Multilingual Disparities in LLM-Based Safety Judgments: Evidence from Brand Safety Applications

Songjiang Liu[1] **Riley Grossman**[1] **Mike Smith**[2] **Cristian Borcea**[1] **Yi Chen**[1]
[1] New Jersey Institute of Technology [2] Indiana University Bloomington
{sl947, rag24, borcea, yi.chen}@njit.edu
ms255@iu.edu

Abstract

Multilingual LLMs are increasingly used as context-aware judges in real-world information systems under the assumption that equivalent content receives equivalent judgments across languages. We examine this assumption through brand safety, a global application where automated ratings can affect advertisers' reputations, publishers' revenues, and users' access to news. We construct a benchmark of LLM-generated safety ratings for 10,467 semantically aligned news articles across 13 languages. We find systematic cross-lingual disagreement appearing in more than 96% of cases where at least one language receives a non-zero risk rating. Suitability ratings differ significantly by language, controlling for run, category, and article. In the main model, English, German, and French content is generally rated more strictly, while Polish, Hungarian, Greek, Turkish, and Persian content is rated more leniently. Robustness checks with two additional LLMs show that significant language effects persist, though directional patterns vary by model. These findings show that multilingual LLM safety judgments can produce unequal outcomes for semantically equivalent content.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in real-world multilingual information systems. Research has explored practical applications across industries, including online advertising (Feizi et al., 2024) and brand safety (Levi et al., 2025). Brand safety systems are used by advertisers to ensure that their ads are not displayed alongside harmful, inappropriate, or otherwise unsuitable content. These systems also gauge the extent to which publishers' content is monetized because many advertisers do not bid to show ads near unsafe content. A recent survey of 200 companies spending more than $10 million annually on advertising found that more than 80% were concerned with brand safety (Johnson et al., 2023).

Although brand safety has historically relied on keyword-based methods or simple natural language processing (NLP) tools, recent advances in LLM capabilities have made brand safety a promising application area for LLMs (Pragad, 2024; Schiff, 2025; Levi et al., 2025). For example, Levi et al. (2025) from ZEFR Inc [1] benchmark the performance of zero-shot multimodal LLMs for brand-safety classifications, and ZEFR's Media Ratings Council accreditation for content-level brand-safety reflects the industry's approval of such systems (ZEFR, 2026).

One open question is whether LLMs are capable of evaluating content safety, especially for news articles, given the multilingual nature of news, where the same story is reported in many languages. Specifically, when the same prompt is used, do LLMs rate semantically equivalent content in different languages equally in terms of brand safety? Prior work has demonstrated 1) cross-lingual disparities in the safeness of LLMs' generated texts (Wang et al., 2024; Yong et al., 2025), and 2) inconsistencies when using LLMs as a judge (Fu and Liu, 2025; Chung and Freienthal, 2026). Although these cross-lingual disparities are typically attributed to lack of training data for low-resource languages, a second issue arises in brand safety judgments: LLMs have been shown to encode the cultural biases of the languages they are trained on (Tao et al., 2024). Thus, multilingual brand safety ratings may exhibit additional inconsistencies from encoded biases that distort an LLM's idea of "safe" based on the content's language.

We specifically evaluate cross-lingual inconsistencies in the context of brand safety ratings for news articles for two main reasons. First, this con-

[1] https://zefr.com/

text provides a clean experimental setup for testing multilingual classification disparities. Large multilingual news publishers, such as Euronews [2], provide professionally translated versions of the same article in several languages, allowing direct cross-lingual comparisons without relying on machine translation. Second, safety classifications of news content have important business implications, yet are underexplored.

We perform a large-scale evaluation of LLM-based brand-safety classifications for 10,467 distinct Euronews articles across 13 languages (totaling more than 70,000 language versions). Each language version is rated for safety relative to 15 safety categories by the cost-effective, general-purpose MiniMax-M2.5 LLM. We show our findings are robust to the choice of LLM by rating a sample of articles with `gemini-3.1-flash-lite-preview` and `deepseek-v4-pro` (Section 5).

The key contributions of this paper are:

- This is the first study to evaluate multilingual disparities in LLM judgment for brand safety, a global application where automated ratings affect advertisers' reputations, publishers' revenues, and users' access to news.
- Our findings show the high prevalence of disparities: Cross-lingual disagreement appears in more than 96% of cases where at least one language receives a non-zero risk rating. These disparities are statistically significant (Section 4.2).
- We identify patterns in the direction and magnitude of the language version disagreements to identify languages that are rated more strictly (e.g., English, German, and French) or more leniently (e.g., Polish, Hungarian, Greek, Turkish, and Persian) (Section 4.3).
- Case studies of model explanations illustrate that some disagreements involve different applications of safety boundaries, rather than obvious misunderstandings of article content (Section 4.4).
- Finally, we release a benchmark of LLM-generated safety ratings (with article URLs, prompts, and model reasoning traces) for 10,467 multilingual news articles to foster future research on LLMs for multilingual safety judgments [3].

Despite our chosen context, our findings are relevant for any application in which LLMs rate content safety, including content moderation (Masud et al., 2024; Kolla et al., 2024).

2 Related Work

Multilingual LLM judgments. Using LLMs to classify the safety of content is an example of LLM-as-a-judge applications. Recent evaluations of multilingual LLM-as-a-judge applications have found inconsistent judgments across languages. In particular, inconsistencies have thus far been observed when LLMs judge the correctness of answers (Fu and Liu, 2025) or the coherence and fluency of synthetic customer support dialogues (Chung and Freienthal, 2026) across multiple languages.

Our evaluation context is distinct because inconsistent safety ratings may in part result from the LLM's cultural biases inherited from training data, rather than just insufficient data in low-resource languages. We further contribute to the literature on multilingual LLM judgments by documenting inconsistencies even when the model provides explanations for each rating. This contrasts with prior work suggesting that inconsistencies can be mitigated by producing explanations alongside judgments (Fu and Liu, 2025). Finally, our analyses have important implications for publishers, advertisers, and social media platforms due to the current and proposed applications of LLMs to rate the safety of news (Schiff, 2025) and user-generated content (He et al., 2026; Hunsberger, 2025).

LLM Classification of Content Safety. Prior work has shown that the most popular brand safety providers (using proprietary NLP systems) are highly inconsistent (Smith et al., 2026) and regularly allow their advertising partners to place ads next to extremely unsafe content (Vekaria et al., 2024). This has motivated the application of LLMs for brand safety classification (Schiff, 2025; Pragad, 2024), with reasonable accuracy demonstrated for video content, albeit only on two safety categories (Levi et al., 2025).

LLMs have similarly been applied to the related area of content moderation for social media platforms. Despite a reasonably high overall accuracy, studies show that LLMs may fail to fol-

[2] https://www.euronews.com

[3] https://github.com/sl947/langdiff

low complex moderation rules (Kolla et al., 2024) and have significantly different outcomes across LLMs (Fasching and Lelkes, 2025). Multilingual assessments have shown success for fine-tuned LLMs (Upadhayay and Behzadan, 2025), but accuracy across languages differs due to insufficient data or translation resources in some languages. Masud et al. (2024) finds that, in zero-shot hate speech detection, providing the language or country of the annotator improved accuracy. This suggests that LLMs are capable of utilizing learned cultural differences to rate the safety of content when prompted.

In this study, we go one step further by showing that language-associated model behavior may affect safety ratings even when cultural context is not explicitly requested. We further differentiate our work from prior studies by comparing LLMs' safety ratings for semantically aligned news content across languages to evaluate cross-lingual consistency.

3 Empirical Setup

Multilingual news collection. We scraped historical multilingual news articles from the Euronews website, where many stories are published in multiple language versions. The collected dataset[4] contains 10,467 unique articles, with 70,041 total language versions across 13 languages. Table 1 reports the number of language versions by language.

Rating protocol and collection. Each language version was rated against the GARM Brand Safety Floor + Suitability Framework[5]. The brand safety floor captures whether the content violates a threshold for unsafe or harmful content, whereas the brand suitability captures a graded risk for sensitive but not necessarily unsafe content. We enforced a constrained output schema to return, for each of the 15 GARM categories (i.e., each category is a content-risk dimension to classify the type of potentially unsafe or unsuitable material): (1) a Brand Safety Floor decision (`safe` or `not_safe`) and (2) a suitability risk level in {`0`, `low`, `medium`, `high`}, which are coded as $\{0, 1, 2, 3\}$. For each language version, we ran the LLM model with the same prompt three times.

Language	Code	Versions	Average Words
English	en	9,666	554.3
Spanish	es	6,549	625.8
Portuguese	pt	6,335	591.2
Turkish	tr	6,028	424.6
Italian	it	5,996	567.0
Greek	el	5,688	576.7
French	fr	5,643	627.2
Polish	pl	5,228	461.5
German	de	5,045	532.7
Russian	ru	5,023	444.6
Hungarian	hu	3,931	399.8
Persian (Farsi)	fa	2,535	579.0
Arabic	ar	2,374	502.7
Total		**70,041**	**536.1**

Table 1: Euronews Multilingual Crawl Statistics

Directional magnitude. To summarize the direction of cross-language disagreement, we compute a signed conditional shift. For each language–category pair (l, c), we compare l with every other available language m for the same article a, retaining only comparisons where the two languages receive different suitability levels:

$$\Delta_{l,c} = \frac{1}{|D_{l,c}|} \sum_{(a,m)\in D_{l,c}} (r_{a,l,c} - r_{a,m,c}),$$

$$D_{l,c} = \{(a, m) : m \neq l,\ r_{a,l,c} \neq r_{a,m,c}\},$$

Because $\Delta_{l,c}$ is conditioned on disagreements involving language l in category c, it measures whether a given language is systematically assigned higher risk (positive values) or lower risk (negative values) than its peers when a cross-language disagreement occurs.

4 Empirical Analysis

Few articles receive floor-level unsafe ratings given Euronews' reputation. Accordingly, we focus our analyses on the suitability-level safety ratings. Section 4.1 quantifies how often cross-lingual disagreement occurs by category, regardless of disagreement magnitude. Section 4.2 then tests whether these differences are statistically significant at both overall and category-specific levels. Section 4.3 characterizes the direction and magnitude of disagreements to explore language differences. Finally, Section 4.4 presents a case study of LLM explanations on three articles with cross-language safety rating differences to help understand why disparities exist.

[4] Almost all articles (99.9%) were published in 2025 or 2026.

[5] Though discontinued in 2024 (World Federation of Advertisers, 2024), GARM's safety categories and high-level definitions remain useful as a brand-safety taxonomy and align with leading vendors' rating systems (Integral Ad Science, 2024).

Category	Disagreement Prevalence	Conditioned on non-0
Adult/Sexual	6.2%	94.3%
Alcohol	3.5%	86.8%
Arms/Ammunition	28.8%	96.9%
Crime	45.1%	97.0%
Death/Injury/Military	56.4%	95.5%
Hate Speech	21.3%	98.3%
Illegal Drugs	4.3%	94.3%
Misinformation	65.6%	98.3%
Obscenity/Profanity	9.1%	99.3%
Online Piracy	1.6%	98.0%
Other	20.5%	99.7%
Sensitive Social Issues	86.4%	95.4%
Spam/Malware	2.8%	96.6%
Terrorism	34.4%	98.9%
Tobacco/Vaping	1.2%	85.4%
Total	25.8%	96.8%

Table 2: Language Disagreements by Category

4.1 Cross-Lingual Disagreements.

We first quantify its prevalence at the article–category level. For each pair, we compare suitability risk levels across all available language versions of the article, marking a disagreement whenever there are at least two different ratings of the suitability risk level. Table 2 reports the *raw prevalence* as a percentage of all article-category pairs and a *conditional prevalence* that removes the article–category pairs where every language version receives a zero-risk rating for the category. We condition on non-zero cases to avoid base-rate dilution in safety categories that rarely appear (e.g., fewer news articles discuss online piracy than sensitive social issues).

Overall, cross-lingual disagreements are common (25.8% of all article-category pairs), and nearly ubiquitous (96.8%) when focusing solely on article-category pairs where at least one non-zero rating is observed. We also observe strong category differences. Unsurprisingly, raw disagreement prevalence is higher for safety categories relating to topics commonly discussed in news such as *Sensitive Social Issues* (86.4%) and *Death/Injury/Military* (56.4%). For other categories with very low disagreement prevalence (e.g., Online Piracy or Alcohol), the overall presence of cross-lingual disagreements is low because it is easier for the ratings to agree when the article clearly discusses nothing related to a safety category. However, when conditioning on at least one unsafe rating (i.e., a proxy for the article discussing topics at least related to the safety category), cross-lingual disagreements reach at least 85.4% for all categories. Categories such as *Online Piracy* and *Tobacco/Vaping* illustrate this pattern most clearly: they have low raw prevalence (1.2-1.6%) yet high conditional disagreement (85.4-98.0%).

4.2 Are Language Differences Statistically Significant?

Panel A: Full FE ANOVA on Suitability Risk Level

Term	Sum Sq	df	F	p
Run	0.29	2	0.983	0.374
Language	14.365	12	8.115	< 0.001
Category	139527.392	14	67559.738	< 0.001
Lang. $\times$ Cat.	508.481	168	20.517	< 0.001
Article	57404.703	9302†	41.834	< 0.001
Residual	457463.213	3101078	–	–

Panel B: Per-Category Language Effect

Category	p_{lang}	q_{FDR}	Sig.
death_injury_military	$< 10^{-4}$	$< 10^{-4}$	Yes
arms_ammunition	$< 10^{-4}$	$< 10^{-4}$	Yes
terrorism	$< 10^{-4}$	$< 10^{-4}$	Yes
obscenity_profanity	$< 10^{-4}$	$< 10^{-4}$	Yes
other	$< 10^{-4}$	$< 10^{-4}$	Yes
crime	$< 10^{-4}$	$< 10^{-4}$	Yes
misinformation	$< 10^{-4}$	$< 10^{-4}$	Yes
sensitive_social_issues	$< 10^{-4}$	$< 10^{-4}$	Yes
hate_speech	0.0003	0.0004	Yes
adult_sexual	0.0020	0.0030	Yes
online_piracy	0.0461	0.0628	No
illegal_drugs	0.4091	0.5114	No
spam_malware	0.7152	0.8007	No
alcohol	0.7594	0.8007	No
tobacco_vaping	0.8007	0.8007	No

Table 3: Overall and per-category ANOVA results to test for language differences. †Article degrees of freedom reflect retained `rss_title` levels with at least two language versions, minus one, not the nominal corpus size.

To test whether the observed cross-lingual disagreements reflect systematic language differences rather than noise, we conduct inferential tests on the risk level for all article–language–category–run observations. Table 3, Panel A reports a full fixed-effects ANOVA, `risk level` $\sim$ `run` + `language` + `category` + `language` $\times$ `category` + `article`. The article fixed effects absorb differences in the stories available in each language. Thus, the language term measures whether safety category ratings for the same stories still significantly differ by language. The interaction term tests whether language disparities vary by category rather than reflecting a single uniform shift.

The results confirm the significance of language differences. First, the language effect is statistically significant even when controlling for stochasticity across the three runs and article fixed effects. The run effect is the only insignificant term, indicating that observed cross-lingual disparities are not

explained by run-to-run stochastic variation. Unsurprisingly, category and article effects are large and highly significant, confirming that suitability risk depends strongly on the safety category and on the specific news story being rated. Finally, the language-by-category interaction is significant, meaning that the language effect is heterogeneous: a language that receives higher (or lower) safety ratings in one category does not necessarily exhibit the same pattern in other categories.

To further explore differences across categories, Panel B in Table 3 reports per-category language tests with the Benjamini–Hochberg correction. Only the five infrequently violated safety categories with a disagreement prevalence of under 5% in Table 2 show nonsignificant language effects.

4.3 Directional Patterns and Magnitudes of Disparity.

ANOVA establishes the significance, but not the direction and magnitude of cross-lingual disparities. We calculate the directional magnitude for each language–category pair as discussed in Section 3 to explore which languages receive relatively stricter (or more lenient) LLM safety ratings for each category. The results are visualized in Figure 1. Crucially, small directional magnitudes do not mean that disagreements occurred rarely, only that the disagreements did not routinely assign stricter (or more lenient) ratings for a particular language. For example, directional magnitudes are relatively low for *Death/Injury/Military*, but this is plausible because the context (e.g., which country's military forces are discussed) could prevent systematically stricter safety ratings for any one language.

Across all safety categories, several trends emerge. Articles in Arabic, Greek, Hungarian, Polish, and Turkish tend to be rated more leniently. In contrast, German, English, and French articles tend to be rated more strictly across almost all categories. One potential explanation is that the greater availability of training resources for Western languages has led to models that are better suited for detecting safety concerns in texts from these high-resource languages.

Within specific categories, there are both unsurprising and counterintuitive findings. Unsurprisingly, articles in French and German are rated most strictly for hate speech. Germany and France have two of the strictest hate speech laws (Lewis, 2022). On the other hand, articles in Arabic and Polish are tied for the most lenient in rating the safety of alcohol content. While Poland has one of the highest per capita alcohol consumption rates, many predominantly Arabic-speaking countries rank at the bottom of the list with strict anti-alcohol laws (World Population Review, 2026). These findings highlight how the observed cross-lingual disagreements could result from inherited cultural biases or from insufficient training resources.

4.4 Case Study: Hate Speech Category.

We provide a case study of cross-lingual disagreements in `Hate Speech` ratings for three different news articles [6].

In Table 4, we present the reasoning excerpts and ratings across all three runs to show the disagreements are persistent, and not due to noise in a single run. Across the three examples, the articles in more strictly rated languages (Arabic, Spanish, and German) consistently result in a medium or high risk level, while the article in the more leniently rated language (English, Hungarian, and Greek) is rated as completely safe in all three runs. In each example, the disagreement is not about whether the article mentions a sensitive topic, but about how the model applies the boundary between hate speech content and news content that reports on events *related* to hate speech.

One possible explanation is that some languages make the model more likely to rate mere discussions of unsafe topics (i.e., as commonly done in breaking news articles) as unsafe. In some cases, this behavior may be due to cultural biases or sensitivities that protect (or disregard) certain groups. For example, Germany's particular sensitivity to anti-Semitic hate speech could explain the harsher rating assigned to the German language version of an article about a man stabbed at a Holocaust memorial (see third example in Table 4).

5 Robustness to Model Choice

To test whether our findings are robust to model choice, we collected brand safety ratings for 1,000 randomly sampled news articles using `deepseek-v4-pro` and `gemini-3.1-flash-lite-preview`. The full fixed-effects ANOVAs in Table 5 show that the effect of language remains significant after controlling for run, category, and article in both models. Thus, cross-lingual disagreements in brand safety ratings are not unique to MiniMax-M2.5.

[6] IDs 1320, 1819, and 4786 in our dataset.

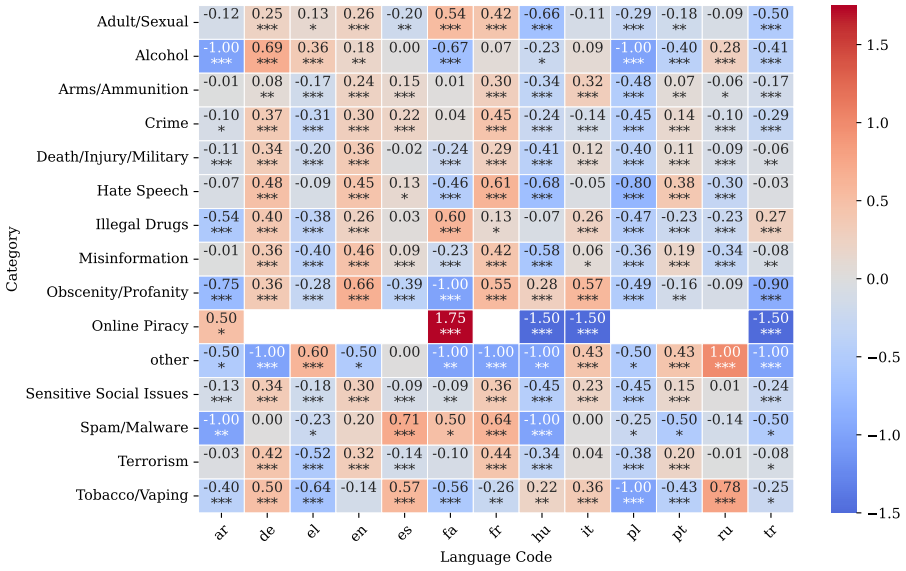

Figure 1: Average directional magnitude of cross-lingual disagreements by language and category. Each cell shows the language-specific average suitability risk level minus the average for all other languages in the same category. Asterisks denote Benjamini–Hochberg FDR-adjusted two-sided Wilcoxon signed-rank tests of each cell's signed shift distribution against zero: $^{*}q<0.05$, $^{**}q<0.01$, and $^{***}q<0.001$.

Article	Pair	Higher Reasoning Excerpt (3 runs)	Lower Reasoning Excerpt (3 runs)
Ai Chatbot Grok Restricts Image Generation After Global Backlash To Sexualised Deepfakes	ar > en	**R1:** `hate_speech`: The deepfakes targeted women; the floor covers content inciting hatred or dehumanizing protected groups; this is High because it dehumanizes women through non-consensual sexual imagery. **R2:** `hate_speech` is a critical category; targeted non-consensual deepfake imagery against women is treated as gender-based harassment and High. **R3:** `hate_speech`: targeting women can constitute hate speech/gender-based violence, but framed as news reporting; assigned Medium.	**R1:** `hate_speech`: the article does not contain hate speech itself; it reports on harmful content; level 0. **R2:** hate speech does not apply here because the focus is sexual exploitation rather than hate targeting protected groups; level 0. **R3:** hate-speech evaluation finds no protected group directly targeted in the article text; no violation / level 0.
Us Suspends Immigrant Visa Processing For Nationals Of 75 Countries White House Says	es > hu	**R1:** `hate_speech`: quotes describing Somalis as "trash" and telling them to "go back" are hateful/denigrating by national origin; assigned High. **R2:** same quoted statements are identified as hate speech targeting a national/ethnic group, but because they are reported in news context, assigned Medium. **R3:** quotes are judged to vilify/dehumanize by national origin; after considering reporting context, reasoning still marks High.	**R1:** `hate_speech`: immigration-policy reporting with no hate speech content; level 0. **R2:** no hate speech; policy discussion does not promote hatred; level 0. **R3:** rhetoric may be politically charged but not hateful/inciting/dehumanizing under the framework; level 0.
Berlin Man Stabbed At Holocaust Memorial Hours After Arrest Over Planned Attack On Israeli Embassy	de > el	**R1:** hate speech is treated as a critical concern: explicit intent to kill Jews is framed as a serious hate crime and High severity. **R2:** `hate_speech`: attacker statement about wanting to kill Jews is recognized as hate-speech context, but reported as news rather than promoted; assigned Medium. **R3:** reasoning flags direct threat against a protected religious group and explicit violent intent, concluding High.	**R1:** `hate_speech`: no direct hate speech content; level 0. **R2:** reasoning focuses on terrorism/death-injury news framing and does not trigger hate-speech criteria; level 0. **R3:** no explicit hate speech or targeted discrimination is identified in the reporting; level 0.

Table 4: Case study with paired high-vs-low language ratings and reasoning excerpts from all three runs. Red text highlights key explanations for the assigned risk level.

Figure 2 shows the corresponding directional magnitudes for both models. In both heatmaps, directional magnitudes are regularly significant and follow patterns similar to Figure 1. Gemini's ratings of German, English, and French content are relatively strict, highly aligned with MiniMax. DeepSeek's ratings are similar for German and English content, but much more lenient for French content. Similarly, all models show lenient ratings for Greek and Polish content. Interestingly, both Gemini and DeepSeek generate relatively stricter ratings for Arabic content than MiniMax. These differences could reflect differences in model safety alignment or the language composition of the training datasets. Importantly, this does not impact our main practical concern for brand-safety or content moderation deployment: content language may affect safety ratings in ways that alter monetization and moderation outcomes for semantically aligned content.

6 Discussion

Implications for multilingual LLM applications. Cross-lingual disagreements are frequent among several consequential categories (Table 2), and language effects remain significant after controlling

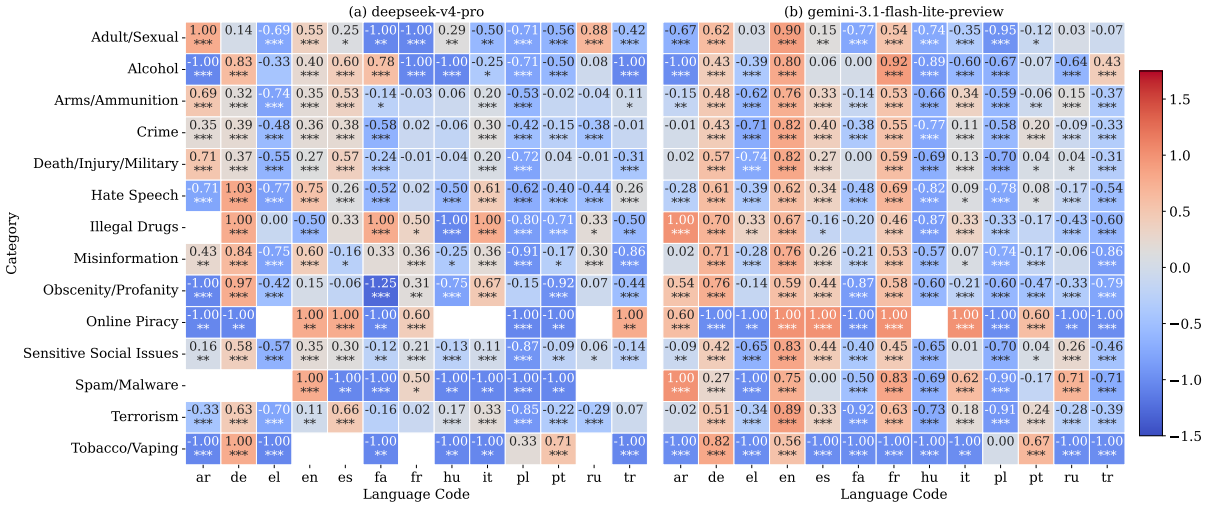

Figure 2: Average suitability-risk-level disparity by language and category for `deepseek-v4-pro` and `gemini-3.1-flash-lite-preview`.

Model	Term	Sum Sq	df	F	p
`deepseek-v4-pro`					
	Run	0.065	2	0.235	0.790
	Language	5.359	12	3.227	< 0.001
	Category	23837.421	14	12303.130	< 0.001
	Article	8303.354	878†	68.335	< 0.001
	Lang. × Cat.	98.860	168	4.252	< 0.001
	Residual	40803.103	294834	–	–
`gemini-3.1-flash-lite-preview`					
	Run	0.000	2	0.000	1.000
	Language	8.703	12	4.575	< 0.001
	Category	34845.565	14	15702.120	< 0.001
	Article	11675.709	878†	83.893	< 0.001
	Lang. × Cat.	129.043	168	4.846	< 0.001
	Residual	46736.843	294848	–	–

Table 5: Cross-lingual language effects persist across additional model choices. Each fixed-effects ANOVA includes run, language, category, article, and a language-by-category interaction. † Article degrees of freedom reflect retained `rss_title` levels with at least two language versions, minus one, not the nominal corpus size.

for category, run, and story (Table 3). Thus, even within a fixed prompt and LLM, semantically equivalent content can receive different ratings depending on its language. Although not every model will exhibit the exact same trends (e.g., Arabic content rated more strictly by DeepSeek and Gemini than MiniMax), our findings in Section 5 show that disparities occur across several contemporary models. Thus, cross-lingual consistency should be measured as part of deployment for any multilingual LLM applications, not assumed from general multilingual capability.

Implications for Publishers, Advertisers, and Society. In the context of brand safety, rating inconsistencies can affect both economic and informational outcomes. For multilingual publishers, cross-lingual inconsistencies may lead to uneven advertising revenues. These revenue disparities may reduce investment in some language editions, limiting information access for their audiences. For advertisers, the presence of several languages that are rated leniently across most or all safety categories (e.g., Turkish or Polish content) means that advertisers placing ads next to that content may be opening themselves up to reputational damage by regularly advertising next to unsafe content.

In the content moderation context, where LLMs may also be used to judge content safety, cross-lingual inconsistencies have several undesirable impacts. First, users may unfairly have their content removed or demonetized (e.g., on YouTube) solely based on the language of their content. Second, platforms may unknowingly allow certain communities to create more toxic and less safe environments by communicating in languages that are more leniently moderated due to limited training resources or cultural biases.

Possible explanations for multilingual disparities. Disagreements may sometimes be the result of inherited cultural biases (Tao et al., 2024). One suggestive pattern is that some language–category directional magnitudes align with known laws (e.g., strict judgments of hate speech in German or French content). However, cross-lingual disparities may also arise from the uneven distribution of languages in the training corpus or safety-alignment datasets. The imbalanced datasets could lead to understandings of safety that are better optimized for English or Western cultures and languages. Given

that these case studies provide qualitative rather than causal evidence, future work should test these conjectures directly. For example, synthetically generated multilingual content could be counterfactually edited to test the safety ratings of articles when the identities of the articles' subjects are changed.

Recommendations. Our findings suggest several practical recommendations for multilingual LLM rating systems. First, systems should be evaluated on semantically aligned multilingual benchmarks before deployment, with results reported by language, category, and their interaction rather than only as aggregate accuracy or agreement. Second, operators should continually track cross-language consistency for high-impact, context-dependent categories where framing judgments are central. Third, when inconsistencies are persistent, systems should route high-disagreement cases to human review and consider language-specific validations, prompts, or thresholds only after checking that such adjustments do not create new inequities.

We also provide several specific recommendations for the brand safety context. Specifically, we believe brand safety companies should transparently share the utilized models, prompts, and evaluation results with advertisers and publishers. Furthermore, mechanisms should be put in place for publishers or advertisers to contest classification decisions. If implemented, such mechanisms would allow multilingual publishers to contest ratings that were inconsistent across languages.

Finally, we note that some advertisers may want different safety ratings based on the content's language because of cultural differences in what content is acceptable. We do not believe that cultural differences should be incorporated by the LLM without being prompted to do so, as we have observed in our analyses. In such cases, advertisers should communicate these preferences to brand safety providers to ensure that they are differentially protected based on content language and/or audience.

7 Conclusion

We presented a benchmark of LLM-based brand-safety ratings for 10,467 multilingual news articles across 13 languages and 15 safety categories. Cross-lingual disagreements are common, especially when at least one language version receives a non-zero risk rating, and language effects remain significant after controlling for run, category, and article. Across the three LLMs used in the paper, several languages are rated more strictly than others. These results show that multilingual LLM safety judgments can produce unequal outcomes for semantically aligned content.

8 Limitations

This study identifies systematic cross-lingual inconsistencies, but we do not identify the causal mechanisms behind each discrepancy. The model's reasoning traces help inspect individual cases, but they should be treated as diagnostic evidence rather than faithful explanations of the rating process. Future work should explore possible mechanisms relating to uneven distributions in training or safety-alignment datasets. Furthermore, future work could use counterfactual evaluations to test whether inherited cultural biases may impact language-specific risk perceptions.

The language coverage is limited by Euronews' publication footprint. The dataset is biased toward European Union languages. Although included, the numbers of Arabic, Persian, Russian, and Turkish articles are lower than those of the other languages. The dataset does not cover many Asian, African, Indigenous, or otherwise underrepresented languages. The results should therefore be read as an evaluation of one multilingual news setting rather than a complete map of cross-lingual brand-safety behavior.

Finally, technical mitigation approaches remain outside the scope of this study. Our mitigation recommendations are focused on practical industry-level interventions that may alleviate some issues arising from cross-lingual safety disagreements in the brand safety context. Future work should evaluate technical implementations (e.g., cultural debiasing or prompting the model to mitigate cultural biases).

Acknowledgments

Research reported in this publication was supported in part by the NSF under Grant No. CNS2237328, the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UM1TR004789, as well as by the Martin Tuchman '62 Chair Endowment and the Leir Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

References

Isaac Chung and Linda Freienthal. 2026. Cross-lingual stability of LLM judges under controlled generation: Evidence from Finno-Ugric languages. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 133–148, Rabat, Morocco. Association for Computational Linguistics.

Neil Fasching and Yphtach Lelkes. 2025. Model-dependent moderation: Inconsistencies in hate speech detection across LLM-based systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22271–22285, Vienna, Austria. Association for Computational Linguistics.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. 2024. Online advertisements with llms: Opportunities and challenges. *Preprint*, arXiv:2311.07601.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Jiahui He, Yiluo Wei, and Gareth Tyson. 2026. Enhancing content moderation with llms: A reddit case study on evaluating and refining human decisions. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 9823–9834, New York, NY, USA. Association for Computing Machinery.

Alice Hunsberger. 2025. How to use llms for content moderation. Musubi.

Integral Ad Science. 2024. Our brand safety and suitability overview. Published March 6, 2024.

Ross W. Johnson, Clay Voorhees, and Farnoosh Khodakarami. 2023. Is your brand protected? assessing brand safety risks in digital campaigns. *Journal of Advertising Research*, 63(3):205–221.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Adi Levi, Or Levi, Sardhendu Mishra, and Jonathan Morra. 2025. Ai vs. human moderators: A comparative evaluation of multimodal llms in content moderation for brand safety. *Preprint*, arXiv:2508.05527.

Marco Lewis. 2022. The netzdg and the avia law: How two different legal systems created two different outcomes from similar laws. *Wis. Int'l LJ*, 40:491.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863, Miami, Florida, USA. Association for Computational Linguistics.

Dev Pragad. 2024. Don't cancel brand safety – improve it. AdExchanger.

Allison Schiff. 2025. Ai is helping brand safety break free from blocklists. AdExchanger.

Michael Smith, Riley Grossman, Antonio Torres-Aguero, Pritam Sen, Cristian Borcea, and Yi Chen. 2026. Inconsistencies in classification of online news articles: A call for common standards in brand safety services. *Preprint*, arXiv:2601.01303.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Bibek Upadhayay and Vahid Behzadan. 2025. X-guard: Multilingual guard agent for content moderation. In *Proceedings of the First Workshop on LLM Security (LLMSEC)*, pages 54–86, Vienna, Austria. Association for Computational Linguistics.

Yash Vekaria, Rishab Nithyanand, and Zubair Shafiq. 2024. The inventory is dark and full of misinformation: Understanding ad inventory pooling in the ad-tech supply chain. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1590–1608.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.

World Federation of Advertisers. 2024. Wfa discontinues garm. Published August 9, 2024.

World Population Review. 2026. Alcohol consumption by country 2026. Accessed: 05-01-2026.

Zheng Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. 2025. The state of multilingual LLM safety research: From measuring the language gap to mitigating it. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15845–15860, Suzhou, China. Association for Computational Linguistics.

ZEFR. 2026. Zefr is the first third-party partner to receive mrc accreditation for content-level brand safety and suitability reporting on youtube. ZEFR Press Release.