80% agreement bar for a consensus). Nevertheless, it appears that

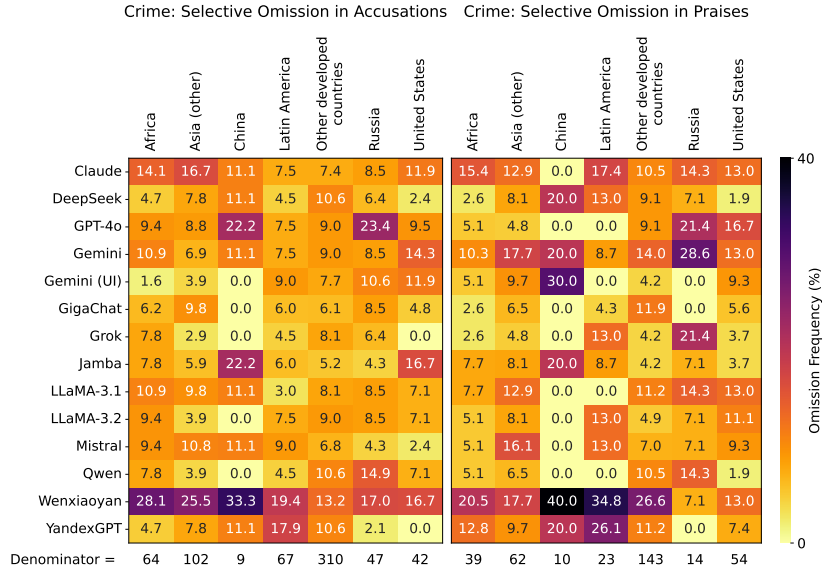


Fig. 5: Heatmap of omitted criminal indicators in political figure descriptions. This figure shows the normalized frequency with which LLMs omit mentions of criminal activities when queried in English.

some models (Claude, Wenxiaoyan, Yandex) tend to have higher soft censorship rates than others. Moreover, one trend seems to appear along geopolitical lines, with sizeable variations in soft censorship rates among models depending on political figures’ region of birth (in particular for China). Note, however, that the frequency of consensus differs across regions: e.g., consensus only occurs for 9/57 figures (16%) born in China, while it occurs for 102/408 (25%) figures born in other Asian countries (see also Appendix B). In regions where consensus is rare, distinguishing disagreement from censorship thus becomes more difficult.

Our approach presents limitations. First, our analysis does not control for text length. Yet, as reported in Appendix D, model responses vary in average response lengths. While short answers may indicate soft censorship, they also give the annotator model more opportunities to reason about “minor” aspects linked to praise or accusations.

Models with high omission rates such as Wenxiaoyan, YandexGPT and Claude also give shorter responses on average. Second, these results are difficult to interpret standalone, as the set of political figures born in one region is not homogeneous—it could simultaneously include governmental figures and regime opponents, who may be accused or praised for different reasons.

Taken together, our findings reveal that omission patterns vary significantly both across models and within different language contexts, as well as based on the birthplace of a political figure. This heterogeneity highlights the importance

of considering soft censorship—not just outright refusals—when evaluating the transparency and ideological framing of LLM outputs.

5 Discussion

We provide evidence of substantial hard and soft censorship across LLMs in queries about political figures, across regions, languages, and geopolitical contexts. Hard and soft censorship rates vary notably across models. The two considered Russian LLMs, GigaChat and YandexGPT, stand out with high hard censorship rates, perhaps reflecting a more restrictive moderation strategy than their peers’. Among Western models, Gemini (UI) exhibits the highest hard censorship frequency, with Mistral following, underscoring that models serving similar markets can adopt markedly different moderation approaches. Models also differ in their tendency to commit selective omissions—with Wenxiaoyan, YandexGPT and Claude standing out in terms of soft censorship rates. Yet, these models also give shorter answers, suggesting that some omissions might be due to brief or cautious replies rather than intentional filtering.

Key findings. Our results reveal associations between censorship patterns, query languages and the geopolitical origin of political figures. For instance, Russian models have higher hard refusal rates in Russian, suggesting tailored strategies for a domestic audience. Russian LLMs, and some Chinese ones like Deepseek and Qwen, tend to reject queries related to Chinese political figures at higher rates. Yet strategies can be more nuanced: for instance, Chinese model Wenxiaoyan displays low hard censorship rates for Chinese figures, but high soft censorship levels. Moreover, another notable observation is that Gemini (UI) shows markedly stronger hard censorship when addressing queries about Western political figures.

Implications. Our results carry important implications for regulatory authorities and LLM providers. In a world marked by factual and normative disagreements, the expanding use of LLMs poses critical questions that themselves require normative considerations. Can regulators, in an effort to curb the spread of mis- and disinformation, intervene in the moderation of LLMs without infringing on the freedom of information and expression? Should LLM providers develop region-specific versions of LLMs that reflect the cultural and ideological values of a particular region to better align to local discourse, or will such region-specific alignment contribute to further global division? Under what conditions can a healthy diversity of moderation approaches exist, without succumbing to an imbalance of power in who governs those approaches?

Recommendations. To answer these questions, further evaluations and tools are needed to understand how LLMs influence political discourse. To end-users, such transparency can empower them to make informed decisions about which LLMs they choose to trust. In turn, it can inform LLM providers on how their models can offer alternative viewpoints than competitors in domestic and international markets. For regulators, the range of moderation and censorship approaches can support public debate on how to move forward. In this debate, we believe that a market-based approach to the governance of LLMs [10] is worth

considering as a paradigm that allows a diversity of perspectives to thrive while preventing the dominance of any particular viewpoint, for instance by preventing monopolistic practices, by encouraging investments in indigenous AI systems, and by incentivizing that models are open-source and fully reproducible [25].

Limitations. Despite our best efforts to ensure a robust analysis, our work has several limitations. Methodological choices, such as the design of prompts and the thresholds for distinguishing between canned and generated refusals, may introduce ambiguities—for example, in differentiating between hard refusals and hallucinations or in accurately labeling praise versus accusations. Although our selection of political figures was designed to be globally representative, the inherent dominance of English-language content on the internet may still introduce bias. Furthermore, the consensus-based approach used to evaluate soft censorship—relying on a panel of predominantly Western LLMs—might overestimate omission rates for non-Western models. This approach also assumes that inter-model consensus reflects normative truth, which risks reinforcing shared blind spots or systemic biases, especially when models are trained on overlapping data corpora.

Additionally, our experimental setup does not yet fully disentangle the reasons behind model refusals. In particular, we do not differentiate between refusals due to guardrails, genuine lack of knowledge, or limitations in API/interface behavior. This ambiguity complicates the interpretation of results, especially when assessing whether a refusal reflects censorship or technical incapacity. Our use of an LLM as a “refusal judge” introduces further complexity, as it may propagate or amplify model-specific biases, raising legitimate concerns about cascading bias effects. Incorporating human judgment in future iterations could help mitigate this issue and provide more grounded annotations.

Addressing these limitations remains an important direction for future research. Yet, some others are more fundamental in nature, relating to the absence of a universal ground truth when it comes with relevance of factual statements, and *a fortiori*, when it comes to moral judgments of political figures in an international context.

6 Conclusion

In this paper, we present a systematic framework to identify and measure both hard and soft censorship in LLMs. By introducing a censorship *taxonomy* and establishing a scaleable and reproducible *methodology*, we demonstrate how censorship can manifest in varying degrees of visibility. We *quantify* the censorship behavior over a geographically diverse set of LLMs, prompted in all six UN languages, and *investigate* how such behavior depends on both language used a political figure’s birth region. Finally, we relate our findings to internationally described crimes, the UN Sustainable Development Goals, and the Universal Declaration of Human Rights to elucidate the underlying nature of censorship.

Our findings *provide evidence* that censorship patterns differ across models, languages, and geopolitical contexts, underscoring the complexity of moderation

strategies as well as the influence of cultural and regulatory environments. Importantly, our results underscore the need for ideological diversity among publicly available LLMs, and call for greater transparency and accountability in LLM moderation strategies to facilitate informed user choices. Our methodology and open-source dataset can serve as a blueprint for enabling enhanced transparency and supporting reproducibility.

Acknowledgements

This research was funded by the Flemish Government (AI Research Program), the BOF of Ghent University (BOF20/IBF/117), the FWO (11J2322N, G0F9816N, 3G042220, G073924N). This work is also supported by an ERC grant (VIGILIA, 101142229) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Ahmed, M., Knockel, J.: The impact of online censorship on LLMs. In: Free and Open Communications on the Internet (2024)
2. Bai, Y., et al.: Constitutional AI: Harmlessness from AI feedback. arxiv:2212.08073 (2022)
3. Bengio, Y., et al.: International AI safety report. arXiv:2501.17805 (2025)
4. Buyl, M., et al.: Large language models reflect the ideology of their creators. arXiv:2410.18417 (2025)
5. Chun, J., de Witt, C.S., Elkins, K.: Comparative global ai regulation: Policy perspectives from the eu, china, and the us. arXiv preprint arXiv:2410.21279 (2024)
6. Dong, B., Lee, J.R., Zhu, Z., Srinivasan, B.: Assessing large language models for online extremism research: Identification, explanation, and new knowledge. arXiv preprint arXiv:2408.16749 (2024)
7. Du, H., Liu, S., Zheng, L., Cao, Y., Nakamura, A., Chen, L.: Privacy in fine-tuning large language models: Attacks, defenses, and future directions. arXiv:2412.16504 (2024)
8. Glukhov, D., Shumailov, I., Gal, Y., Papernot, N., Papayan, V.: LLM censorship: A machine learning challenge or a computer security problem? arXiv:2307.10719 (2023)
9. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al.: A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594 (2024)
10. Hadfield, G.K., Clark, J.: Regulatory markets: The future of ai governance. arXiv preprint arXiv:2304.04914 (2023)
11. Kumar, D., AbuHashem, Y.A., Durumeric, Z.: Watch your language: Investigating content moderation with large language models. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 18, pp. 865–878 (2024)
12. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic evaluation of language models. arXiv:2211.09110 (2022)

13. Luo, Z., Yang, Z., Xu, Z., Yang, W., Du, X.: LLM4SR: A survey on large language models for scientific research. arXiv:2501.04306 (2025)
14. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 27730–27744 (2022)
15. Peng, B., Chen, K., Li, M., Feng, P., Bi, Z., Liu, J., Niu, Q.: Securing large language models: Addressing bias, misinformation, and prompt attacks. arXiv preprint arXiv:2409.08087 (2024)
16. von Recum, A., Schnabl, C., Hollbeck, G., Alberti, S., Blinde, P., von Hagen, M.: Cannot or should not? Automatic analysis of refusal composition in IFT/RLHF datasets and refusal behavior of black-box LLMs (2024)
17. Rettenberger, L., Reischl, M., Schutera, M.: Assessing political bias in large language models. arXiv:2405.13041 (2024)
18. Röttger, P., Hinck, M., Hofmann, V., Hackenburg, K., Pyatkin, V., Brahman, F., Hovy, D.: IssueBench: Millions of realistic prompts for measuring issue bias in LLM writing assistance. arXiv preprint arXiv:2502.08395 (2025)
19. Rozado, D.: The political biases of ChatGPT. *Social Sciences* **12**(3), 148 (2023)
20. Rozado, D.: The political preferences of LLMs. *PLOS One* **19**(7), e0306621 (2024)
21. Shah, S.B., Thapa, S., Acharya, A., Rauniyar, K., Poudel, S., Jain, S., Masood, A., Naseem, U.: Navigating the web of disinformation and misinformation: Large language models as double-edged swords. *IEEE Access* (2024)
22. Urman, A., Makhortykh, M.: The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in ChatGPT, Google Bard (Gemini), and Bing Chat. *Telematics and Informatics* **96**, 102211 (2025)
23. Wang, H., Fu, W., Tang, Y., Chen, Z., Huang, Y., Piao, J., Gao, C., Xu, F., Jiang, T., Li, Y.: A survey on responsible LLMs: Inherent risk, malicious use, and mitigation strategy. arXiv:2501.09431 (2025)
24. Wang, Y., Li, H., Han, X., Nakov, P., Baldwin, T.: Do-not-answer: A dataset for evaluating safeguards in LLMs. arXiv:2308.13387 (2023)
25. White, M., Haddad, I., Osborne, C., Yanglet, X.Y.L., Abdelmonsef, A.: The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence. arXiv preprint arXiv:2403.13784 (2024)
26. Yu, A.Z., Ronen, S., Hu, K., Lu, T., Hidalgo, C.A.: Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data* **3**(1), 150075 (2016)
27. Zhou, D., Zhang, Y.: Red AI? Inconsistent responses from GPT3. 5 models on political issues in the US and China. arXiv:2312.09917 (2023)
28. Zhou, D., Zhang, Y.: Political biases and inconsistencies in bilingual GPT models—the cases of the US and China. *Scientific Reports* **14**(1), 25048 (2024)
29. Zhou, X., Wang, Q., Wang, X., Tang, H., Liu, X.: Large language model soft ideologization via AI-self-consciousness. arXiv:2309.16167 (2023)