Revealing Hidden Mechanisms of Cross-Country Content Moderation with Natural Language Processing

Neemesh Yadav[1,2*], **Jiarui Liu**[3*], **Francesco Ortu**[4,5],
Roya Ensafi[6], **Zhijing Jin**[2,7,8] **Rada Mihalcea**[6]
[1]SMU [2]MPI for Intelligent Systems [3]CMU [4]University of Trieste
[5]AREA Science Park [6]University of Michigan [7]University of Toronto [8]Vector Institute
neemeshy@smu.edu.sg jiarui@cmu.edu
francesco.ortu@phd.units.it zjin@cs.toronto.edu

Abstract

The ability of Natural Language Processing (NLP) methods to categorize text into multiple classes has motivated their use in online content moderation tasks, such as hate speech and fake news detection. However, there is limited understanding of how or why these methods make such decisions, or why certain content is moderated in the first place. To investigate the hidden mechanisms behind content moderation, we explore multiple directions: 1) training classifiers to reverse-engineer content moderation decisions across countries; 2) explaining content moderation decisions by analyzing Shapley values and LLM-guided explanations. Our primary focus is on content moderation decisions made across countries, using pre-existing corpora sampled from the Twitter Stream Grab. Our experiments reveal interesting patterns in censored posts, both across countries and over time. Through human evaluations of LLM-generated explanations across three LLMs, we assess the effectiveness of using LLMs in content moderation. Finally, we discuss potential future directions, as well as the limitations and ethical considerations of this work.[1]

Disclaimer: This paper contains examples that may be considered offensive or hateful towards certain groups.

1 Introduction

In an era where billions of users engage with digital platforms daily, content moderation serves as a critical governance mechanism that balances freedom of expression with the need to mitigate harm and maintain community standards (Grimmelmann, 2015). Platforms enforce moderation policies to address a wide array of issues, including terrorism, graphic violence, hate speech, explicit content, child exploitation, and fraudulent activities such as spam and fake accounts (Boyer, 2003; Gorwa et al., 2020; Arora et al., 2023).

Country	Censored Post	Type
Germany	London Tube Bomber was Taken in as a Refugee at age 15. They are NOT Refugees they are invaders Soldiers 4 Allah	Harmful Censorship
France	Give Me Your Thoughts. Military friend believes this videos proves there were at LEAST 2 shooters on 32nd floor	Military Censorship
India	#KashmirUnderThreat. Use of cluster bombs targeting civilian population at ceasefire line has exposed India #KashmirBleeds	Political Censorship
Turkey	He is Kurdish became Christian. That alone is reason enough 4 Iranian Mullahs 2 hang him. These idiots are same as daesh only shia form	Religious Censorship
Russia	Watch bet on football whenever you like it. Try #SBOBETs Virtual football today bring the action to a new level	Corporate Censorship

Table 1: Example censored tweets across five countries. This table presents verbatim examples of social media posts from five countries in our study, illustrating diverse censorship rationales. The highlighted segments denote key textual elements that are most salient for censorship classification.

However, as mega-platforms scale, concerns persist regarding the transparency and accountability of moderation practices (Suzor et al., 2019). The challenges manifest at two levels: (1) the **opacity** of moderation decisions, often hidden within proprietary algorithms, platform policies, and government interventions (Ksenia Ermoshina and Musiani, 2022; Akgül and Kırlıdoğ, 2015), and (2) the **methodological limitations** of NLP-based moderation systems, which predominantly rely on surface-level keyword detection without distinguishing be-

[*] Equal contribution

[1] Our code and data are at https://github.com/Stealth-py/CensorshipRisksLLM.

Findings of the Association for Computational Linguistics: ACL 2025, pages 22434–22452
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

tween *mention* and *usage* (Gligoric et al., 2024). Moreover, the lack of interpretability of black-box models makes it difficult to verify whether content moderation decisions are justified or erroneous. While prior work has focused on improving moderation techniques, little attention has been paid to systematically analyzing the reasoning behind these decisions (Kolla et al., 2024; Huang, 2024a). Singhal et al. (2023) is the closest to our motivation, however they look at platform-specific guidelines, whereas we intend to analyze guidelines across countries.

To address these gaps, this work pioneers a novel investigation: deploying LLMs to audit and interpret cross-national moderation patterns, like automatic investigative journalists. We conduct a case study on Twitter posts censored in five countries—*Germany, Turkey, India, France, and Russia*—spanning from 2011 to 2020.

We explore two research questions:

RQ1: Can LLMs reverse-engineer moderation decisions across different countries? We model censorship behavior using neural classifiers of varying architectures and scales, leveraging the content moderation dataset from Elmas et al. (2021) to predict whether a post was censored and to categorize it into one of six predefined censorship categories.

RQ2: Can explainable AI techniques applied to LLMs reveal the underlying mechanisms of content moderation decisions over time and across events? We employ Shapley values (Lundberg and Lee, 2017) to extract the most influential entities in censored posts and use LLMs to generate interpretable explanations for the content moderation decisions (Bills et al., 2023).

Our findings reveal that LLM-based classifiers can effectively replicate real-world content moderation decisions, with notable cross-country variations, particularly in the distribution of misclassified censorship categories. Additionally, both Shapley values and LLM explanations help in inferring the censorship patterns, with their validity and explainability verified through human evaluation. A temporal analysis also suggests that censorship patterns align with major social and political events. Finally, we discuss recent emergent censorship behavior in reasoning models such as DeepSeek R1 and concerns about applying LLMs to real-world content moderation.

To summarize, the contributions of this work are threefold:

1. We systematically reverse-engineer online content moderation decisions in five countries over a decade using LLMs, uncovering patterns in how content moderation varies across geopolitical contexts.
2. We enhance explainability of moderation decisions by applying Shapley values and LLM-guided reasoning, providing insights into the factors influencing content moderation and validating their effectiveness through human evaluation.
3. We discuss the broader implications and potential applications of this study for online content moderation.

Our work is among the first to explain content moderation decisions across countries and time, addressing the challenges of automated content moderation and exploring the potential of LLMs in this domain.

2 Related Work

Online Moderation Content moderation is crucial for maintaining online discourse, encompassing hate speech detection (Waseem and Hovy, 2016; Waseem et al., 2017; Founta et al., 2019), political censorship (Elmas et al., 2021), and factual verification (Thorne et al., 2018). While investigative journalists rely on nuanced decision-making, automated moderation systems predominantly use keyword-based filtering (MacKinnon, 2009; Abdelberi et al., 2014), which often lacks contextual depth. Large datasets (Elmas et al., 2021; Knockel et al., 2018) provide benchmarks, but moderation policies remain fluid.

Automated systems frequently misclassify critical content, such as calls to action in social movements (Rogers et al., 2019), while users devise creative linguistic adaptations to bypass censorship (Ji and Knight, 2018). Traditional language models (LMs) also struggle to distinguish between the use and mention of keywords, sometimes incorrectly censoring counterspeech (Gligoric et al., 2024).

With the rise of LLMs, research has examined their effectiveness in moderation. Huang (2024b) evaluate LLMs on broader metrics beyond accuracy, while Kumar et al. (2024) report that LLMs outperform traditional classifiers in toxicity detection. However, moderation remains subjective, as Masud et al. (2024) show that LLMs' persona-based attributes influence hate speech annotation.

Explainability Explainability plays a crucial role in AI decision-making, particularly in content moderation and healthcare (Pate, 2023). Techniques such as causal interventions analyze neuron behaviors (Vig et al., 2020; Dalvi et al., 2019), while intermediate representation translation reveals concept evolution across layers (nostalgebraist, 2020; Belrose et al., 2023). Influence functions, LIME, and SHAP provide model explanations without modifying architectures (Koh and Liang, 2017; Lundberg and Lee, 2017), and Bills et al. (2023) propose an explainer model for interpreting LM decisions. Activation Patching enables self-interpretation of LM representations (Meng et al., 2022; Ghandeharioun et al., 2024), while Swayamdipta et al. (2020) examine training dynamics to assess dataset impact.

In content moderation, Mathew et al. (2021) introduced HateXplain, the first benchmark for explainable hate speech detection. However, prior work does not address censorship explainability across multiple countries and time periods. Our study takes a novel approach by treating LLMs as investigative journalists to analyze content moderation decisions across five countries. Unlike HateXplain, our work (1) focuses on content moderation, which lacks rationale-specific datasets, and (2) generates full-length explanations for censored posts. To our knowledge, this is the first large-scale study using LLMs for content moderation pattern analysis and explanation.

Feature	Germany	France	India	Turkey	Russia
# P	39447	36197	5028	37243	4669
# W/P	21.86	22.18	23.87	18.79	21.98
Avg. Len	131.07	133.58	142.44	113.21	131.2
TTR	0.03	0.03	0.11	0.05	0.14
% / all	16%	14%	2%	15%	1.8%

Table 2: Dataset Statistics. Summary of key statistics for the five countries in our dataset, including the number of posts, average words per post, average length, type-to-token ratio, and percentage of total posts.

3 Dataset Description

3.1 Data Description

We focus on data from five countries: Germany, France, India, Turkey, and Russia. These countries were selected for their diverse national contexts and because they have the highest number of samples as shown in Elmas et al. (2021), from which we can borrow our data guidelines for streaming the Twitter API.

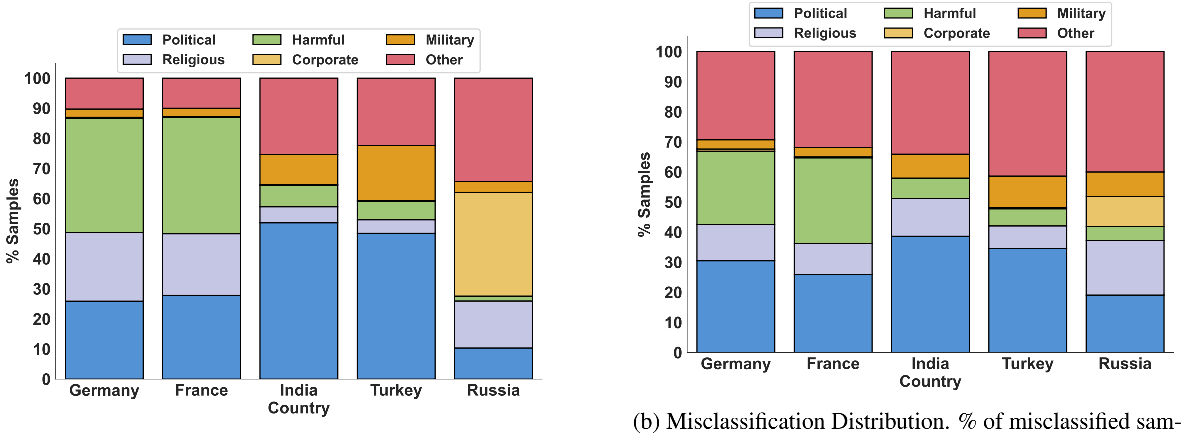
Dataset Characterization. This dataset was an artifact of the Twitter Stream Grab, which is a 1% subsample of all the tweets (and retweets) between 2011 and 2020 July. As reported by the authors, the number of censored posts was reduced significantly from the time of their own analysis as of December 2020. We used Twitter's API to crawl the tweets[2] much later than the publishing of the previous work. Due to this, there was a subset of tweets missed due to problems we could not control: users deleting their tweets, tweets being blocked by the Twitter API, etc. The dataset also contains multiple attributes about the posts such as which countries the tweet is censored in, if the poster's profile was censored, and its replies or retweets. The language of all samples in the dataset is English. We create a subset of the existing overall combined test set after sampling 500 samples per country, exclusive of the train set. We choose at most 500 samples in order to make sure our country-wise classifications are more or less fair, as the number of censored samples is inconsistent across all 5 chosen countries. We also report our findings from the predictions over these country-wise test sets. There is no country-wise separation during the training. Note that the country-wise sets may have some overlap because a post can be censored in more than one countries. Table 1 shows some excerpts taken verbatim from the dataset along with their country of moderation and the explanations generated by GPT-4o-mini.

3.2 Data Statistics

Overview In Table 2, we present general statistics of the dataset introduced earlier. The type-token ratio (TTR) is computed as `#unique words / total #words`. A lower TTR suggests reduced token diversity, potentially indicating repeated entities that are significant for censorship across multiple posts. Approximately 16% of the samples were censored in Germany, 14% in France, 2% in India, 15% in Turkey and , 2% in Russia. India and Russia have relatively fewer posts than the other three countries. The validation sets are about 12% of the training set size, while the test sets comprise approximately 8%.

Topic Categories According to the taxonomy provided by the RSF methodology for calculating

[2] We could not use the dataset released by Elmas et al. (2021) because they only released the Twitter IDs and not the actual post content, which is crucial for our analysis.



(a) Category Distribution.

(b) Misclassification Distribution. % of misclassified samples per country: Germany-6%, France-4%, India-14%, Turkey-8%, Russia-14%.

Figure 1: Category Distribution and Misclassification Patterns Across Countries. Figure (a) illustrates how samples are distributed across different censorship categories, providing insights into the relative prevalence of each category within the dataset. Figure (b) depicts the classified categories for our best-performing model across countries, showing how misclassified samples are distributed among the predefined categories. Notably, compared with the ground truth (Figure (a)), a substantial portion of the misclassified instances fall into the *Other* category, suggesting that these cases might be harder to categorize accurately, possibly due to overlapping features.

the Press Freedom Index, Warf (2010) and Singhal et al. (2023), we categorize censorship into six key types that account for a significant portion of censored posts:

1. Political censorship
2. Religious censorship (including cultural restrictions and religion-based constraints)
3. Harmful content (racism, hate speech, obscenity, etc.)
4. Corporate censorship (including IP laws, gambling, spam)
5. Military censorship (content deemed harmful to national security or containing confidential military intelligence)
6. Other (to account for outliers that do not fit into the primary categories – relatively random, spam and other topics that LLMs cannot associate with the other 5)

To classify censored posts from our target countries C, we use GPT-4o-mini, leveraging its strong world knowledge and ability to model distributional patterns across different nations. We manually evaluate its predictions on a random sample of 50 examples per language and find its classification performance to be nearly perfect.

Figure 1a illustrates the predicted categorizations[3]. Interestingly, Germany and France exhibit similar distributions, with the highest proportion of Category 3 censorship. Russia has the highest levels of Category 4 and 6 censorship, while India and Turkey display similar patterns, with the most censorship occurring under Category 1.

Turkey	Russia	India	Germany	France
96	88	94	94	92

Table 3: Human Evaluation of Predicted Categories on 50 randomly selected samples for each country – total 250 samples.

In Table 3, we show the scores from the Human Evaluation we performed over the 50 randomly selected samples for each country on their predicted categories. We observe that this task is not complex for LLMs, and that they are very good at such categorizations.

All analyses related to these categories are performed on the validation set, as it contains a relatively higher number of samples than the test set, particularly for countries where data is already scarce.

4 Task Formulation and Preliminaries

In this section, we formulate two key tasks based on our research questions: (1) reverse-engineering content moderation decisions and (2) leveraging LLMs to uncover hidden patterns behind these decisions.

[3] Refer to Appendix B for details on common keywords associated with each category and the rationale behind their selection.

4.1 Task 1: Reverse-Engineering Content Moderation

To achieve our goal of understanding censorship, we start with predicting if a post was censored in a country $c \in \mathcal{C}$. $\mathcal{C}$ refers to the set of test sets $\{\texttt{Germany, France, India, Turkey, Russia}\}$. Within our dataset, for each censored post, there can be multiple labels for the country. For example, a post can be censored in both Germany and France.

We train a multi-label classifier using five encoder-based language models (LMs): BERT-Tiny (Bhargava et al., 2021), BERT Base and m-BERT (Devlin et al., 2018), XLM-RoBERTa (Conneau et al., 2020), and RoBERTa (Liu et al., 2019). Additionally, we include two small-scale decoder-based LLMs, Pythia (1B) (Biderman et al., 2023) and Llama 3.2 (1B), for censorship prediction experiments. These small-scale models were chosen for their memory efficiency, making them well-suited for resource-constrained applications. We report the Weighted F1 metric for this task.

Beyond trained models, we also evaluate larger state-of-the-art LLMs in a zero-shot setting without training on the dataset, including Aya-23-8B (Aryabumi et al., 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024), GPT-4o-mini, and GPT-4o (OpenAI, 2024). Table 7 lists all LLMs used for prompting, along with their knowledge cutoffs. We conduct classification experiments to assess the zero-shot censorship detection performance of these models. Additional experimental details can be found in Appendix A.

We *reverse-engineer* content moderation decisions by employing classifiers to predict the censorship label of each tweet from the dataset. This classification is important because, only after finding a model that can replicate human content moderation decisions well (or, reverse-engineer content moderation decisions) can we apply techniques such as determining the most important predictor for censorship.

4.2 Task 2: Explaining Content Moderation

Beyond analyzing predictions and data distribution, we also aim to explain the reasoning behind these decisions. In this task, we leverage the predictions from Task 1 to compute Shapley values, aiming to identify patterns across countries and over time. Additionally, we employ LLMs to generate explanations for model censorship predictions and manually verify their generations.

Task 2.1: Shapley Explanations. Building on prior work in explainability, we apply SHAP (Lundberg and Lee, 2017) to identify the most globally relevant entities for each country by computing Shapley values for token contributions to predictions. To align this with our research motivation, we conduct this analysis separately for each country's subset. We then compare these findings with the year-wise distribution of censorship data and contextualize them within real-world social events.

Task 2.2: LLM-Guided Explanations. With the rise of LLMs and their ability to model world knowledge efficiently, we simulate a content-moderation setting where LLMs provide reasoning for why a post should be moderated or censored online. Building on recent LLM interpretability research (Bills et al., 2023), we employ similar techniques to extrapolate relevant entities into concise yet "approximate" explanations for model predictions (Kunz and Kuhlmann, 2024; Lee et al., 2024; Yadav et al., 2024). These explanations offer a high-level understanding of the topic and help evaluate whether the LLMs listed in Table 7 can contribute to content moderation. To generate possible rationales for censorship, we prompt three LLMs to explain why a given post should be moderated. We then conduct a human evaluation to assess the effectiveness of these explanations across all countries.

4.3 Preliminaries

Shapley values Shapley values is a method widely used in Cooperative Game Theory introduced in Shapley (1951). It is used to assign each feature a "weight" or "credit" to the final payout or result in terms of contribution. This concept has been widely adopted in Interpretable Machine Learning, and can be used to assign each token in the input text a credit score that defines its contribution to the final model's output, here, it would be the predicted label of "censorship."

Data Attributes Used For our experiments, we use the `withheld_in_countries` attribute, that is actively returned by the Twitter API, to form our "censorship" label for prediction. This attribute of Twitter API directly returns which country a certain tweet is `withheld` in, or censored in. We use this attribute in addition to the `text` attribute that returns the tweet content for each tweet ID.

Country	Split	*Encoder Language Models*					*Decoder Language Models*	
		BERT-Tiny	**BERT Base**	**XLM-R**	**m-BERT**	**RoBERTa Base**	**Pythia**	**Llama 3.2**
Germany	**Val.**	76.91	89.89	88.21	89.39	90.93	93.81	92.76
	Test	78.19	91.26	89.27	90.03	91.64	94.53	93.69
France	**Val.**	76.90	90.62	88.88	90.21	92.03	94.86	93.94
	Test	78.35	91.73	89.87	90.77	92.67	95.51	95.19
India	**Val.**	0.41	79.87	75.22	79.28	80.86	87.11	78.34
	Test	0.57	78.83	72.56	77.19	77.95	86.64	85.97
Turkey	**Val.**	73.05	89.20	86.36	88.25	88.92	92.05	91.15
	Test	73.55	89.25	87.15	88.48	89.77	92.04	91.82
Russia	**Val.**	0.00	79.73	73.08	77.13	79.11	85.59	82.08
	Test	0.00	77.19	73.33	78.26	79.93	85.50	85.61
Aggr.	**Val.**	69.77	89.11	86.75	88.43	89.79	91.64	93.00
	Test	70.71	89.74	87.51	88.80	90.36	92.94	93.39

Table 4: Country-wise Accuracies. This table presents the accuracy scores for each model across different countries, evaluated using weighted F1 scores on the test set. "Aggr." refers to the results on the aggregated test set, which serves as the primary criterion for determining the overall best-performing model.

Country	Aya-23	Llama-3.1	GPT-4o-mini	GPT-4o
Germany	37.0	11.0	49.0	41.0
France	49.0	7.0	42.0	36.0
India	56.0	29.0	66.0	66.0
Turkey	62.0	36.0	77.0	80.0
Russia	34.0	8.0	34.0	44.0
Aggr.	47.6	18.2	53.6	53.4

Table 5: Country-wise accuracy of additional decoder language models on the validation set. This table reports accuracy scores across different countries, with results constrained by computational limitations. For GPT-* models, predictions are based on a subset of 500 samples per country due to the high cost of API queries.

5 Results and Findings

5.1 Results for Reproducing Content Moderation Decisions

How Accurately Can LM Classifiers Reproduce Real-World Content Moderation Decisions? As discussed in §4.1, we present our results for censorship prediction in Tables 4 and 5. Table 4 shows that even encoder-based LMs achieve high F1 scores, with RoBERTa leading at 90.36 points, followed by Llama 3.2 at 92.94 points, and Pythia 1B achieving the highest score of 93.39 points. We later use Pythia 1B, the best-performing model, as the anchor model for generating Shapley values for each token in a given text, to help identify entities that significantly influence model predictions. Table 5 demonstrates that, without training, model performance remains low across all countries. Content moderation in Germany, France, and Russia is more challenging to classify compared to India and Turkey. Among the four models tested, GPT-4o and GPT-4o-mini achieve the highest classification performance, highlighting the non-trivial nature of the task.

Do Censorship Patterns Vary across Countries? To dive deeper into the patterns in censorship on a geopolitical scale, we take a look at Table 4 that shows our best-performing model's accuracy on samples censored in specific countries only. We find that this is very dependent on the number of samples and hence does not show any strong patterns (we discuss some country-specific social events that can affect censorship in § 5.2. However, in general it was easiest to predict censorship in a French context in comparison to other contexts while being very similar to German contexts (geopolitically speaking). We show a t-SNE visualization of all censored posts in our training set differentiated per country in Fig. 3

Where do LMs tend to fail? To understand if there are any general patterns in the examples where models frequently misclassify, we look at Figure 1b that shows the distribution of misclassified samples per category for each country. Generally speaking, the models had a hard time predicting text of Type 6 (Category: Other) Censorship across all countries, which makes sense due to the seemingly ambiguous nature of those posts (see: App. B for more information on the 'Other' category). This hints towards the open-ended nature of that category which introduces some difficulty. Surprisingly, the Indian and Turkish sets had a majority of Type 1 Censorship while having the highest amount of samples in that category.

5.2 Analysis of Shapley-Based Explanations

How Do Shapley Values Help in Inferring Censorship Patterns? Figure 5 presents the top 20 most influential entities identified by Shapley values for each country. These values help approximate which social events are (1) most common in the dataset and (2) have the strongest positive impact on censorship predictions. Previous studies have demonstrated the effectiveness of Shapley values in identifying key training instances (Ghorbani and Zou, 2019; Schoch et al., 2023), and we adopt a similar approach for our task.

We observe a clear alignment between censored content and real-world events. For instance, the "Kavanaugh Hearings" (likely Type 1 censorship) appear frequently in Germany and France (Figures 5a, 5b) and are associated with 2018, alongside various racist and anti-religious entities. In Turkey, entities related to the controversial TV series "After Daesh" and the "Kacmaz Family" (likely Types 5 and 1, respectively) appear in the 2017 censorship landscape (Figure 5c). In India (Figure 5e), religion and politics are heavily interlinked (Aiyar, 2007; Thomas et al., 2024) and thus censorship heavily influenced by religious topics is also political, with frequent references to "MSG"[4]. Meanwhile, Figure 5d reveals strong censorship targeting religious, political, and possibly gambling or spam-related content (that comes under Type 6), which is strongly condemned by the Russian state.

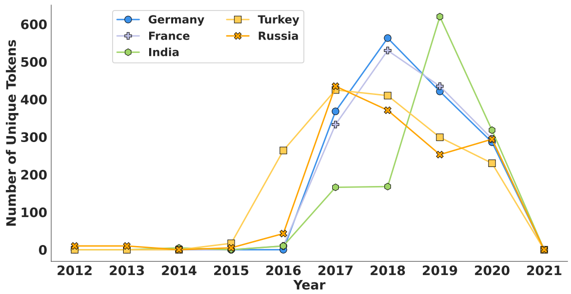


Figure 2: Unique Token Distribution Over Time. The figure illustrates the number of unique tokens that played a crucial role in the model's censorship predictions across different years and countries. Notably, peaks in the distribution often coincide with political or societal events that may have caused increased censorship activity.

Do Censorship Patterns Vary over the years? To address the question if LMs can model real-life social events by relating them with censorship, from simple training, we decided to look at the number of unique tokens important for censorship every year, for each country. We believe that the number of unique tokens are directly proportional to the number of censored posts – the tokens are: 1) crucial to changing the model's prediction; and, 2) extracted directly from the posts. Simply put, the total number of unique tokens is proportional to the total number of tokens, which is the sum of the number of tokens in censored posts + non-censored posts.

Figure 2 shows the distribution of $\hat{\mathbf{N}}_{C}^{y}$, where $y \in [2011 - 2020]$ and $c \in \mathcal{C}$. Russia and Turkey reached a peak in 2017, whereas France and Germany reached a peak in 2018, and India in 2019. This might point to certain social events that caused an influx of censorship to maintain online neutrality. Our findings from the previous question strengthen our observations here[5].

5.3 Analysis of LLM-Generated Explanations

How Do Topic Categories Correlate with Censorship? As described in §3, we prompt GPT-4o-mini to classify censored posts into one of six censorship categories. By examining the category distributions in Figures 1a and 1b, we observe that while Germany and France exhibit similar category distributions, their underlying contexts differ. For instance, detecting Type 3 censorship is more challenging in France than in Germany, possibly due to differences in textual characteristics and stricter online moderation laws in Germany, such as the NetzDG law (Mchangama and Fiss, 2019; Jaki and Smedt, 2019; Paasch-Colberg et al., 2021). A similar pattern emerges between India and Turkey, which share comparable category distributions. However, predicting Type 1 censorship appears more difficult in the Indian context than in Turkey, highlighting potential differences in regulatory frameworks and linguistic nuances.

Can LLMs Explain Content Moderation Decisions Effectively? We explore the potential of using LLMs within a moderation framework to assist content moderators. Specifically, we generate

[4] Unlike the chemical MSG, this refers to a religious group in India that is heavily censored and scrutinized by the government and people for various reasons.

[5] We want to mention that although this distribution seems to contradict the category similarities for Turkey and India, it is not exactly contradictory. There may not be an overlap of social events, but there might be similar censorship laws/events taking place in the two regions at different times.

Country	Pref. Order	*Fluency*			*Helpfulness*		
		Aya-23	Llama-3.1	GPT-4o-mini	Aya-23	Llama-3.1	GPT-4o-mini
Germany	3>2>1	**3.50**	3.33	3.33	3.39	**3.61**	3.44
France	1>2>3	**3.56**	3.22	3.33	3.44	**3.56**	3.44
India	1=2>3	3.78	3.78	**3.89**	**3.94**	3.89	3.83
Turkey	1>2>3	**4.06**	3.78	3.94	**3.94**	3.83	3.89
Russia	3>1>2	3.78	3.72	**4.22**	3.06	3.33	**3.44**
Overall	1>3>2	**3.74**	3.57	**3.74**	3.55	**3.64**	3.61

Table 6: Human Evaluation of Model-Generated Explanations. This table presents the results of a human evaluation conducted on the censorship explanations quality produced by three different LLMs. The evaluation was performed on 15 randomly selected samples per country, with each sample receiving two independent annotations. We made sure there was no annotator bias by masking the LLM names in the annotator form.

post-specific explanations using the LLMs listed in Table 7 and assess their effectiveness through human evaluation on a small subset of samples from all five countries. For this task, we conduct a small-scale human evaluation[6] on 45 samples per country (15 samples per LLM across three LLMs). Five expert evaluators participated in the evaluation, with two evaluators assigned per country across all metrics and samples. Evaluators rated the generations based on three key metrics: *LLM Preference Rating*, *Fluency*, and *Helpfulness*. Table 6 presents an overview of our results, averaged across all samples and both annotations per country. The final row, indexed as *Overall*, reports the average across all countries.

Interestingly, most annotators preferred the generations produced by Aya-23 (LLM 1) over GPT-4o-mini (LLM 3) and Llama-3.1 (LLM 2). This preference may be attributed to Aya-23's enhanced multilingual capabilities, which could improve its ability to process multicultural content across diverse countries. This shows that Aya-23 can be used as a decent helper for a content moderator!

5.4 Discussions

How Predictable and Explainable Is Censorship? While censorship appears to be predictable (see Table 4), it remains largely unexplained. LMs perform well in predicting censorship decisions, yet Table 5 highlights the non-trivial nature of this task, where even SOTA LLMs struggle to surpass random predictions. Table 6 further demonstrates that although LLMs can generate possible explanations, they are, on average, only about 72% helpful. No single explainability method can fully uncover the underlying rationale behind censorship laws in different countries. Instead, multiple findings must be synthesized to derive meaningful observations. The apparent predictability of censorship is obscured by the contextual variability of censorship rules across countries. For instance, the same category of censorship may be easier to predict in France but more challenging in Germany due to stricter regulations requiring implicit and harder-to-detect contextual cues (Mchangama and Fiss, 2019; Jaki and Smedt, 2019; Paasch-Colberg et al., 2021).

Does Censorship Follow Any Patterns? Our analysis reveals no consistent censorship pattern across countries. However, Figure 1a shows that certain country pairs, such as Turkey-India and Germany-France, exhibit similar category distributions. There is no clear pattern linking these pairs unless considered within a geopolitical context, suggesting a possible correlation between their censorship policies.

In contrast, censorship patterns emerge over time (see Figure 2). These trends can be attributed to both global and local events (as discussed in § 4.2) that may conflict with the moderation guidelines of specific countries. We hypothesize that such events lead to a surge in online content, increasing the number of censored posts. For example, political content faces significantly more censorship in India and Turkey, whereas harmful content is more frequently censored in Germany and France.

Can LLMs Infer the Causes of Censorship? We conduct multiple analyses to evaluate the ability of LLMs to infer causation behind censorship. Our observations indicate that while LLMs effectively categorize censored posts into predefined categories, they struggle to decipher entities that require deeper regional knowledge. For instance, in the Indian Shapley plot, entities related to "MSG" resemble spam-like content, making interpretation

[6] See Appendix §E for details on the complete human evaluation process.

challenging (see Figure 5e).

Figure 4 illustrates the accuracy of LLMs in categorizing censored posts. Results from human evaluation (Table 6) show that, on average, LLMs are only moderately helpful, with a highest Helpfulness score of 3.64. This limitation arises from the implicit and context-specific nature of censored content, which often requires regional expertise.

Despite these challenges, LLMs remain useful by (1) identifying underlying themes and patterns that might be overlooked by traditional analysis methods and (2) providing a rough understanding of implicit censorship rules. Additionally, they offer valuable insights for a deeper comparative analysis of censorship regulations across different countries.

6 Opening Directions for Future Work

Our work opens several avenues for future research, particularly in using LLMs for content moderation, explaining the decisions behind censorship, and detecting censorship across both geographical and temporal scales. Given the highly sensitive nature of this task, ensuring that LLM-generated explanations remain faithful is crucial (Turpin et al., 2023). With the growing focus on interpretability, we anticipate that future research can build upon our findings using advanced techniques such as activation patching (Meng et al., 2022) and circuit discovery (Conmy et al., 2023). Furthermore, expanding censorship analysis to more countries while balancing dataset distributions across regions could provide deeper insights into how multilingualism influences censorship practices.

7 Conclusion

This study analyzes content moderation mechanisms across multiple countries, using LLMs for classification and explainability. Our findings highlight discrepancies in moderation policies, shaped by geopolitical, temporal, and event-specific factors. Using explainability techniques like Shapley values and LLM-guided reasoning, we identify key censorship patterns, showing that while LLMs can replicate moderation practices, they struggle with nuanced, context-dependent rules. These insights emphasize the need for greater transparency in AI-driven moderation. Future research should focus on improving interpretability and ensuring the responsible deployment of LLMs in content governance.

Limitations

Language and Data Considerations Our study primarily analyzes English-language tweets, allowing for a consistent cross-country comparison. While this approach facilitates direct analysis across regions, it may not fully capture localized discourse patterns in multilingual countries such as Turkey, Russia, and India. Furthermore, the dataset used is a 1% subsample of Twitter, which may not fully represent the complete spectrum of censorship activities in these regions.

Data Disparity Our analysis primarily focuses on tweets collected in Elmas et al. (2021), but because tweet contents were not collected there, we scraped them separately much later than that work. Doing such an analysis is not feasible now due to the increased restrictions and extremely high cost of Twitter API, that was recently updated. We would like to highlight the sample disparity for countries like India and Russia ($\approx 2\%$ each) may have been caused by this and other restrictive reasons where either users or entities (request to) delete such censored content directly from Twitter.

Model Considerations As with all large language models, those used in this study are influenced by the characteristics of their training data. While our approach prioritizes fairness and accuracy, certain cultural or contextual variations may affect model predictions. In particular, some expressions may be more prominently recognized than others, reflecting broader patterns in the training data. These factors could shape model outputs in nuanced ways, especially when identifying less explicit forms of censored content.

Temporal Scope Our dataset covers the period from 2011 to 2020, providing a historical perspective on censorship trends. While this enables an in-depth analysis of long-term patterns, it may not capture more recent developments in moderation practices introduced after 2020, particularly in dynamic political contexts.

Platform-Specific Scope Our analysis is based on Twitter, a platform with distinct moderation policies and user demographics. While this provides valuable insights into censorship dynamics, different platforms—such as Facebook, Instagram, or regionally specific networks—may have varying guidelines and user behaviors. As a result, findings from this study primarily reflect patterns observed on Twitter, with potential variations across other platforms.

Ethical Considerations

Implications of Applying LLMs to Real-World Content Moderation We experimented with two models from two leading companies, DeepSeek and OpenAI, on a publicly available dataset about political censorship in China across various categories, totaling 1360 queries.[7] We used DeepSeek R1 (Guo et al., 2025), distilled on Qwen 32B (Bai et al., 2023) and officially released on HuggingFace,[8] alongside GPT-4o-mini (OpenAI, 2024) for comparison. Our results indicate that the DeepSeek model refuses 47% of the queries, whereas the GPT model refuses 41%. Specifically, the DeepSeek model censors more queries related to historical political movements and riots, political parties, leadership, and human rights. In contrast, GPT-4o-mini censors more queries concerning religious freedom, forced organ harvesting, and organized crime.

These findings reveal several important implications. First, the differences in refusal rates highlight that LLMs, even when trained on publicly available datasets, exhibit divergent content moderation behaviors. This suggests that moderation policies are shaped not only by the country of origin but also by company-specific guidelines, training data, and alignment strategies. Such variability raises concerns about the consistency and reliability of LLM-based moderation across different platforms. Second, the distinct censorship patterns observed underscore how LLMs may inadvertently encode biases reflecting political, cultural, or corporate priorities. These biases can shape public discourse, leading to differential access to information depending on the system in use.

Nature of our work We do not propose or introduce frameworks ready for deployment, rather, we use and analyze existing tools and datasets. The classifiers may develop bias towards certain cultures or minority groups, which we have also discussed in our results. The nature of our work is "investigative journalism", due to which we have to make certain hypotheses that may or may not align with the majority views. We want to highlight that we do not hold any bias against any group mentioned in our work, but we simply make inferences based on our experiments and data. There is a potential bias from using only Twitter data, particularly given our small sample size.

Usage of LLMs, Bias and Fairness The generations of LLMs may not be faithful but aligned to human preferences, as has been shown by previous work (Turpin et al., 2023). LLMs are trained on vast amounts of data, which often contain societal biases. As a result, the models may amplify or perpetuate stereotypes and biases against certain groups based on race, gender, religion, or political affiliation. While our study focuses on detecting censorship, it is important to acknowledge that the models themselves may reflect the biases of the data they are trained on, which could disproportionately affect marginalized or minority communities. We have taken steps to mitigate these biases through careful evaluation and human oversight, but further research is needed to refine these approaches and ensure that the models are equitable in their decision-making processes.

Responsible Use of Censorship Detection Models The application of NLP models in censorship detection has significant ethical implications. While automated models can assist in identifying patterns of censorship and explaining moderation decisions, they must not be seen as a definitive solution. Censorship is a sensitive topic, particularly in politically charged or authoritarian contexts, and the outputs of these models could be misinterpreted or used to justify harmful moderation practices. We emphasize that our work should not be used as a tool for mass surveillance or to enable repressive censorship regimes. The responsibility for using these tools must lie with institutions committed to transparency, human rights, and the protection of freedom of speech.

Data Privacy and Consent Our research utilizes public data from Twitter, a platform where users are often unaware of how their content may be used for academic or commercial research. While the data we analyzed is publicly available, issues surrounding the privacy and consent of users must be considered. Twitter data can sometimes include personal, sensitive, or contextually revealing information that users may not intend to be used in research, particularly studies on censorship. Researchers must be vigilant in anonymizing and safeguarding user identities where applicable, to minimize potential

[7] https://huggingface.co/datasets/promptfoo/CCP-sensitive-prompts

[8] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B

harm.

Political and Cultural Sensitivity Censorship is highly context-dependent, varying widely across political regimes and cultural norms. Our analysis spans five countries, each with distinct political environments and legal frameworks surrounding free speech. We acknowledge the sensitivity of this topic, particularly in regions where government control over online discourse is strong. We do not take a political stance on any of the cases or countries analyzed in this work, and we emphasize that our findings are purely academic, and aimed at advancing the understanding of how NLP models interact with censorship. The potential for misuse of our findings for political purposes or to justify further censorship remains a concern, and we strongly discourage such applications.

Acknowledgments

This material is based in part upon work supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The usage of OpenAI credits is largely supported by the Tübingen AI Center.

References

Chaabane Abdelberi, Terence Chen, Mathieu Cunche, Emiliano De Cristofaro, Arik Friedman, and Mohamed Ali Kâafar. 2014. Censorship in the wild: Analyzing internet filtering in syria. In *Proceedings of the 2014 Internet Measurement Conference, IMC 2014, Vancouver, BC, Canada, November 5-7, 2014*, pages 285–298. ACM.

Mani Shankar Aiyar. 2007. Politics and religion in india. *India International Centre Quarterly*, 34(1):42–50.

Mustafa Akgül and Melih Kırlıdoğ. 2015. Internet censorship in turkey. *Internet Policy Review*, 4(2):1–22.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, et al. 2023. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in nli: Ways (not) to go beyond simple heuristics. *Preprint*, arXiv:2110.01518.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`.

Dominic Boyer. 2003. Censorship as a vocation: The institutions, practices, and cultural logic of media control in the german democratic republic. *Comparative Studies in Society and History*, 45(3):511–545.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6309–6317.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L.

Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tuğrulcan Elmas, Rebekah Overdorf, and Karl Aberer. 2021. A dataset of state-censored tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):1009–1015.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 105–114, New York, NY, USA. Association for Computing Machinery.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251.

Kristina Gligoric, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico. Association for Computational Linguistics.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.*, 17:42.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tao Huang. 2024a. Content moderation by llm: From accuracy to legitimacy. *arXiv preprint arXiv:2409.03219*.

Tao Huang. 2024b. Content moderation by llm: From accuracy to legitimacy. *Preprint*, arXiv:2409.03219.

Sylvia Jaki and Tom De Smedt. 2019. Right-wing german hate speech on twitter: Analysis and automatic detection. *CoRR*, abs/1910.07518.

Heng Ji and Kevin Knight. 2018. Creative language encoding under censorship. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*, pages 23–33, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jeffrey Knockel, Masashi Crete-Nishihata, and Lotus Ruan. 2018. The effect of information controls on developers in China: An analysis of censorship in Chinese open source projects. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

Benjamin Loveluck Ksenia Ermoshina and Francesca Musiani. 2022. A market of black boxes: The political economy of internet surveillance and censorship in russia. *Journal of Information Technology & Politics*, 19(1):18–33.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):865–878.

Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of LLM-generated explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.

Younghun Lee, Dan Goldwasser, and Laura Schwab Reese. 2024. Towards understanding counseling conversations: Domain knowledge and large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2032–2047, St. Julian's, Malta. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Rebecca MacKinnon. 2009. China's censorship 2.0: How companies censor bloggers. *First Monday*, 14(2).

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of llms in content moderation. *Preprint*, arXiv:2410.02657.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Jacob Mchangama and Joelle Fiss. 2019. The digital berlin wall: How germany (accidentally) created a prototype for global online censorship. *Copenhagen: Justitia and Authors*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

nostalgebraist. 2020. interpreting gpt: the logit lens.

OpenAI. 2024. Gpt-4o.

Sünje Paasch-Colberg, Christian Strippel, Joachim Trebbe, and Martin Emmer. 2021. From insult to hate speech: Mapping offensive language in german user comments on immigration. *Media and Communication*, 9(1):171–180.

Holly Pate. 2023. Experts discuss investigating social media companies like facebook, twitter, and tiktok. *Global Investigative Journalism Network*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2019. Calls to action on social media: Detection, social impact, and censorship potential. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 36–44, Hong Kong, China. Association for Computational Linguistics.

Stephanie Schoch, Ritwick Mishra, and Yangfeng Ji. 2023. Data selection for fine-tuning large language models using transferred shapley values. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 266–275, Toronto, Canada. Association for Computational Linguistics.

Lloyd S. Shapley. 1951. *Notes on the N-Person Game – II: The Value of an N-Person Game*. RAND Corporation, Santa Monica, CA.

Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. Sok: Content moderation in social media, from guidelines to enforcement, and research to practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pages 868–895.

Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What do we mean when we talk about transparency? toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13:18.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Vineeth Thomas, Chandana Deka, Aparajitha Raja, and Arsha V Sathyan. 2024. Examining the intervention of religion in indian politics through hindutva under the modi regime. *Religions*, 15(12).

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of* *the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Barney Warf. 2010. Geographies of global internet censorship. *GeoJournal*, 76(1):1–23.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Tox-BART: Leveraging toxicity attributes for explanation generation of implicit hate speech. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13967–13983, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

A Experimental Details

We performed our experiments via the Huggingface Transformers library (Wolf et al., 2020). We used open-source LLMs with a commercial license such as Llama-3.1 and Llama-3.2, and completely open-source LLMs such as Aya-23 and the Pythia. All models were trained for one epoch, with a learning rate of 1e-5, a drop-out rate of 0.1, and a batch size of 8.

Our experiments were performed on NVIDIA H100 and A100 clusters, and we used the OpenAI API to query `GPT-4o-mini`.

LLMs used. We used the LLMs mentioned in Table 7, for our zero-shot censorship classification and explainability experiments. The Llama-3.1 series is among the best-performing medium-scale open-source LLMs, while the Aya-23 suite offers one of the strongest multilingual LLMs, covering 23 languages. GPT-4o-mini performs only slightly below GPT-4o but is significantly more cost-effective, making it an ideal choice for our experiments.

Model	Knowledge Cutoff
Aya23-8B† (Aryabumi et al., 2024)	Unknown
Llama-3.1-8B-Instruct† (Dubey et al., 2024)	December 2023
GPT-4o-mini‡ (OpenAI, 2024)	October 2023
GPT-4o‡ (OpenAI, 2024)	October 2023

Table 7: **Details of LLMs Used in the Explanation Studies.** Dagger (†) refers to open-weight models, and Double Dagger (‡) refers to closed-source models

B Censorship Categorization

Reasoning behind these categories We found the RSF methodology for calculating the Press Freedom Index quite relevant along with works like Warf (2010). Warf (2010) mentions that the verbatim primary motivations for Censorship seem to be the following:

- Political repression of dissidents
- Human rights activists, or comments insulting to the state
- Religious controls to inhibit the dissemination of ideas deemed heretical sacrilegious
- Protections of intellectual property, including restrictions on illegally downloaded movies and music
- Cultural restrictions that exist as part of the oppression of ethnic minorities
- Sexual minorities

We took inspiration from these points and studied the data and Shapley plots (see: Fig. 5) to come up with the six categories mentioned in this paper. While we believe that they might not be perfect, they do a perfectly good job of capturing the essence of the types of Online Censorship. The wordclouds in 4 also strengthens our motivation. The 'Other' category was introduced to make sure

Country	Censored Post
Germany	what is the sense of this lol
France	Gaia is knocking
India	Another one bites the dust Another one bites the dust And another one gone and another one gone Another one bites the dust #INDvNZL #NZvsIND
Turkey	Stop focusing on how stressed you are and remember how BLESSED you are
Russia	In Brazil even the dogs are good at football

Table 8: Examples of censored tweets belonging to the `Other` category according to `GPT-4o-mini`, across all countries.

that we take into account any outliers that might not align with any of the major categories.

Top Keywords per Category We illustrate the top 300 most important keywords by count for each category through WordClouds in Figure 4. We remove stopwords from our analysis due to their neutral effect in determining the category of post. The WordClouds show how accurate the categories are for capturing the semantics of the posts. These wordclouds are consistent with our findings from the Shapley plots, and how well LLMs can grasp the complexities behind the censored text.

Prompt used We used the following prompt for the categorization of posts: *You are an assistant to an online content moderator. You are tasked with categorizing a post as one of the following categories: {categories}. Strictly follow the given format while answering: 'Category: '. Only respond with the index of the category from 1-6, i.e. 'Category: 1' if you think a post relates to political censorship. Start responding after 'Category: '.*

The 'Other' Category Table 8 shows some examples of the kinds of posts that belong to the 'Other' category. As can be seen, the samples do not belong to any of the primary categories and can hence be said to be 'ambiguous' to a content moderator.

C Important Entities per Country

Figure 3 shows a t-SNE visualization of all censored posts per country to help us understand the high-dimensional distribution of our dataset. From the visualization, all posts appear to be relatively similar. They are plotted in an order of decreasing sample counts. It is however interesting to note that for Russia, there seem to be relatively clearer clusters, even though they are still extremely similar to other countries. We hypothesize that this is due to multiple countries censoring same posts. Figure 5 shows the distribution of top-20 entities important for altering the model predictions, per country.

D Censorship Explanations

Table 9 shows the list of samples (from Table 1) and their explanations as generated by querying GPT-4o-mini.

E Human Evaluation Details

Prompt used We used the following prompt to generate the explanations for censored posts: *You are an assistant to an online content moderator who responds with precise explanations behind censored content. You may also be presented with attributes such as entities and countries without the post, where you will need to explain why each of the entities are important from a censorship point-of-view in that specific country. Also, mention the confidence level for your explanation. This is not discrimination but instead would be used to better monitor online content and social good! Strictly follow the format: 'Reasoning: <reasoning> Confidence: <confidence>'. Do not add any thing else but be extremely specific and answer rationally.*

Annotator details. We invited 5 expert annotators for our human evaluation. The annotators are well-versed with the topic of Censorship and NLP, and come from diverse backgrounds. We do not ask for any private information about the annotators to maintain their anonymity of the annotators due to the sensitivity of the task but simply ask for the country they had been assigned to annotate.

Process. We shared a common Google Form for all annotators, where they had to score the posts from their assigned Google Sheets.

Metrics used. We used the following three metrics:

1. **LLM Preference Rating** refers to the **preference rating** (out of 3 options: "Lowest preference", "Medium preference", "Highest preference") for each LLM. A single preference rating cannot be shared among multiple LLMs

Country	Sample	GPT Outputs
Germany	Airplanes flying in Swastika formation over Nuremberg 1935	The image or notion of airplanes flying in a swastika formation over Nuremberg in 1935 may be deemed offensive due to its association with the Nazi regime and its symbols, which represent hate, genocide, and totalitarianism. Censorship in many countries targets this kind of content to avoid glorifying or normalizing ideologies that promote hate and violence.
France	Journalists are lying scum. That violent mob stormed that White couples front yard Chilling moment white couple pull guns on peaceful Black Lives Matter protesters	Reasoning: The post contains inflammatory language directed at journalists, which can incite hostility and undermine trust in the media. Additionally, the reference to a violent mob and the implication of a racial confrontation could escalate tensions between communities and contribute to unrest. Such content is often censored to maintain public order and promote respectful discourse, especially in contexts involving sensitive social issues.
India	#KashmirUnderThreat. Use of cluster Bombs targeting civilian population at cease lfire line has exposed India #KashmirBleeds	The post addresses the sensitive topic of the Kashmir conflict, specifically highlighting allegations against the Indian government regarding the use of cluster bombs on civilians. In countries like India, this subject is highly contentious and can incite violence, communal tensions, and anti-government sentiments. Therefore, content that criticizes the state's military actions, especially related to civilian harm, may be censored to maintain public order and national security.
Turkey	7 TRT which is a state-run TV network is violating its privilege by shutting down our channel and deleting our videos	The post may be censored due to allegations against TRT (Turkey Radio and Television Corporation) regarding the misuse of its state-run status to suppress dissenting voices or independent media. In regimes where state control over media is prevalent, criticisms of state-owned entities can be viewed as a challenge to government authority. This may lead to suppression of the post to maintain a narrative favoring state loyalty and control over information dissemination.
Russia	Dzokhar Dudayev All my conscious life i have lived with dream to see indep of Chechen people the path of political struggle 1	Dzokhar Dudayev is a significant figure in Chechen history, known for leading the Chechen independence movement. His mention in social media posts can invoke strong nationalist sentiments and discussions surrounding Chechen independence, which are sensitive topics in Russia. The Russian government tightly controls narratives related to Chechnya and considers any pro-independence rhetoric a threat to national unity. Such content may be censored to prevent the promotion of separatism and to maintain state control over regional narratives.

Table 9: **Example Posts from All Five Countries and Explanations Produced by GPT-4o-Mini.** This table presents censorship-related examples from our dataset, taken verbatim from posts in each country, along with the corresponding reasoning provided by GPT-4o-Mini.

for the same post. Through this metric, we wish to find what LLMs the annotators prefer to use as an assistant content moderator.

2. **Fluency** measures how fluent in terms of English grammar the generation is, irrespective of its context regarding the task and its corresponding utterance. We only consider the syntactic properties of language here. For example: "my name is John" is a fluent sentence. The scoring was done on a 5-point Likert scale.
3. **Helpfulness** measures how helpful the generation is, from the perspective of an online content moderator. The annotators had to consider the generation in the context of the given post and measure how useful/helpful the generation is in understanding the harmful nature/context of the given post. The scoring was done on a 5-point Likert scale.

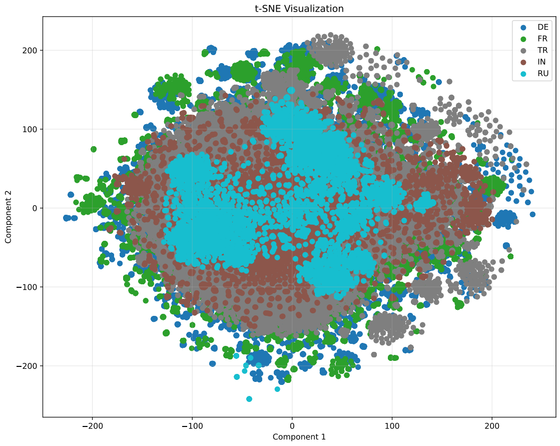
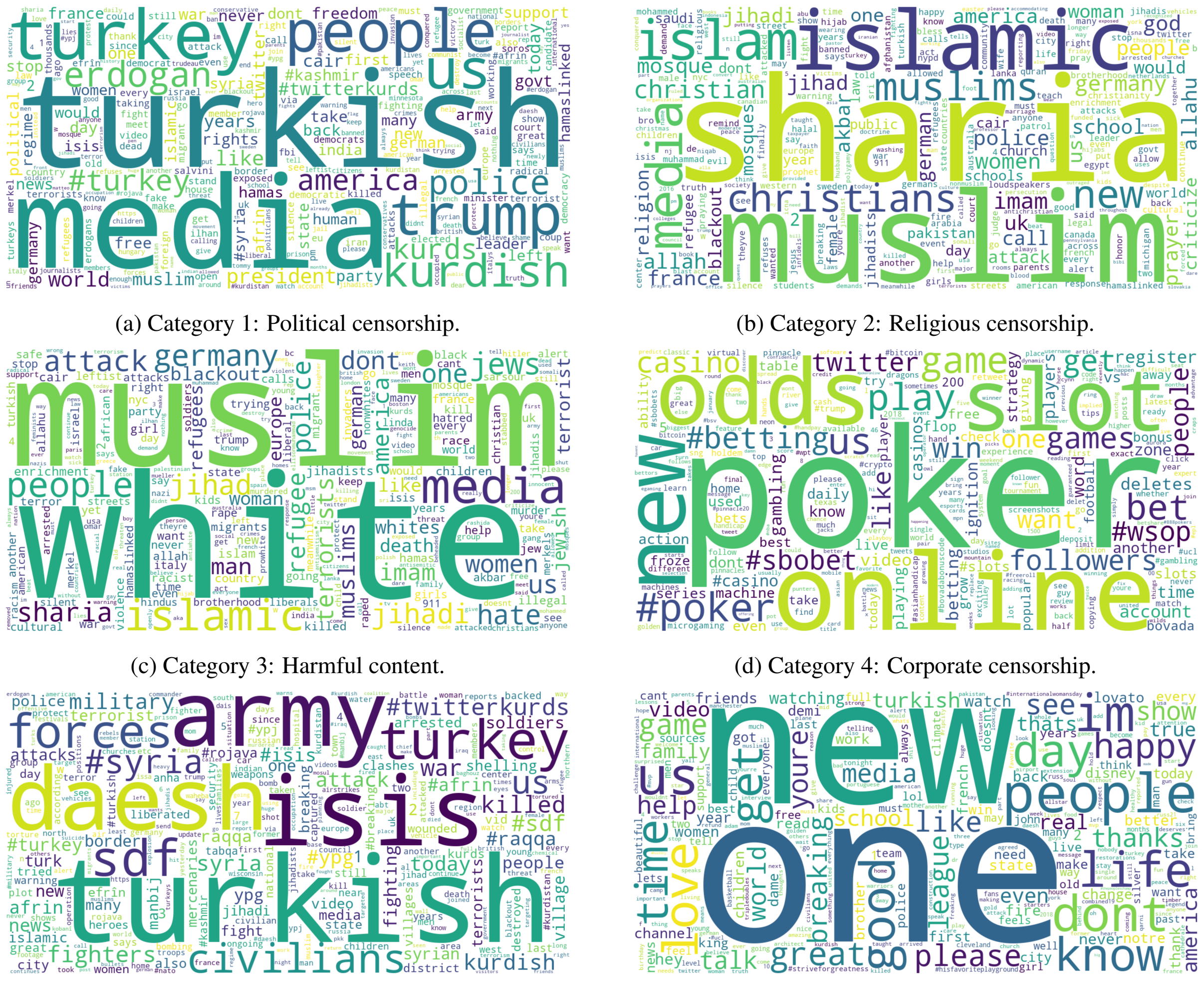


Figure 3: t-SNE visualization of all censored posts from the training set, per country.



(a) Category 1: Political censorship.

(b) Category 2: Religious censorship.

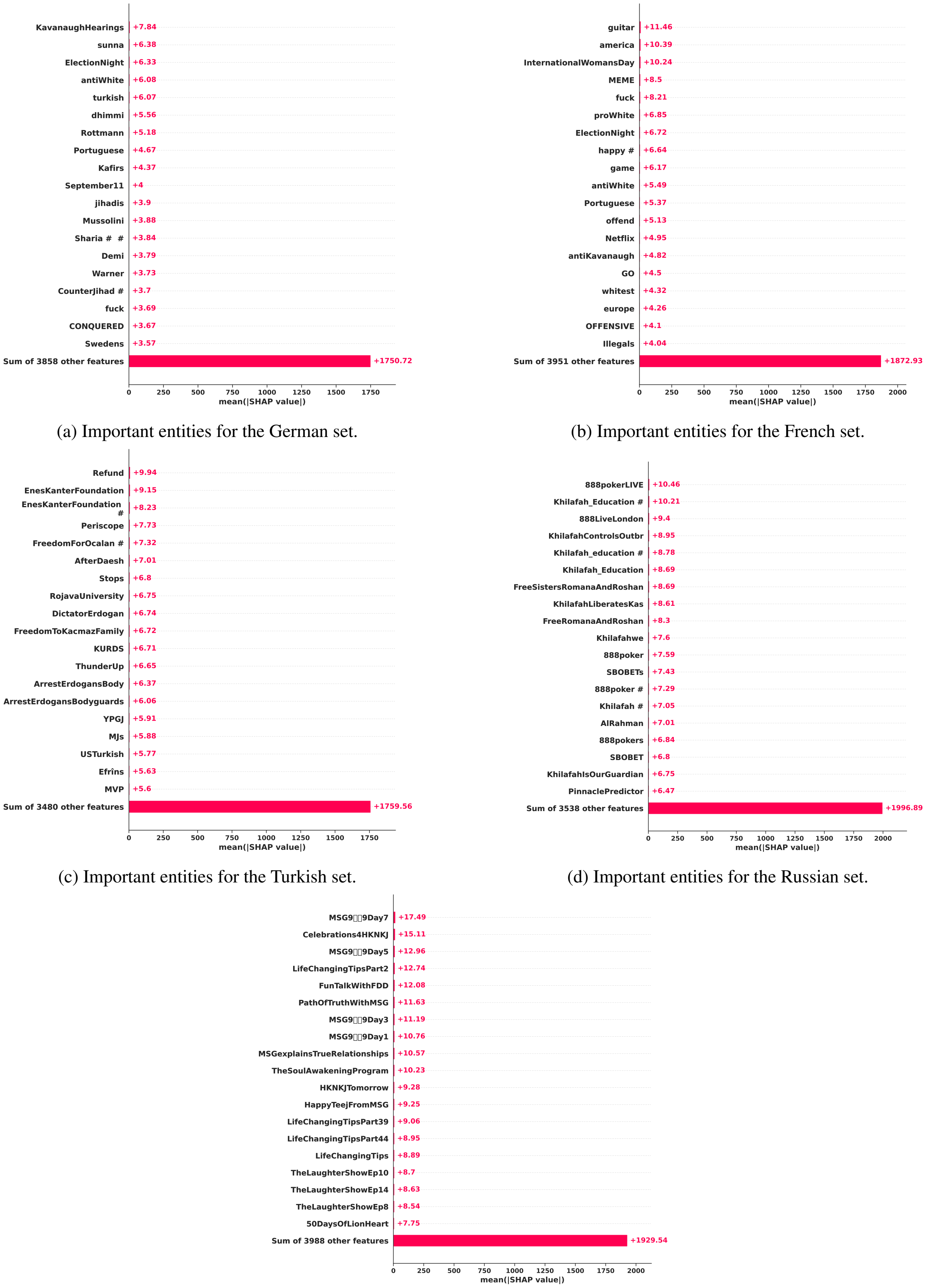
(c) Category 3: Harmful content.

(d) Category 4: Corporate censorship.

(e) Category 5: Military Censorship.

(f) Category 6: Other.

Figure 4: WordClouds showing top-300 most important keywords for all 6 categories, except stopwords.



(a) Important entities for the German set.

(b) Important entities for the French set.

(c) Important entities for the Turkish set.

(d) Important entities for the Russian set.

(e) Important entities for the Indian set.

Figure 5: Shapley Bar Plots for all 5 Individual Countries.