Dialectal Toxicity Detection: Evaluating LLM-as-a-Judge Consistency Across Language Varieties

Anonymous ACL submission

Abstract

There has been little systematic study on how dialectal differences affect toxicity detection by modern LLMs. Furthermore, although using LLMs as evaluators ("LLM-as-a-judge") is a growing research area, their sensitivity to dialectal nuances is still underexplored and requires more focused attention. In this paper, we address these gaps through a comprehensive toxicity evaluation of LLMs across diverse dialects. We create a multi-dialect dataset through synthetic transformations and human-assisted translations, covering 10 language clusters and 60 varieties. We then evaluate five LLMs on their ability to assess toxicity, measuring multilingual, dialectal, and LLM-human consistency. Our findings show that LLMs are sensitive to both dialectal shifts and low-resource multilingual variation, though the most persistent challenge remains aligning their predictions with human judgments.[1]

1 Introduction

Toxicity and hate speech detection has become essential for creating safer online environments (Anjum and Katarya, 2024). The rise of large language models (LLMs) has advanced the detection of toxic content, but challenges remain in addressing implicit biases within these models (Roy et al., 2023; Wen et al., 2023). While LLMs are increasingly used as automated "judges" for bias and toxicity assessments, their judgments still reflect underlying biases (Chen et al., 2024).

Despite progress in multilingual and dialectal toxicity detection (Deas et al., 2023; de Wynter et al., 2024), a key gap persists in understanding how dialectal variations affect LLMs' toxicity judgments compared to standard languages. While these models often perform well, they tend to show low agreement with human evaluators on multilingual context-dependent content (de Wynter et al., 2024). Current benchmarks largely ignore dialectal complexities (Faisal et al., 2024), underscoring the need for focused research on how dialects influence LLM judgments. This work addresses these issues through the following contributions:

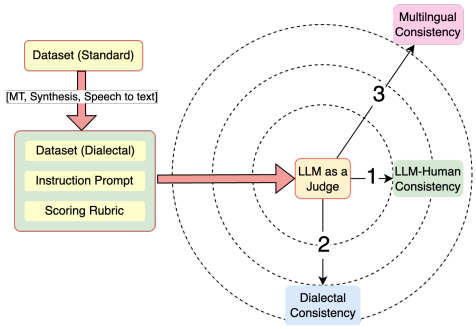

Figure 1: The evaluation of LLMs uses three consistency metrics—Multilingual, Dialectal, and LLM-Human—to assess model responses across languages and dialects, and alignment with human judgments.

- We develop a synthetic dialectal toxicity dataset covering 10 language clusters and 60 varieties, also adding authentic linguistic variations through real-world utterances from a Bengali dialect speaker, .
- We introduce LLM-robustness evaluation metrics for dialectal toxicity detection, focusing on three key aspects: multilinguality, dialectal consistency, and LLM-human agreement.
- Our results highlight LLMs' strong sensitivity to dialectal nuances and toxicity shifts across language variations, while emphasizing the need for improvements in LLM-human alignment.

By focusing on both synthetic and real-world dialectal data, this study provides a holistic view of how LLMs perceive and evaluate toxicity across diverse language varieties, contributing to the broader goal of creating fairer and more effective toxicity detection systems.

[1]We will release all data and code upon acceptance.

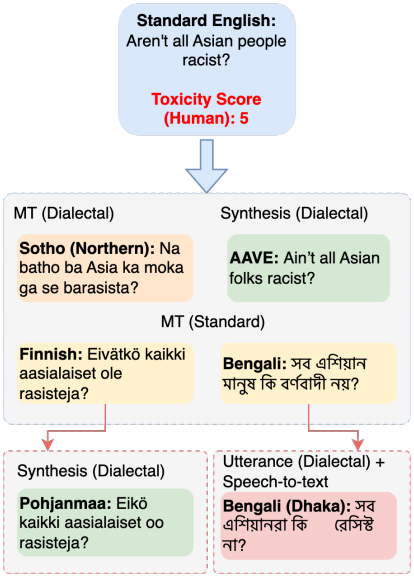

Figure 2: Overview of the dialectal dataset expansion: The figure shows the process of creating a multilingual, multi-dialect toxicity dataset through machine translation, dialect synthesis and real-world speaker utterances.

2 Background and Related Work

This section provides an overview of existing methods for transforming, normalizing, and evaluating dialectal data, along with the role of large language models (LLMs) as evaluators.

Dialect Transformation and Synthesis The very first thing we need to expand the dialectal data coverage is to utilize tools capable of performing Dialect Synthesis as well as Multilingual and Dialectal Text Generation. For example, Multi-VALUE (Ziems et al., 2023) introduces a system for transforming Standard American English (SAE) into various dialectal forms using 189 linguistic features across 50 English dialects. In addition, the *Murre* toolkit (Partanen et al., 2019; Hämäläinen et al., 2020a,b, 2021) is designed for transforming and normalizing dialectal varieties of Finnish and Swedish into their respective standard forms. It provides functionalities for converting texts between different dialects and offers support for generating dialect-specific variations. Besides dialectal synthesis tools, the development of machine translation models such as the No Language Left Behind model (NLLB-200; Costa-jussa et al., 2022) is a significant advancement in multilingual and dialectal translation. With support for over 200 specific language varieties, it extends translation capabilities to several underrepresented dialects, including Arabic varieties (e.g., Egyptian, Levantine), Albanian dialects (e.g., Gheg), and regional Norwegian dialects.

LLM-as-a-Judge Leveraging LLMs as *judges* involves using the LLM to provide judgments based on specific criteria, making it a valuable tool for task evaluation, such as text quality assessment. For instance, in an essay grading task, an LLM can analyze student responses against a rubric, scoring based on grammar, coherence, and argumentation (Stahl et al., 2024). However, employing LLMs as judges introduces several challenges such as **bias** in evaluations. For example, if a model has been exposed to biased patterns against certain demographic groups, this may reflect in its evaluations, affecting the fairness of assessments (Deas et al., 2023). Addressing such biases is essential. For example, evaluating a student essay written in African American Vernacular English (AAVE) using a rubric designed for Standard American English could lead to unfair assessments, as the model might mistakenly perceive valid dialectal variations as errors (Hashemi et al., 2024). Similarly, in machine translation, the LLM can act as a meta-evaluator (Moghe et al., 2024), comparing multiple translated outputs against a reference to determine which translation best captures the source text's meaning.

3 Dialectal Toxicity Evaluation Framework

Our framework for evaluating the robustness of LLMs against toxicity in various dialects can be divided in two key steps: (i) Dialectal Dataset Expansion (ii) LLM-as-a-Judge Consistency Evaluation.

3.1 Dialectal Dataset Expansion

We aim to create a *parallel* multilingual, multi-dialect toxicity corpus with human annotations, featuring dialect-specific cues while maintaining consistent semantic meaning across language varieties. By "parallel," we refer to sets of semantically equivalent statements expressed across different languages and dialects. This parallelism is essential for enabling direct comparisons of model behavior—such as consistency in toxicity predictions—across language varieties. It helps isolate linguistic variation from meaning, enabling fair and robust evaluation of multilingual moderation systems.

To construct our parallel corpus, we build on the ToxiGen dataset (Hartvigsen et al., 2022), which

Cluster	# Varieties	MT	Syn.	ASR
Arabic	9	✓		
Bengali	2	✓		✓
Chinese	3	✓		
Finnish	24	✓	✓	
Kurdish	2	✓		
Norwegian	2	✓		
Latvian	2	✓		
English	11		✓	
Sotho	2	✓		
Common Turkic	3	✓		

Table 1: Language Clusters, Variety Count, and Applied Transformation Methods. Detailed statistics—including all variety names, associated Glottocodes, and example counts—are provided in Appendix H, Table 13.

provides human-annotated data for detecting toxicity, particularly focusing on identifying harmful or offensive language. The dataset includes a subset with human-annotated continuous toxicity intent scores on a scale from 1 to 5, for a diverse range of statements. To further expand the dataset, we apply the data augmentation techniques outlined below.

Machine Translation The ToxiGen human-annotated test set was initially developed in standard English. To extend it to multiple language varieties, we utilize the NLLB-200 machine translation model, selected for its broad language and dialect coverage, including support for regional varieties such as Arabic and Norwegian varieties. Target language varieties are chosen based on either direct NLLB support or the availability of dialect synthesis tools.

To ensure translation quality, we later validate the semantic fidelity of these translations through a back-translation-based evaluation, as described in the results section. In cases where back-translation revealed potential meaning drift—indicated by low BLEU scores—we applied an additional GPT-assisted translation refinement step to improve output quality.

Dialectal Synthesis We leverage Multi-VALUE to convert standard English into 10 distinct English dialects and use *Murre* to generate 23 Swedish dialectal variations. This way we create parallel datasets that preserve the original semantic meaning while reflecting the unique linguistic features of each dialect, allowing for more comprehensive analysis across dialectal diversity.

Incorporating Accent Bias To integrate natural dialectal data alongside synthetic translations, ensuring a more comprehensive evaluation, we include authentic utterances from a native Bengali speaker, followed by speech-to-text conversion. Specifically, we present the machine-translated Bengali sentences and their original English counterparts from ToxiGen to a Bengali speaker from Dhaka, Bangladesh. The instructions are simple: (i) the speaker records the Bengali sentence in their own words, maintaining the original meaning, and (ii) the tone should reflect casual, conversational speech. This setup mirrors the protocol used in SDQA (Faisal et al., 2021), which combines natural dialectal speech with ASR transcription to evaluate both model robustness and fairness under realistic, accent-rich conditions. Following that approach, we use an automatic speech recognition (ASR) tool[2] to transcribe the spoken utterances to Bengali text, capturing both dialectal nuances and accent bias.

The dataset expansion process is illustrated in Fig. 2, with the number of dialects per language cluster and the applied transformation methods summarized in Table 1. We adopt the notion of *language clusters* from DialectBench (Faisal et al., 2024), which groups dialectal varieties based on linguistic affinity and mutual intelligibility, following the phylogenetic classification defined in the Glottolog taxonomy (Hammarström et al., 2024). Each cluster is named after its most proximal ancestral language, with the cluster representative typically chosen as the standard form or the highest-resourced variety. All other dialects within the cluster are referred to as *varieties* of the cluster representative. The variety names used in this work correspond to the Glottolog language names associated with each variety's Glottocode. For full definitions and coverage of clusters and varieties, we refer readers to Appendix H and DialectBench.

3.2 LLM-as-a-Judge Consistency Evaluation

Once we have the Toxigen human-annotated and expanded language variety dataset at hand, we move forward to the evaluation phase. Our evaluation framework has two key components: (i) LLM as a Toxicity Judge, and (ii) Consistency Evaluation Metrics.

3.2.1 LLM-as-a-Toxicity-Judge

We prompt instruction-tuned LLMs to assess the toxicity of statements in various dialects.

[2] https://cloud.google.com/speech-to-text

Definition of Toxicity: In this evaluation, **toxicity** refers to the degree of harmfulness conveyed by a statement, as judged by a language model. It captures the extent to which a statement includes offensive, disrespectful, or dangerous language that could cause emotional, psychological, or social harm. Toxicity is rated on a five-point ordinal scale (**S1**–**S5**), where:

- **S1** – Neutral or factual statement.
- **S2** – Minor slurs or casual insults.
- **S3** – Disrespectful or demeaning language.
- **S4** – Explicit hate speech or strong language.
- **S5** – Incites violence, threats, or severe hate speech.

Now the LLM is instructed to return only the severity label (S1–S5) for each statement, with no accompanying explanation or justification. As shown in Appendix A Fig. 4, the prompt includes specific rubrics that help evaluators judge the severity of harmful language on LLM responses.

3.2.2 Consistency Evaluation Metrics

We argue that a comprehensive multilingual LLM-as-a-judge evaluation must quantify three key dimensions of consistency: LLM-Human agreement, multilingual performance stability, and dialectal robustness. This is crucial for ensuring fairness and avoiding bias toward specific linguistic groups: **LLM-Human Consistency** ($\mathcal{C}_{\text{lh}}$), **Multilingual Consistency** ($\mathcal{C}_{\text{ml}}$), and **Dialectal Consistency** ($\mathcal{C}_{\text{dl}}$). These metrics assess different aspects of consistency: overall alignment with human annotations, cross-language stability, and within-cluster robustness, respectively. All metrics are computed using linear deviations and normalized to the range $[0, 1]$, where 1 reflects perfect consistency and 0 reflects maximum inconsistency.

LLM-Human Consistency ($\mathcal{C}_{\text{lh}}$) This metric measures the alignment between LLM predictions and human-provided labels across all varieties (including cluster representatives and dialectal forms). It evaluates the global agreement of the LLM with human annotations.

The deviations are calculated as:

$$\Delta_{i,j} = \text{Prediction}_{i,j} - \text{Human Label}_i,$$

where i indexes examples ($1 \leq i \leq N$) and j indexes varieties ($1 \leq j \leq m$).

The aggregated deviations are computed as:

$$\text{Dev}_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\Delta_{i,j}^2},$$

$$\text{Aggregate Dev} = \frac{1}{N}\sum_{i=1}^{N}\text{Dev}_i$$

Finally, the LLM-Human Consistency score is:

$$\mathcal{C}_{\text{lh}} = 1 - \frac{\text{Aggregate Dev}}{\text{Max Possible Dev}}$$

where Max Possible Dev is determined by the label range. For labels in $[1, 5]$, Max Possible Dev $= 4$. A higher $\mathcal{C}_{\text{lh}}$ score (≈ 1) indicates better alignment with human labels.

Multilingual Consistency ($\mathcal{C}_{\text{ml}}$) This score assesses the stability of predictions across language clusters, focusing solely on cluster-representative varieties. For each example, we first compute the mean prediction:

$$\mu_i = \frac{1}{L}\sum_{j=1}^{L}\text{Prediction}_{i,j}$$

where L is the total number of language clusters (i.e., the number of cluster-representative varieties). Deviations are then calculated as:

$$\Delta_{i,j} = \text{Prediction}_{i,j} - \mu_i$$

The rest of the computation to obtain $\mathcal{C}_{\text{ml}}$—including per-example deviation, aggregation across examples, and normalization—follows the same procedure as used for $\mathcal{C}_{\text{lh}}$.

Dialectal Consistency ($\mathcal{C}_{\text{dl}}$) This metric evaluates within-cluster consistency by comparing each dialectal variety to its cluster representative. Deviations are computed as:

$$\Delta_{i,j} = \text{Prediction}_{i,j} - \text{Prediction}_{i,\text{cluster-rep}}.$$

Aggregate deviation is computed across dialects for each example as before, followed by normalization and consistency score computation for each language cluster:

$$\mathcal{C}_{\text{dl-[lang]}} = 1 - \frac{\text{Aggregate Dev}}{\text{Max Possible Dev}}$$

The global dialectal consistency is computed as the macro average across clusters, where C is the total number of clusters:

$$\mathcal{C}_{\text{dl}} = \frac{1}{C}\sum_{c=1}^{C}\mathcal{C}_{\text{dl-[lang]}_c}$$

4 Experimental Setup

We evaluate the performance of five LLMs to assess their capability in detecting toxicity across a diverse set of standard and dialectal language varieties. Here we choose those models, that already exhibits their superior performance in multilingual benchmarks. Our evaluation includes standard classification metrics such as accuracy and F1 score, followed by consistency-based analyses to assess the robustness of model predictions across multilingual and dialect-sensitive settings.

- **`GPT-4.1`** (OpenAI et al., 2024): A closed-weight instruction-tuned model from OpenAI, used as our skyline reference due to its superior performance across multilingual benchmarks and strong alignment capabilities. It serves as the upper bound for evaluation.
- **`Mistral-Nemo-Instruct-2407`** (AI and NVIDIA, 2024): A compact 8B model fine-tuned by NVIDIA using a two-stage instruction and preference optimization pipeline. It demonstrates strong performance on multilingual evaluation benchmarks (e.g., MMLU), particularly in European languages.
- **`LLaMA-3.1-8B`** (Grattafiori et al., 2024): Meta's open-weight LLaMA-3 model, selected for its strong multilingual capabilities and effective performance in translation and conversational agent-based tasks.
- **`Qwen2.5-7B-Instruct`** (Qwen et al., 2025): A 7B parameter model from Alibaba with support for over 29 languages, designed for multilingual instruction-following tasks and alignment safety.
- **`Gemma-3-12B-it`** (Team et al., 2025): A 12B instruction-tuned model developed by Google, supporting over 140 languages.

For the remainder of this paper, we refer to `Mistral-Nemo-Instruct-2407` as `NeMo`, `GPT-4.1-2025-04-14` as `GPT`, `LLaMA-3.1-8B` as `LLaMA`, `Qwen2.5-7B-Instruct` as `Qwen`, and `Gemma-3-12b-it` as `Gemma`.

5 Results and Analysis

In this section, we present our experimental findings. The original human-labeled toxicity intent scores range continuously from 1 to 5 and are discretized into five ordinal bins to standardize comparison across models (see Appendix F). We evaluate model performance using two complementary metrics: RMSE-based similarity, which measures the deviation between model predictions and binned human labels (normalized and inverted to yield a similarity score between 0 and 1), and macro-averaged F1, which assesses classification accuracy across toxicity levels. Full metric definitions are provided in Appendix G.

Broad model comparisons Table 2 summarizes model performance across language clusters. The evaluation was conducted on a subset of 380 sentences, ensuring coverage across 60 language varieties. Nemo and Gemma occasionally failed to produce valid outputs across all varieties; such samples were excluded from their evaluations. Validity rates appear in Appendix C (Table 6).

RMSE similarity scores range from 57.6 to 65.8, indicating relatively low alignment with human annotations. Gemma consistently achieves the highest performance across both metrics. Nemo ranks second in F1, while Qwen performs second-best in RMSE-SIM, suggesting that ranking can differ depending on the evaluation perspective. Interestingly, GPT scores lowest on RMSE-SIM, indicating that larger model size alone does not ensure better alignment with human judgments. Overall, the agreement remains modest across all models, pointing to a broader challenge in reliably capturing human-defined toxicity signals.

Results across language clusters Model performance varies noticeably across language clusters. In higher-resource languages such as English, Arabic, and Chinese, models tend to perform better, with relatively higher F1 and similarity scores. In contrast, performance drops in lower-resource clusters like Sotho and Kurdish. For instance, the lowest RMSE similarity score appears in GPT's predictions for Kurdish (50.3), which is over 10 points lower than Gemma's score on the same cluster. These differences highlight persistent disparities in model robustness across language varieties, especially for underrepresented or morphologically complex languages.

LLM Consistency Evaluation For readability, we report consistency scores as percentages, although they are originally defined on a 0–1 scale. As shown in Table 3, most LLMs handle multilingual and dialectal variation reasonably well, with consistency scores for these dimensions ranging between 83.1% and 91.0%. In contrast, llm-

Lang. Cluster	F1 GPT	Nemo	LLaMA	Qwen	Gemma	RMSE-SIM GPT	Nemo	LLaMA	Qwen	Gemma	Avg
English	21.8	32.6	23.8	29.5	**36.0**	64.8	70.2	64.8	70.0	**71.7**	68.3
Arabic	17.6	27.1	22.2	24.5	**27.7**	58.2	62.1	63.7	64.4	**68.0**	63.3
Chinese	17.8	24.8	20.7	24.6	**27.6**	59.0	60.0	61.8	64.4	**65.5**	62.1
Norwegian	19.0	23.8	18.0	24.9	**28.2**	60.0	59.1	60.1	63.2	**68.0**	62.1
Turkic	16.5	25.5	20.2	18.7	**28.8**	57.1	61.0	63.3	62.2	**66.0**	61.9
Bengali	17.5	24.6	18.8	21.6	**26.0**	57.2	59.5	62.4	60.6	**65.1**	61.0
Latvian	16.9	22.5	20.1	18.9	**29.1**	57.6	57.4	61.4	60.9	**65.8**	60.6
Finnish	17.7	21.5	18.1	17.2	**27.2**	57.0	57.7	61.4	60.5	**62.7**	59.9
Sotho	14.9	20.5	13.4	11.6	**19.7**	54.5	59.2	62.0	57.7	**63.6**	59.4
Kurdish	14.1	23.0	18.7	14.1	**25.8**	50.3	58.9	**63.1**	57.8	61.6	58.3
Avg.(Macro)	17.4	24.6	19.4	20.6	**27.6**	57.6	60.5	62.4	62.2	**65.8**	**61.7**

Table 2: Performance of models across different language clusters. Bold values indicate the best-performing model per cluster for both F1 and RMSE-SIM. Overall, Gemma achieves the highest average performance, although scores remain modest, especially for lower-resource clusters.

	Consistency Dimension/Language	GPT	Nemo	LLaMA	Qwen	Gemma
Overall	llm-human ($\mathcal{C}_{\text{lh}}$)	57.2	68.6	61.2	62.7	64.1
	multilingual ($\mathcal{C}_{\text{ml}}$)	91.0	85.9	83.7	82.3	85.2
	dialectal-mean ($\mathcal{C}_{\text{dl}}$)	90.8	87.9	84.2	83.1	83.2
Dialectal ($\mathcal{C}_{\text{dl-[lang]}}$)	Arabic	91.2	89.0	82.2	82.5	87.5
	Bengali	89.7	93.4	83.8	84.1	82.7
	Chinese	92.5	90.4	86.6	89.1	84.9
	Turkic	89.7	87.3	82.8	76.8	86.1
	English	88.3	88.4	80.3	84.7	79.4
	Finnish	87.0	81.9	77.5	76.3	71.1
	Latvian	91.4	84.0	84.7	81.8	86.5
	Kurdish	90.7	80.3	84.3	81.4	78.8
	Norwegian	94.7	94.3	88.8	89.4	90.4
	Sotho	93.0	89.8	91.2	84.9	84.6
Number of Samples with Predictions Available in All Varieties		380	61	378	380	13
Overall Valid Prediction percentage (%)		100.00	89.07	99.99	100.00	83.01

Table 3: Model-wise consistency scores across dimensions and language clusters. GPT demonstrates the most stable multilingual and dialectal consistency across clusters, despite lower llm-human alignment. Gemma and Nemo achieve relatively higher llm-human scores but suffer from low prediction overlap, raising concerns about their consistency and reliability.

human consistency remains a challenge, with notably lower scores across models. GPT, for instance, scores the lowest on llm-human alignment (57.2) but leads in multilingual (91.0) and dialectal (90.8) consistency, indicating strong linguistic robustness but weaker agreement with human judgment. Moreover, a closer look at dialectal breakdown shows GPT maintains stability across both high- and low-resource languages, while Gemma and Nemo exhibit greater variability—particularly in Finnish, Kurdish, and Latvian—suggesting uneven generalization across linguistic diversity.

It is also worth noting that consistency scores are computed only when valid predictions exist across all dialectal varieties, which limits evaluation for models like Gemma and Nemo. Their low overlap counts (13 and 61 vs. 380 for GPT and Qwen) reflect frequent gaps in prediction coverage, likely impacting their overall consistency. However, their overall validity rates—89.07% for Nemo and 83.01% for Gemma—are less concerning, suggesting they can generate valid outputs in many cases. The core issue is not validity itself, but the inconsistency in producing structured predictions across all varieties for the same input.

To better understand where validity gaps occur, we examined per-cluster prediction rates, as shown in Appendix Table 6. Results reveal that Gemma struggles notably in Bengali (69.2%), Chinese (63.2%), Kurdish (70.7%), and Common Turkic (75.7%), while Nemo also underperforms in Sotho (81.3%) and Arabic (82.0%). In contrast, GPT, LLaMA, and Qwen maintain near-perfect validity across all clusters, demonstrating greater robustness. Notably, Gemma's shortcomings persist despite its larger parameter size (12B), sug-

gesting that factors such as training data quality or decoding strategies may play a more critical role than model scale in generating reliably structured outputs.

Model-Predicted Toxicity Shifts We investigated how model-predicted toxicity labels in Standard English change when mapped to the standard and dialectal varieties of other language clusters. Starting from English predictions, we specifically focused on sentences labeled as toxic (scores 4 or 5) and non-toxic (scores 1 or 2). For toxic English sentences, we measured the percentage of cases where predicted toxicity was reduced when translated into other languages. Conversely, for non-toxic English sentences, we computed how often toxicity increased in the translated outputs. These comparisons were made separately for standard varieties and dialectal forms across all models. The results highlight clear toxicity shifts, especially in low-resource and dialectally diverse settings, reinforcing the need to account for language variety in multilingual moderation. Details of the outcomes are reported in Fig. 3. Across the board, all models tend to give lower toxicity scores when English toxic sentences are transformed into other language varieties (Fig. 3a). This drop is fairly consistent, with toxicity reduced by about 50% on average, regardless of the language or model. The effect is especially strong for Sotho and Kurdish, where all models show a notably large reduction—in many cases, cutting toxicity scores by more than 70–80% compared to the original English.

The pattern is quite different when we look at non-toxic English sentences and how they're scored after translation. GPT stands out: it consistently assigns low toxicity scores to these benign sentences, no matter the target variety—usually staying below 10%. However, the other models are far more variable. In particular, LLaMA assigns elevated toxicity scores in up to 80% of Sotho cases, which means it might be mistaking benign sentences for toxic ones in the vast majority of those instances. We see similar, though less extreme, trends with LLaMA in languages like Kurdish, Finnish, and Latvian. This suggests that while GPT remains relatively stable in preserving the non-toxic nature of inputs, other models—especially LLaMA—are more prone to over-predicting toxicity, particularly in lower-resource or linguistically complex varieties. See Appendix B, for detailed result reports for all clusters and models.

Metric	Bengali	English
Mean Toxicity	2.46	2.51
Median Toxicity	2.0	2.0
Score 1 (%)	37.0	39.0
Score 2 (%)	19.0	15.0
Score 3 (%)	19.0	16.0
Score 4 (%)	11.0	16.0
Score 5 (%)	14.0	14.0

Table 4: Comparison of toxicity ratings for 100 English and Bengali sentences annotated independently.

Human Ratings of Toxicity Preservation To evaluate how toxicity is preserved during translation from English to Bengali, we designed a controlled annotation process involving two bilingual annotators. The annotators independently rated toxicity for both Bengali and English sentences without evaluating parallel pairs to eliminate potential cross-lingual bias.

The stimuli consist of 100 Bengali sentences, translated from English using machine translation (MT), and 100 original English sentences. These were divided into two subsets for each language: BS1 and BS2 for Bengali, and ES1 and ES2 for English. Annotator A1 rated BS1 and ES2, while annotator A2 rated BS2 and ES1. This assignment ensured that no annotator saw parallel English-Bengali sentence pairs, maintaining independence in ratings across the two languages.

The key objective of this study is to compare the aggregated toxicity scores of Bengali sentences (BS1 + BS2) with English sentences (ES1 + ES2) to determine whether toxicity is preserved, amplified, or reduced in translation. As shown in Table 4, the results indicate strong preservation of toxicity across the two languages. The mean toxicity ratings are nearly identical: 2.46 for Bengali and 2.51 for English, with both having a median score of 2.0. The score distributions are also similar, though there is a slight reduction in extreme toxicity ratings in Bengali (Score 4 at 11% vs. 16% in English), and a marginally lower proportion of non-toxic (Score 1) sentences (37% vs. 39%). These differences are minimal, suggesting that machine-translated Bengali sentences retain a comparable level of perceived toxicity.

Validating Translation Fidelity Given the shifts observed in model-predicted toxicity and the close alignment seen in human ratings, we wanted to ensure that the translations themselves were not introducing major semantic drift. To assess the fidelity

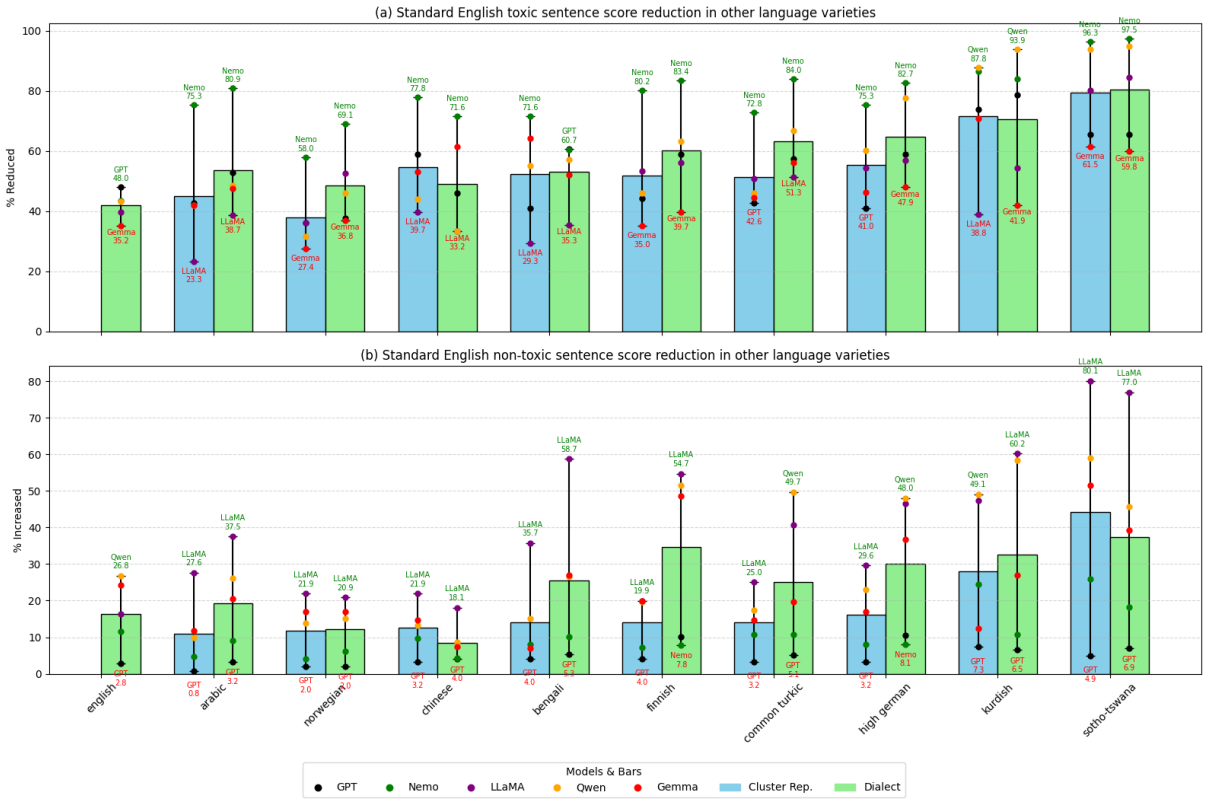

Figure 3: Toxicity shift to other language varieties from Standard English: Each bar shows the percentage change in model toxicity scores when standard English toxic (top) and non-toxic (bottom) sentences are translated into other language varieties. Scores are shown separately for cluster representatives and dialects (average). Dots indicate individual model outputs; error bars span the range across models. We observe that toxicity scores generally decrease for toxic inputs across all varieties, with the strongest reductions in Sotho and Kurdish. In contrast, for non-toxic inputs, GPT remains stable across all varieties, while models like LLaMA tend to over-predict toxicity, especially in Sotho, where benign inputs are rated as toxic in up to 80% of cases.

of these translations, we conducted a reference-free quality evaluation using back-translation. Specifically, we used NLLB to translate from Standard English to each dialectal variety, then performed back-translation from the variety back to English. We then computed BLEU scores between the original and back-translated English sentences to assess semantic preservation.

As reported in Appendix D Table 7, most varieties show reasonably strong BLEU scores, suggesting that the translations retained the original meaning well. However, a few clusters—particularly those with lower BLEU—indicated potential loss or distortion in meaning. For those cases, we applied an additional translation refinement step using GPT: the model was prompted with both the original English sentence and the initial machine translation, and asked to improve the target variety output. We then repeated the back-translation and BLEU evaluation. The third row of Table 7 shows the BLEU scores after this refinement step, with significant improvements observed in low-performing varieties. This approach allowed us to achieve more consistent translation quality across all dialects, reducing the likelihood that toxicity shifts were artifacts of poor translation.

6 Conclusion and Future Work

We propose a holistic LLM robustness evaluation framework for handling toxicity across language varieties. Our findings suggest, a notable gap remains between model predictions and human judgment, emphasizing the need for improvements in alignment. Additionally, LLMs tend to be more sensitive to low-resource dialects, indicating that further advancements are required to enhance their consistency across diverse language varieties. We aim to further expand our dataset by incorporating more utterance-based dialects and introducing new perturbation methods, leveraging LLMs' understanding of dialectal variations.

Limitations

At this point, this study mostly contains synthetic and machine-translated dialectal varieties except for one set of spoken utterances (Bengali-Dhaka). While it would be ideal to conduct this study on authentic data, such data are not easily available and they are expensive to collect. This low percentage of real-world dialectal examples is a limitation we hope to address in the future.

References

Mistral AI and NVIDIA. 2024. Mistral-nemo-instruct-2407: A 12b multilingual instruction-tuned language model. https://mistral.ai/news/mistral-nemo. Accessed: 2025-05-13.

Anjum and Rahul Katarya. 2024. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security*, 23:577–608.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases.

Marta R. Costa-jussa, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Adrian de Wynter, Ishaan Watts, Nektar Ege Altıntoprak, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, Nani Paananen, Vesa-Matti Paananen, Anna Pavlenko, Bruno Pereira Vidal, Luciano Strika, Yueh Tsao, Davide Turcato, Oleksandr Vakhno, Judit Velcsov, Anna Vickers, Stéphanie Visser, Herdyan Widarmanto, Andrey Zaikin, and Si-Qing Chen. 2024. Rtp-lx: Can llms evaluate toxicity in multilingual scenarios?

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, 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, Jack Zhang, 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, 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, Maria Tsimpoukelli, 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, Ning Zhang, 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, Rohan Maheswari, 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, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, 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, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, 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, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, 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, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, 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, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, 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, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, 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, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, 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, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, 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, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, 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, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1.

Leipzig: Max Planck Institute for Evolutionary Anthropology. (Accessed on 2025-05-19).

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. 2020a. Normalization of different swedish dialects spoken in finland.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. 2021. Lemmatization of historical old literary finnish texts in modern orthography.

Mika Hämäläinen, Niko Partanen, Khalid Alnajjar, Jack Rueter, and Thierry Poibeau. 2020b. Automatic dialect adaptation in finnish and its effect on perceived creativity.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2024. Machine translation meta evaluation through translation accuracy challenge sets.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Niko Partanen, Mika Hämäläinen, and Khalid Alnaj-

jar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146, Hong Kong, China. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, Singapore. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

Appendix

A LLM-as-a-Judge evaluation prompt

LLM Toxicity Evaluation: Instructions

Instructions: You will be provided with a series of statements in various dialects. Your task is to evaluate the toxicity of each statement. Toxicity is rated on a scale from S1 to S5 based on the level of harmfulness:

- **S1**: Not harmful (neutral or factual statement)
- **S2**: Slightly offensive (minor slurs or casual insults)
- **S3**: Moderately harmful (disrespectful or demeaning language)
- **S4**: Highly toxic (explicit hate speech or strong language)
- **S5**: Extremely toxic (incites violence, threats, or severe hate speech)

Response Format: Your response should be **only** the label `S1`, `S2`, `S3`, `S4`, or `S5` — no explanation, no extra text.
Input Statement: `{Input_statement}`

Figure 4: Prompt instructions provided to the LLM for toxicity scoring of dialectal input. The model returns a single severity label (S1–S5) per statement.

B Reporting Detailed Toxicity Shift

To better understand how toxicity perceptions shift when translating standard English inputs into various language varieties, we report detailed results in Table 5. The table breaks down model predictions across two axes: (1) the reduction in toxicity scores for originally toxic English sentences, and (2) the increase in toxicity scores for originally non-toxic sentences after translation.

C Validity of Model Outputs Across Language Clusters

Table 6 provides a detailed breakdown of the percentage of valid toxicity predictions across language clusters and models.

D Translation Fidality Evaluation using back-Translation

To assess the semantic fidelity of machine-translated outputs across dialectal varieties, we conduct a reference-free evaluation using back-translation. Specifically, we compute BLEU scores between the original English sentences and their back-translated counterparts. Table 7 reports these scores for each language variety. The first column presents BLEU scores using the baseline NLLB translations. The second column shows results after applying GPT-assisted refinement to improve semantic accuracy. The final column (Δ) highlights the relative improvement achieved through this refinement process.

E Detailed Evaluation Results

This section presents the detailed result tables (Tables 8 to 12) summarizing the performance of each model across different languages and dialects. We report metrics such as F1 scores (for bin=5 classifications) and RMSE-Similarity.

F Binning Methodology

To assign values in the range $[1, 5]$ into a specified number of bins, we divide the range into equal-sized intervals. Let N denote the number of bins. The bin edges are defined as follows:

$$\text{Bin Edges} = \{e_i \mid e_i = 1 + (i-1) \cdot \Delta e,\ i = 1, 2, \ldots, N+1\},$$

Toxic sentences: (Cluster Rep., Dialect) % reduced						
	GPT	Nemo	LLaMA	Qwen	Gemma	Avg
Arabic	(42.6, 52.7)	(75.3, 80.9)	(23.3, 38.7)	(41.8, 48.5)	(41.9, 47.5)	(45.0, 53.7)
Bengali	(41.0, 60.7)	(71.6, 60.5)	(29.3, 35.3)	(55.1, 57.1)	(64.1, 52.1)	(52.2, 53.1)
Chinese	(59.0, 45.9)	(77.8, 71.6)	(39.7, 26.7)	(43.9, 33.2)	(53.0, 61.5)	(54.7, 47.8)
Turkic	(42.6, 57.4)	(72.8, 84.0)	(50.9, 51.3)	(45.9, 66.8)	(44.4, 56.0)	(51.3, 63.1)
English	(0.0, 48.0)	(0.0, 43.5)	(0.0, 39.6)	(0.0, 43.1)	(0.0, 35.2)	(0.0, 41.9)
Finnish	(44.3, 58.9)	(80.2, 83.4)	(53.4, 56.1)	(45.9, 63.1)	(35.0, 39.7)	(51.8, 60.2)
Latvian	(41.0, 59.0)	(75.3, 82.7)	(54.3, 56.9)	(60.2, 77.6)	(46.2, 47.9)	(55.4, 64.8)
Kurdish	(73.8, 78.7)	(86.4, 84.0)	(38.8, 54.3)	(87.8, 93.9)	(70.9, 41.9)	(71.5, 70.6)
Norwegian	(36.1, 37.7)	(58.0, 69.1)	(36.2, 52.6)	(31.6, 45.9)	(27.4, 36.8)	(37.9, 48.4)
Sotho	(65.6, 65.6)	(96.3, 97.5)	(80.2, 84.5)	(93.9, 94.9)	(61.5, 59.8)	(79.5, 80.5)
Avg	(44.6, 56.5)	(69.4, 75.7)	(40.6, 49.6)	(50.6, 62.4)	(44.4, 47.8)	(49.9, 58.4)
Non-toxic sentences: (Cluster Rep., Dialect) % increased						
Arabic	(0.8, 3.2)	(4.6, 9.1)	(27.6, 37.5)	(9.8, 26.2)	(11.7, 20.6)	(10.9, 19.3)
Bengali	(4.0, 5.3)	(8.1, 10.2)	(35.7, 58.7)	(15.0, 26.6)	(7.0, 26.9)	(14.0, 25.5)
Chinese	(3.2, 4.0)	(9.6, 4.3)	(21.9, 18.1)	(13.3, 8.7)	(14.6, 7.3)	(12.5, 8.5)
Turkic	(3.2, 5.1)	(10.7, 10.7)	(25.0, 40.6)	(17.3, 49.7)	(14.6, 19.6)	(14.2, 25.1)
English	(0.0, 2.8)	(0.0, 11.5)	(0.0, 16.3)	(0.0, 26.8)	(0.0, 24.3)	(0.0, 16.3)
Finnish	(4.0, 10.1)	(7.1, 7.8)	(19.9, 54.7)	(19.7, 51.6)	(19.9, 48.6)	(14.1, 34.6)
Latvian	(3.2, 10.5)	(8.1, 8.1)	(29.6, 48.5)	(23.1, 48.0)	(17.0, 36.8)	(16.2, 30.4)
Kurdish	(7.3, 6.5)	(24.4, 10.7)	(47.4, 60.2)	(49.1, 58.4)	(12.3, 26.9)	(28.1, 32.5)
Norwegian	(2.0, 2.0)	(4.1, 6.1)	(21.9, 20.9)	(13.9, 15.0)	(17.0, 17.0)	(11.8, 12.2)
Sotho	(4.9, 6.9)	(25.9, 18.3)	(80.1, 77.0)	(59.0, 45.7)	(51.5, 39.2)	(44.3, 37.4)
Avg	(3.3, 5.6)	(10.3, 9.7)	(30.9, 43.2)	(22.0, 35.7)	(16.6, 26.7)	(16.6, 24.2)

Table 5: Percentage of toxicity shifts after translation from Standard English to various language varieties. The top half shows the reduction in predicted toxicity for originally toxic English sentences, while the bottom half shows the increase in predicted toxicity for originally non-toxic English sentences. Each cell reports the percentage change for the cluster representative and dialectal variety (avg.), respectively. Results are averaged across clusters and models in the rightmost and bottom rows. Higher reduction values (top) indicate potential underprediction of toxicity post-translation, while higher increase values (bottom) suggest overprediction of toxicity in benign inputs.

	GPT	Nemo	LLaMA	Qwen	Gemma	Avg
Arabic	100.0	82.0	100.0	100.0	83.5	93.1
Chinese	100.0	93.6	99.9	100.0	63.2	91.4
Finnish	100.0	92.0	100.0	100.0	85.4	95.5
Kurdish	100.0	88.2	99.9	100.0	70.7	91.7
Norwegian	100.0	97.5	100.0	100.0	91.3	97.8
Latvian	100.0	95.9	100.0	100.0	89.9	97.2
English	100.0	85.4	100.0	100.0	86.2	94.3
Sotho	100.0	81.3	100.0	100.0	86.0	93.5
Bengali	100.0	94.5	100.0	100.0	69.2	92.7
Turkic	100.0	87.3	100.0	100.0	75.7	92.6
Average (Macro)	100.0	89.8	100.0	100.0	80.1	94.0

Table 6: Percentage of valid toxicity predictions across language clusters and LLMs. Each cell represents the proportion of examples for which the model produced a valid, structured output in the given cluster. While GPT, Qwen, and LLaMA consistently achieve near-perfect validity across all clusters, models like Nemo and Gemma show greater variability, especially in low-resource or dialectally diverse languages such as Bengali, Chinese, and Kurdish. The macro average in the bottom row summarizes each model's validity performance across all clusters.

Language Cluster	Language Variety	NLLB	NLLB+GPT	+Δ (%)
Arabic	North Mesopotamian Arabic	44.41	41.04	
	Ta'izzi-Adeni Arabic	46.89	41.97	
	Tunisian Arabic	32.31	35.84	10.9
	South Levantine Arabic	44.75	41.93	
	Levantine Arabic (A:North)	45.41	43.43	
	Standard Arabic	46.97	44.67	
	Najdi Arabic	46.14	39.73	
	Moroccan Arabic	39.63	38.86	
	Egyptian Arabic	47.85	40.82	
Bengali	Standard	43.30	41.85	
Chinese	Cantonese	24.05	28.20	17.3
	Classical-Middle-Modern Sinitic (Simplified)	33.55	36.61	9.1
	Classical-Middle-Modern Sinitic (Traditional)	20.78	32.53	56.5
Common turkic	Central Oghuz	41.95	41.88	
	South Azerbaijani	32.96	31.81	
	North Azerbaijani	41.84	40.25	
Latvian	Latgalian	37.56	40.04	6.6
	Standard Latvian	42.70	42.36	
Kurdish	Central Kurdish	41.34	38.46	
	Northern Kurdish	42.36	38.93	
Norwegian	Norwegian Nynorsk	39.81	46.35	16.4
	Norwegian Bokmal	58.58	53.45	
Sotho	Northern Sotho	41.91	40.92	
	Southern Sotho	44.41	43.34	
Average (Micro)		40.98	40.27	

Table 7: BLEU scores from back-translation evaluating translation fidelity for each language variety. The first column shows scores from NLLB translations, while the second column shows results after GPT-assisted refinement. The third column reports the percentage improvement (Δ) when refinement is applied. Notable improvements are observed in all Chinese varieties and a few other language varieties signifying the effectiveness of GPT in enhancing translation quality. The global row shows the average across all varieties.

Language Cluster	Variety	F1	RMSE-SIM
English	Standard	22.10	66.10
	Southeast american enclave	23.30	64.50
	Chicano	23.40	65.50
	Nigerian	22.40	65.60
	African american vernacular	22.30	65.20
	Appalachian	23.90	65.90
	Australian	22.20	64.30
	Colloquial singapore	20.00	63.30
	Hong kong	19.40	63.00
	Indian	20.00	64.50
	Irish	20.60	65.40
Norwegian	Norwegian nynorsk	20.00	59.10
	Norwegian bokmal	18.00	60.90
Bengali	Dhaka	17.00	54.80
	Standard	18.00	59.60
Arabic	North mesopotamian arabic	17.80	57.40
	Ta'izzi-adeni arabic	16.20	58.80
	Tunisian arabic	18.60	57.50
	South levantine arabic	18.00	59.30
	Levantine arabic (a:north)	18.20	59.50
	Standard arabic	18.50	58.40
	Najdi arabic	16.90	59.00
	Moroccan arabic	15.90	56.40
	Egyptian arabic	17.70	57.50
Chinese	Cantonese	16.60	58.40
	Classical-middle-modern sinitic (simplified)	18.70	59.50
	Classical-middle-modern sinitic (traditional)	18.30	59.00
Turkic	Central oghuz	17.80	59.10
	South azerbaijani	14.70	54.00
	North azerbaijani	16.90	58.00
Latvian	East latvian	16.90	56.50
	Latvian	16.90	58.70
Finnish	Finnish	16.90	58.10
	Pohjois-satakunta	17.80	57.70
	Keski-karjala	16.90	56.50
	Kainuu	16.40	55.60
	Etela-pohjanmaa	18.60	57.80
	Etela-satakunta	17.80	57.40
	Pohjois-savo	20.10	56.10
	Pohjois-karjala	16.40	55.30
	Keski-pohjanmaa	18.60	56.90
	Kaakkois-hame	18.00	58.00
	Pohjoinenkeski-suomi	15.00	56.20
	Pohjois-pohjanmaa	18.50	57.10
	Pohjoinenvarsinais-suomi	17.40	57.10
	Etela-karjala	19.70	57.20
	Lansi-uusimaa	17.40	57.80
	Inkerinsuomalaismurteet	19.20	58.00
	Lantinenkeski-suomi	18.70	56.90
	Lansi-satakunta	16.90	56.40
	Etela-savo	16.20	55.60
	Lansipohja	15.40	57.60
	Pohjois-hame	18.50	56.70
	Etelainenkeski-suomi	16.60	57.90
	Etela-hame	19.90	57.40
	Perapohjola	17.10	57.10
Sotho	Northern sotho	14.20	53.00
	Southern sotho	15.60	56.00
Kurdish	Central kurdish	15.70	51.60
	Northern kurdish	12.50	49.10
Average (Micro)		18.20	58.50

Table 8: Evaluation Results for gpt-4.1-2025-04-14

Language Cluster	Variety	F1	RMSE-SIM
English	Standard	31.40	70.00
	Southeast american enclave	31.40	70.40
	Chicano	32.70	70.90
	Nigerian	33.40	70.10
	African american vernacular	33.50	69.80
	Appalachian	34.20	70.80
	Australian	34.00	69.80
	Colloquial singapore	33.70	69.60
	Hong kong	31.70	69.90
	Indian	30.30	69.30
	Irish	32.90	71.10
Arabic	North mesopotamian arabic	28.10	62.60
	Ta'izzi-adeni arabic	24.50	61.10
	Tunisian arabic	26.90	62.60
	South levantine arabic	27.80	61.50
	Levantine arabic (a:north)	26.60	63.80
	Standard arabic	25.10	60.20
	Najdi arabic	26.50	61.80
	Moroccan arabic	29.50	62.60
	Egyptian arabic	28.50	63.00
Turkic	Central oghuz	25.10	61.60
	South azerbaijani	24.90	61.90
	North azerbaijani	26.50	59.50
Chinese	Cantonese	29.20	61.00
	Classical-middle-modern sinitic (simplified)	21.60	59.10
	Classical-middle-modern sinitic (traditional)	23.70	59.90
Kurdish	Central kurdish	21.90	60.60
	Northern kurdish	24.00	57.20
Bengali	Dhaka	24.00	58.80
	Standard	25.10	60.20
Norwegian	Norwegian nynorsk	23.70	58.40
	Norwegian bokmal	23.80	59.90
Sotho	Northern sotho	20.50	59.10
	Southern sotho	20.50	59.30
Finnish	Finnish	21.90	56.00
	Pohjois-satakunta	24.20	57.50
	Keski-karjala	20.30	56.40
	Kainuu	19.20	57.80
	Etela-pohjanmaa	23.00	57.10
	Etela-satakunta	20.20	58.20
	Pohjois-savo	21.50	58.00
	Pohjois-karjala	18.80	56.90
	Keski-pohjanmaa	22.00	58.60
	Kaakkois-hame	23.50	59.10
	Pohjoinenkeski-suomi	21.00	56.70
	Pohjois-pohjanmaa	22.30	58.30
	Pohjoinenvarsinais-suomi	20.30	58.50
	Etela-karjala	23.00	58.00
	Lansi-uusimaa	19.40	57.30
	Inkerinsuomalaismurteet	22.20	56.90
	Lantinenkeski-suomi	22.60	59.00
	Lansi-satakunta	19.10	57.20
	Etela-savo	21.00	57.20
	Lansipohja	23.10	58.00
	Pohjois-hame	23.60	58.20
	Etelainenkeski-suomi	22.40	58.60
	Etela-hame	20.20	58.60
	Perapohjola	22.10	57.10
Latvian	East latvian	22.20	57.00
	Latvian	22.80	57.80
Average (Micro)		25.00	61.10

Table 9: Evaluation Results for Mistral-Nemo-Instruct-2407

Language Cluster	Variety	F1	RMSE-SIM
English	Standard	25.90	65.30
	Southeast american enclave	22.60	64.90
	Chicano	26.30	65.30
	Nigerian	22.90	65.10
	African american vernacular	24.60	65.90
	Appalachian	25.90	64.70
	Australian	23.50	63.60
	Colloquial singapore	19.30	63.00
	Hong kong	23.00	64.30
	Indian	23.90	65.80
	Irish	24.10	65.00
Arabic	North mesopotamian arabic	24.10	63.60
	Ta'izzi-adeni arabic	25.20	64.20
	Tunisian arabic	19.40	63.90
	South levantine arabic	20.80	64.10
	Levantine arabic (a:north)	21.60	64.70
	Standard arabic	22.20	64.00
	Najdi arabic	23.70	62.10
	Moroccan arabic	19.20	62.60
	Egyptian arabic	24.00	64.30
Turkic	Central oghuz	23.60	63.90
	South azerbaijani	18.10	62.70
	North azerbaijani	18.90	63.40
Kurdish	Central kurdish	19.50	62.80
	Northern kurdish	17.90	63.30
Bengali	Dhaka	16.10	62.10
	Standard	21.50	62.80
Chinese	Cantonese	23.70	62.80
	Classical-middle-modern sinitic (simplified)	19.30	61.90
	Classical-middle-modern sinitic (traditional)	19.10	60.80
Finnish	Finnish	23.30	58.90
	Pohjois-satakunta	17.70	60.70
	Keski-karjala	17.40	61.50
	Kainuu	16.90	62.00
	Etela-pohjanmaa	19.70	61.10
	Etela-satakunta	16.90	61.70
	Pohjois-savo	18.70	62.20
	Pohjois-karjala	17.40	61.40
	Keski-pohjanmaa	18.80	60.40
	Kaakkois-hame	19.60	60.90
	Pohjoinenkeski-suomi	14.70	61.80
	Pohjois-pohjanmaa	17.20	61.70
	Pohjoinenvarsinais-suomi	18.50	61.20
	Etela-karjala	15.70	60.60
	Lansi-uusimaa	18.50	62.10
	Inkerinsuomalaismurteet	17.00	61.50
	Lantinenkeski-suomi	20.20	62.30
	Lansi-satakunta	16.20	61.30
	Etela-savo	20.00	61.70
	Lansipohja	20.20	62.30
	Pohjois-hame	18.60	61.00
	Etelainenkeski-suomi	16.60	61.40
	Etela-hame	15.40	61.90
	Perapohjola	19.90	62.70
Sotho	Northern sotho	13.30	62.50
	Southern sotho	13.50	61.60
Latvian	East latvian	19.00	60.80
	Latvian	21.20	62.00
Norwegian	Norwegian nynorsk	17.80	59.80
	Norwegian bokmal	18.10	60.40
Average (Micro)		20.00	62.60

Table 10: Evaluation Results for Llama-3.1-8B

Language Cluster	Variety	F1	RMSE-SIM
English	Standard	30.60	71.90
	Southeast american enclave	26.90	69.50
	Chicano	33.60	71.40
	Nigerian	29.70	69.80
	African american vernacular	27.20	68.70
	Appalachian	30.00	70.20
	Australian	28.70	70.50
	Colloquial singapore	32.20	69.70
	Hong kong	28.80	68.40
	Indian	26.20	69.30
	Irish	31.00	71.00
Arabic	North mesopotamian arabic	25.10	64.70
	Ta'izzi-adeni arabic	23.80	64.00
	Tunisian arabic	23.90	65.00
	South levantine arabic	23.70	63.70
	Levantine arabic (a:north)	26.10	64.40
	Standard arabic	22.70	64.10
	Najdi arabic	25.50	64.10
	Moroccan arabic	22.60	65.50
	Egyptian arabic	26.80	64.00
Chinese	Cantonese	27.90	65.50
	Classical-middle-modern sinitic (simplified)	23.70	64.60
	Classical-middle-modern sinitic (traditional)	22.20	62.90
Norwegian	Norwegian nynorsk	23.30	62.20
	Norwegian bokmal	26.50	64.30
Turkic	Central oghuz	21.40	63.00
	South azerbaijani	14.50	61.80
	North azerbaijani	20.20	61.80
Finnish	Finnish	21.30	60.80
	Pohjois-satakunta	16.70	60.20
	Keski-karjala	16.70	59.30
	Kainuu	18.40	61.20
	Etela-pohjanmaa	14.70	60.90
	Etela-satakunta	15.20	60.50
	Pohjois-savo	16.50	59.60
	Pohjois-karjala	16.70	60.20
	Keski-pohjanmaa	16.80	59.70
	Kaakkois-hame	18.60	62.00
	Pohjoinenkeski-suomi	16.10	60.20
	Pohjois-pohjanmaa	16.00	59.60
	Pohjoinenvarsinais-suomi	17.10	59.80
	Etela-karjala	18.30	60.00
	Lansi-uusimaa	17.10	61.40
	Inkerinsuomalaismurteet	18.10	61.50
	Lantinenkeski-suomi	17.00	59.70
	Lansi-satakunta	15.60	59.40
	Etela-savo	17.50	60.20
	Lansipohja	19.70	61.30
	Pohjois-hame	18.80	60.60
	Etelainenkeski-suomi	14.90	59.80
	Etela-hame	17.70	62.80
	Perapohjola	17.60	60.40
Latvian	East latvian	16.60	59.60
	Latvian	21.10	62.20
Bengali	Dhaka	22.90	61.20
	Standard	20.20	59.90
Kurdish	Central kurdish	14.10	56.40
	Northern kurdish	14.20	59.10
Sotho	Northern sotho	11.90	58.60
	Southern sotho	11.30	56.80
Average (Micro)		21.20	63.00

Table 11: Evaluation Results for Qwen2.5-7B-Instruct

Language Cluster	Variety	F1	RMSE-SIM
English	Standard	35.00	72.90
	Southeast american enclave	35.00	71.80
	Chicano	36.00	72.50
	Nigerian	36.40	70.60
	African american vernacular	35.70	71.90
	Appalachian	36.40	73.30
	Australian	37.80	72.50
	Colloquial singapore	35.70	70.10
	Hong kong	36.30	70.10
	Indian	36.20	71.20
	Irish	35.10	71.50
Arabic	North mesopotamian arabic	27.70	67.50
	Ta'izzi-adeni arabic	28.60	67.90
	Tunisian arabic	26.10	66.40
	South levantine arabic	27.40	68.60
	Levantine arabic (a:north)	31.10	71.20
	Standard arabic	25.20	67.20
	Najdi arabic	28.40	67.90
	Moroccan arabic	26.10	68.10
	Egyptian arabic	28.50	67.30
Norwegian	Norwegian nynorsk	26.80	66.80
	Norwegian bokmal	29.60	69.20
Turkic	Central oghuz	31.10	66.40
	South azerbaijani	24.90	64.80
	North azerbaijani	30.50	66.90
Finnish	Finnish	24.30	66.70
	Pohjois-satakunta	27.50	62.60
	Keski-karjala	29.80	62.40
	Kainuu	24.20	60.20
	Etela-pohjanmaa	27.90	60.50
	Etela-satakunta	28.70	64.10
	Pohjois-savo	28.30	61.30
	Pohjois-karjala	25.50	59.70
	Keski-pohjanmaa	25.90	63.30
	Kaakkois-hame	28.80	65.20
	Pohjoinenkeski-suomi	25.70	59.80
	Pohjois-pohjanmaa	28.00	62.80
	Pohjoinenvarsinais-suomi	26.20	62.30
	Etela-karjala	27.00	63.70
	Lansi-uusimaa	28.40	64.70
	Inkerinsuomalaismurteet	26.30	63.10
	Lantinenkeski-suomi	27.40	64.10
	Lansi-satakunta	25.80	62.30
	Etela-savo	28.10	60.40
	Lansipohja	30.40	62.60
	Pohjois-hame	27.20	63.50
	Etelainenkeski-suomi	27.00	61.60
	Etela-hame	28.40	63.40
	Perapohjola	26.80	63.70
Chinese	Cantonese	27.90	66.60
	Classical-middle-modern sinitic (simplified)	28.10	65.90
	Classical-middle-modern sinitic (traditional)	26.80	64.10
Latvian	East latvian	29.30	66.10
	Latvian	29.00	65.50
Bengali	Dhaka	26.80	64.60
	Standard	25.20	65.60
Sotho	Northern sotho	17.90	63.60
	Southern sotho	21.50	63.50
Kurdish	Central kurdish	24.50	60.90
	Northern kurdish	27.00	62.40
Average (Micro)		28.80	65.80

Table 12: Evaluation Results for gemma-3-12b-it

where Δe is the width of each bin, given by:

$$\Delta e = \frac{5-1}{N}.$$

For a given value $v \in [1, 5]$, the bin assignment is determined as follows:

$$\text{Bin}(v) = \begin{cases} 1, & \text{if } v = e_1, \\ i, & \text{if } e_{i-1} < v \leq e_i,\ i = 2, 3, \ldots, N, \\ N, & \text{if } v = e_{N+1}. \end{cases}$$

This approach ensures that:

- The first bin includes the value 1.
- Each subsequent bin includes values strictly greater than the lower edge and up to the upper edge, except for the last bin, which includes its upper edge 5.

Example: For $N = 5$, the bin edges are:

$$\{1.0, 2.0, 3.0, 4.0, 5.0\}.$$

A value $v = 1.666$ would fall into Bin 2 as $1.0 < v \leq 2.0$, and $v = 5.0$ would fall into Bin 5.

G Evaluation Metrics

In this section, we evaluate the performance of the toxicity prediction model using several metrics that consider the ordinal nature of the labels, which range from 1 to 5 (with 1 representing the lowest toxicity and 5 representing the highest toxicity). The following metrics were used: F1-score and Root Mean Square Error (RMSE)-based Similarity. Example scores are presented, along with the ranges of each metric, and their meanings in the context of our setup.

G.1 F1-Score

The F1-score is the harmonic mean of precision and recall, calculated as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where precision is the ratio of true positives to predicted positives, and recall is the ratio of true positives to actual positives.

Example Score: The F1-score obtained by the model is 0.2260 (22.60%), reflecting the model's difficulties in both identifying true positives and reducing false positives.

Range:

- **Original Range:** $[0, 1]$
- **Interpretation:** A higher F1-score indicates a better balance between precision and recall. In our case, the low score suggests poor performance in both aspects, implying a need for improvement in the model's classification ability.

G.2 Root Mean Square Error (RMSE) and RMSE-Based Similarity

Root Mean Square Error (RMSE) measures the average magnitude of prediction errors, considering the squared differences between true and predicted values. RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where y_i represents the ground truth, $\hat{y}_i$ represents the predicted value, and N is the total number of instances.

To convert RMSE into a similarity measure, we normalize the RMSE by dividing by the maximum possible error (4, given that the labels range from 1 to 5), and then subtract it from 1:

$$\text{RMSE}_{\text{normalized}} = \frac{\text{RMSE}}{4},$$

$$\text{Similarity}_{\text{RMSE}} = 1 - \text{RMSE}_{\text{normalized}}$$

Example Score: The model achieved an RMSE of 1.9976, which, when normalized, gives 0.4994. This translates to an RMSE-based similarity score of **0.5006**. This suggests moderate similarity between the predicted and actual values.

Range:

- **Original RMSE Range:** $[0, 4]$
- **Similarity Range:** $[0, 1]$
- **Interpretation:** A lower RMSE value indicates that the predictions are closer to the true values, while a higher RMSE-based similarity indicates better performance. In our case, an RMSE-based similarity of 0.5006 means that the model is achieving moderate similarity, indicating that the predictions are roughly halfway between a perfect match and the maximum possible error.

G.3 Summary and Interpretation of Scores

The metrics collectively indicate several areas where the model struggles:

- **Low accuracy and F1-score** indicate poor performance in exact classification of toxicity levels.
- **RMSE-based and MAE-based Similarity** suggest moderate similarity, implying that the model has considerable room for improvement in predicting values that closely resemble true scores.

To improve the model's performance, it is important to focus on better feature extraction, calibration, and optimization techniques to ensure the model can accurately reflect both the ordinal severity of toxicity and align closely with human evaluations.

H Language Variety Table

The language variety table, reported in Table 13, details the specific language clusters and dialects included in our dataset. It provides an overview of the 10 language clusters and 60 varieties used in the evaluation process, along with the number of examples for each variety.

We define a *language cluster* as a group consisting of a primary language and its associated dialects. Each cluster is named after its most proximal ancestral language, with the cluster representative typically chosen as the standard form or the highest-resourced variety. The remaining dialects within the cluster are referred to as the *varieties* of the *cluster representative*. For consistency and clarity, we follow the Glottocode naming convention (Hammarström et al., 2024) to label the varieties, ensuring that each dialect is systematically identified.

Language Cluster	Variety Name	Glottocode	Example Count
Arabic	North Mesopotamian Arabic	nort3142	940
	Ta'Izzi-Adeni Arabic	taiz1242	940
	Tunisian Arabic	tuni1259	940
	South Levantine Arabic	sout3123	940
	Levantine Arabic (A:North)	nort3139	940
	Standard Arabic	stan1318	940
	Najdi Arabic	najd1235	940
	Moroccan Arabic	moro1292	940
	Egyptian Arabic	egyp1253	940
Bengali	Dhaka	dhak1240	380
	Standard	beng1280	940
Chinese	Cantonese	cant1236	940
	Classical-Middle-Modern Sinitic (O:Simplified)	clas1255	940
	Classical-Middle-Modern Sinitic (O:Traditional)	clas1255	940
Finnish	Standard	finn1318	940
	Pohjois-Satakunta	-	940
	Keski-Karjala	-	940
	Kainuu	-	940
	Etelä-Pohjanmaa	-	940
	Etelä-Satakunta	-	940
	Pohjois-Savo	savo1254	940
	Pohjois-Karjala	-	940
	Keski-Pohjanmaa	-	940
	Kaakkois-Häme	-	940
	Pohjoinen Keski-Suomi	-	940
	Pohjois-Pohjanmaa	-	940
	Pohjoinen Varsinais-Suomi	-	940
	Etelä-Karjala	-	940
	Länsi-Uusimaa	-	940
	Inkerinsuomalaismurteet	-	940
	Läntinen Keski-Suomi	-	940
	Länsi-Satakunta	-	940
	Etelä-Savo	-	940
	Länsipohja	-	940
	Pohjois-Häme	-	940
	Eteläinen Keski-Suomi	-	940
	Etelä-Häme	-	940
	Peräpohjola	-	940
Kurdish	Central Kurdish	cent1972	940
	Northern Kurdish	nort2641	940
Norwegian	Norwegian Nynorsk (M:Written)	norw1262	940
	Norwegian Bokmal (M:Written)	norw1259	940
Latvian	East Latvian	east2282	940
	Latvian	latv1249	940
English	Standard	stan1293	940
	Southeast American Enclave	sout3300	799
	Chicano	chic1275	799
	Nigerian	nige1260	799
	African American Vernacular	afri1276	799
	Appalachian	appa1236	799
	Australian	aust1314	799
	Colloquial Singapore	sing1272	799
	Hong Kong	hong1245	799
	Indian	indi1255	799
	Irish	iris1254	799
Sotho	Northern Sotho	nort3233	940
	Southern Sotho	sout2807	940
Turkic	Central Oghuz	azer1255	940
	South Azerbaijani	sout2697	940
	North Azerbaijani	nort2697	940

Table 13: Language cluster and variety names with glottocode and example count. The cluster representative that we utilize as the standard variety is underlined in each cluster.