PerQ: Efficient Evaluation of Multilingual Text Platform-Adaptation Quality

Dominik Macko and Andrew Pulver
Kempelen Institute of Intelligent Technologies
dominik.macko@kinit.sk, andrew.pulver@intern.kinit.sk

Abstract

Since no metrics are available to evaluate specific aspects of a text, such as its personalization quality or quality of adaptation to a specific form, researchers often rely solely on large language models to evaluate such texts. Due to internal biases of individual language models, it is recommended to use multiple of them for combined evaluation, which directly increases costs of such evaluation. In this paper, a computationally efficient method (called PerQ) for evaluation of personalization quality (in terms of platform adaptation) of a given text (generated by a language model) is introduced, which is based on predicting the LLM judgments (i.e., a kind of metric distillation). A case study of comparison of generation capabilities of large and small language models shows the usability of the proposed metric in research, effectively reducing the waste of resources.

1 Introduction

Due to a lack of existing metrics for various aspects of qualitative analysis of the texts and the time-intensive nature of human evaluation (especially challenging in multilingual settings), researchers often resort to using general-purpose large language models (LLMs) for annotation purposes. Given the scope and variety of applied research tasks, such usage of highly resource-intensive LLMs represents a huge waste of resources globally. Therefore, resource-efficient methods must be developed to evaluate specific aspects of the texts.

In this work, we focus on quality evaluation of platform adaptation, i.e., adjusting a given text to suit the stylistic and communicative conventions of a specific distribution platform (e.g., social media channels). While this can be viewed as a specific form of personalization, we distinguish it from user-level personalization, which typically involves conditioning on individual user preferences, profiles, or interaction history. Instead, our setting assumes a fixed target (platform) and focuses on adapting content accordingly (i.e., a higher quality refers to better adaptation). Since there is no standard metric to evaluate this aspect, we propose a new metric (PerQ) that is able to reflect the majority judgment of three diverse LLMs (various sizes and architectures) in an efficient manner (our fine-tuned quality predictors are orders of magnitude faster and much less resource demanding)[1].

The contributions of this work are as follows:

- the **proposal of a new metric** for resource-efficient estimation of text platform-adaptation quality,
- a **methodology for metric training** with a potential to be easily extensible to other textual aspects (depending on LLM capabilities),
- a **comparison and evaluation of multiple base models** offering different trade-offs between performance and inference costs,
- a **case study** showcasing the utility of the proposed metric in research.

2 Related Work

There is no standard metric focused directly on platform-adaptation quality. Traditional metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or BERTScore (Zhang et al., 2020) require reference texts and measure similarity towards those; thus, limiting their utilization to evaluate purely generated content. Such metrics have been used to evaluate personalization (based on user history profile), for example, in LaMP (Salemi et al., 2024), identifying the need for new metrics. The AuPEL method (Wang et al., 2023) uses LLMs as the evaluators of personalized text generation and shows that it is superior to the traditional text similarity metrics. The ExPerT framework (Salemi et al., 2025) also utilizes LLMs to extract atomic aspects from the generated and

[1] https://github.com/kinit-sk/PerQ

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5990–6003
November 6-10, 2026 ©2026 Association for Computational Linguistics

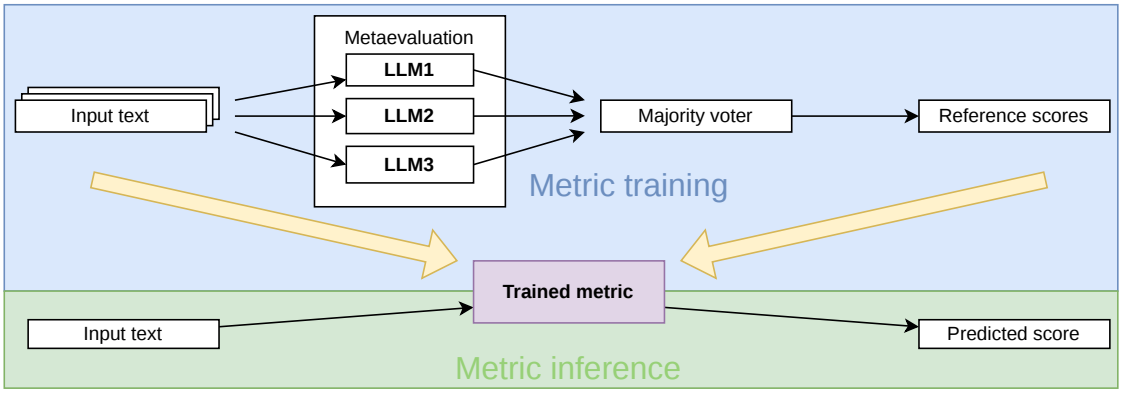


Figure 1: Framework of the proposed metric training and inference.

reference texts and compare them based on content and writing style. The PREF framework (Fu et al., 2025) does not rely on reference texts to evaluate personalized text quality. The personalization is rather focused on a user preference, i.e. user-specific alignment. In PerDisNews study (Zugecova et al., 2025), the target-group personalization quality has been evaluated using three LLMs in zero-shot manner (without reference texts) in order to limit LLMs internal biases. Such approach has shown strong correlation with human judgment.

The LLMs are nowadays commonly used for evaluation of various aspects of a text, such as linguistic quality, coherence, hallucinations, even in multilingual manner (Chang et al., 2025; Hada et al., 2024; Macko et al., 2025). There are also specialized LLM judges trained to evaluate the text; among those, M-Prometheus (Pombal et al., 2025) being one of the first to operate multilingually and offering also reference-free evaluation. There are also attempts to distill quality metrics (even multiple at once) generated by an LLM in a zero-shot manner by using smaller trained encoder models (Alam et al., 2026), which however propagate internal biases of the source LLM into the evaluator.

However, the reliance on huge LLMs (even multiple inferences for a single text) to evaluate personalization quality (or other aspects of a text) is infeasible in many cases, especially for exploratory research, where quick feedback is required to continue. These challenges cannot be solved just by modern LLM acceleration (though invaluable), utilizing advanced caching mechanisms, low-precision inference and quantization. They do not eliminate the fundamental latency and resource overhead inherent in generative architectures. There is a critical need for efficient, standalone metrics that provide the best of both worlds: the nuanced understanding of LLMs with the computational agility of lighter architectures.

3 The Proposed Evaluation Metric

To address these challenges, we propose a reference-free metric to evaluate platform-adaptation quality, based on a smaller fine-tuned language model. The efficiency of the fine-tuning process is ensured by utilization of the **QLoRA** (Dettmers et al., 2023) parameter-efficient finetuning (**PEFT**) approach with 4-bit **quantization**. The metric is trained as a classifier (sequence classification task) based on majority-voted scores from three different LLMs, effectively **distilling** their aggregated judgments. The diverse combination of LLMs reduces internal biases of individual LLMs (as suggested by Zugecova et al., 2025), in contrast to approaches that distill a single model (e.g., OmniScore by Alam et al., 2026. The resulting metric is conceptually similar to the reference-free variant of the COMET metric (Rei et al., 2021), used nowadays for evaluation of machine translation. However, unlike COMET, which performs regression, our method formulates the problem as classification, predicting discrete quality scores (classes).

The LLMs are used only once to score a sufficient number (recommended at least 100 samples per class, based on Pecher et al., 2025) of texts. Afterwards, the metric is trained and provides the score predictions with sufficient accuracy at considerably lower costs (e.g., result obtained in seconds in comparison to hours). It can be directly reused for the same scoring scale (e.g., the one provided in Table 1) on different data without further training.

Score	Description
0 - Not adapted at all to the platform	The text shows no adaptation to the stylistic or structural norms of the platform. a) The format is incompatible with the platform (e.g., long-form text posted to a platform that favors brevity like X/Twitter). b) The tone is formal or generic and does not reflect the platform's communication style. c) There is no use of features native to the platform (e.g., hashtags, tags, emojis, memes, short videos, interactive polls). d) The content could be copied and pasted onto any platform with no change in impact.
1 - Low adaptation to the platform	The text fits on the platform but lacks adaptation in tone, format, or engagement style. a) The post length fits within the platform limits but does not leverage platform-specific formatting (e.g., uses no line breaks or bullets on LinkedIn). b) The tone is only slightly adjusted (e.g., slightly more casual for Instagram but still generic). c) Minimal or generic use of platform features (e.g., hashtags that are too broad or not trend-specific). d) Content could be slightly edited for another platform with minimal effort.
2 - Moderately adapted to the platform	The text shows clear effort to fit the platform's tone and structure, though some elements could be more naturally integrated. a) Includes relevant platform features (e.g., a few well-chosen hashtags, a meme format, a call to action) but doesn't use them to their full potential. b) The tone is mostly aligned with the platform but may feel slightly off or inconsistent in parts. c) The format is generally appropriate for the platform, but could be better optimized (e.g., a caption on Instagram that is a bit too long, or a thread on X that doesn't build momentum). d) Engagement prompts are present but not especially compelling or natural for the platform's audience.
3 - Well-adapted to the platform	The text is naturally and effectively tailored to the specific platform's norms, tone, and engagement style. a) Uses platform-specific features effectively (e.g., carousel for LinkedIn or Instagram, threads on X, storytelling formats on TikTok). b) Tone is fully aligned with what the platform's audience expects (e.g., professional yet authentic on LinkedIn, snappy and meme-savvy on X, visual and emotive on Instagram). c) Leverages platform-specific engagement cues (e.g., "comment below," "share with a friend," "stitch this"). d) The message is concise, formatted appropriately, and makes full use of the platform's strengths. e) Likely to generate high engagement because of its alignment with the platform's user behavior and content culture.

Table 1: Scoring schema for evaluation of platform-adaptation quality.

As illustrated in Fig. 1, in the inference stage (i.e., metric usage), only the input text is provided, and the trained metric predicts the score regarding the platform-adaptation quality of the given text. The quality score reflects the trained classes based on scoring scale (e.g., class 0, class 1, class 2). The proposed approach for metric development can be reused for various scoring scales as well as for various text analysis metrics (e.g., general text quality, fluency, coherency). This is useful especially for multilingual evaluation, where usually non-English text metrics are largely not available.

3.1 PerQ

The research question targeted by PerQ metric development is (**RQ1**) *"To what extent can a lightweight trained evaluation metric approximate the judgments of an ensemble of LLM evaluators when scoring platform adaptation quality?"*.

For training of the PerQ metric, a dataset of multilingual texts adapted to various targets is required, containing sufficient amount of various adaptation quality scores. Since no such dataset is available, we have created a new dataset by using multilingual news articles from MassiveSumm (Varab and Schluter, 2021).

We have pseudo-randomly selected 100 samples per each of the 7 selected languages (English, German, French, Italian, Slovak, Russian, and Hungarian). The titles (headlines) of the articles have been used for platform-adapted text generation and the contents of the articles have been used for adaptation of the already existing text (i.e., it is a modification/adjustment of the provided input text). As targets, we have selected 3 social-media platforms (Twitter/X, Telegram, and Signal). We have used the English instructions in the LLM prompts (templates available in Appendix C). Finally, we have used smaller and larger variants of 3 LLM architectures (Gemma-3, Qwen3, Llama-3) to generate the platform-adapted texts. In this way, we have ended up with sufficient amount and variety of different texts of various quality. Using all combinations of the above mentioned 7 languages, 2 adaptation types, 3 platforms, and 6 LLMs, resulted in 25,200 generated (possibly personalized) texts (see Table 6 for English examples of generated texts, scored with the highest and lowest platform-adaptation quality).

After the texts have been generated, we have used 3 different LLMs (Mistral-Small-3.1-24B-Instruct-2503, Aya-Expanse-32B, and QwQ-32B) to evaluate the quality of the texts by providing a detailed scoring schema (see Table 1). Using majority voting (to reduce the effect of individual LLM biases), the received scores have been modified to a single score per text (the lowest score used when majority was indecisive – i.e, each LLM predicted different score – providing a rather pessimistic view on the output quality since the scale

Score	Absolute distribution	Relative distribution
3	4,135	0.1641
2	11,360	0.4508
1	2,813	0.1116
0	6,874	0.2728
N/A	18	0.0007

Table 2: Distribution of majority evaluation scores.

is ordered). Other aggregation alternatives have been evaluated in the ablation study in Section 5.1.

Out of all generated texts, all three LLM-based evaluators agreed (i.e., a total inter-annotator agreement) on the score in 43% of cases, and all three disagreed (tie-breaking) in 8% of cases. They agreed the most on assignment of the score of 2 (5,207 samples) and the least on assignment of the score of 1 (153 samples). Both the Krippendorff's Alpha and Fleiss' Kappa are of 0.42, while average Spearman's Rho between the LLM-based evaluators is 0.72. The majority score distribution is provided in Table 2. In some cases, the automated parsing of the evaluation scores from the LLMs' outputs has been difficult (representing one more challenge of using LLMs for scoring), providing ambiguous values resulting in no meaningful majority score. Fortunately, the number of such cases that were unresolvable even by a manual check was not high, where this occurred only in under 0.1% of cases.

To answer our research question, we are pseudo-randomly splitting this data into label-balanced (based on majority evaluation score) subsets for training (1,300 samples per class), validation (500 samples per class), and testing (1,000 samples per class). The training and validation splits are used for language model fine-tuning for multiclass classification task, while the testing split is used for evaluation. Since the testing split is balanced based on labels (the same number of samples in the classes), we are using the standard evaluation metrics of Accuracy and macro average of F1 score (MacroF1).

The results in the following section (especially Table 4) show that the metric for evaluation of platform-adaptation quality of the multilingual texts can be successfully trained (answering RQ1), with the accuracy well above chance (worst-case accuracy of 0.642 vs random-classifier accuracy of 0.25). The Spearman correlation coefficient between the predicted scores and majority evaluation scores on test split is above 0.8 for all tested fine-tuned models, i.e. representing **strong correlation**.

The used pseudo-random construction of the splits ensures the diversity across generators, languages, platforms, and adaptation types. Since the same seed articles have been used for data generation in both train and test splits (enabling coverage of all factors in each splits), we have analyzed potential biases due to data leakage. In both splits 98% of seed articles have generated texts with at least two different score values and all the generated texts are different. Moreover, all the generators produced text of all scores. Thus, potential classification shortcuts are minimized. To verify this, we additionally evaluated seed-disjoint splits, where all samples originating from a given source article are assigned exclusively to a single data split (while keeping the numbers of samples per class as mentioned above). This setting provides a stricter evaluation of generalization across unseen content. The evaluation of the best-performing Gemma-2-2B model on test split in this setting resulted in the accuracy of 0.6997 (vs. 0.6973) and Spearman's Rho of 0.85 (vs. 0.85). These comparable results suggest that the original setting does not introduce significant bias. Consequently, we proceed with the original pseudo-random configuration to maintain higher diversity across seed articles.

3.2 Human Sanity Check

Although a full-scale human study was not feasible in our research (multilingual settings, time consuming, questionable replicability, costs), we have conducted a small-scle human sanity check to validate that the evaluators are actually evaluating the quality of adaptation to the platform, reflecting the provided scoring schema. For this purpose, we have pseudo-randomly sampled 10 samples for each majority evaluation score from the above-mentioned test set for English and Slovak languages (for multilingual evaluation). It resulted into 80 samples that have been annotated by a single human based on the same scoring scale (by providing the adjusted text and the target platform description, without knowing evaluation scores). The Spearman correlation coefficient values of each of the used evaluators (along with p-values reflecting the statistical significance) when comparing to human annotation are provided in Table 3. Even by using such a small number of samples for human check, the results indicate **moderate (statistically significant for p-values < 0.05) correlation with a human** of the majority evaluation as well as the best trained PerQ metrics. On the other hand, smaller LLMs

Model	ρ	p-value
majority-else-lowest evaluation	0.4316	0.0001
Mistral-Small-3.1-24B (MistralAI, 2025)	0.4260	0.0001
Gemma-2-2b (PerQ) (Team et al., 2024)	0.4199	0.0001
Qwen3-0.6B-Base (PerQ) (Team, 2025a)	0.4135	0.0001
XLM-R-Large (PerQ) (Conneau et al., 2020)	0.4061	0.0002
Qwen3-1.7B-Base (PerQ) (Team, 2025a)	0.3889	0.0004
mDeBERTa-v3-Base (PerQ) (He et al., 2022)	0.3742	0.0006
DeBERTa-v3-Large (PerQ) (He et al., 2022)	0.3666	0.0008
(A) majority-else-mean evaluation	0.3666	0.0008
(A) rounded mean evaluation	0.3498	0.0015
Qwen3-4B-Base (PerQ) (Team, 2025a)	0.3488	0.0015
M-Prometheus-14B (Pombal et al., 2025)	0.3347	0.0024
XLM-R-Base (PerQ) (Conneau et al., 2020)	0.3241	0.0034
(A) rounded mean prediction	0.2966	0.0076
(A) Gemma-2-2b Mistral-prediction	0.2959	0.0077
Aya-Expanse-32b (Dang et al., 2024)	0.2853	0.0103
(A) majority-else-lowest prediction	0.2804	0.0118
QwQ-32B (Team, 2025b)	0.2700	0.0154
(A) majority-else-mean prediction	0.2607	0.0195
(A) Gemma-2-2b Aya-prediction	0.2386	0.0331
M-Prometheus-3B (Pombal et al., 2025)	0.2190	0.0509
(A) Gemma-2-2b QwQ-prediction	0.2100	0.0615
M-Prometheus-7B (Pombal et al., 2025)	0.1952	0.0828
Gemma-3-4b-it (Team et al., 2025)	0.1930	0.0863
Qwen3-4B (Team, 2025a)	0.1189	0.2937

Table 3: Human sanity check of platform-adaptation evaluation using Spearman correlation (ρ). (A) denotes ablation study.

and smaller M-Prometheus models (< 10B) show weak correlation that is not statistically significant. We further show **strong correlation with human judgments** on other existing annotated data (quality of target-group personalization) in Appendix B.

4 Experimental Results

The following experiment is focused on comparison of multiple language models when trained for the PerQ metric. The selection of base models for fine-tuning covers multiple sizes and architectures to explore the trade-off between various aspects (classification performance vs size/inference costs). The results are summarized in Table 4 and Table 5.

Model	Accuracy	MacroF1
Gemma-2-2B	0.6973	0.6955
Qwen3-4B	0.6963	0.6954
Qwen3-0.6B	0.6888	0.6846
DeBERTa-v3-Large (0.4B)	0.6878	0.6926
Qwen3-1.7B	0.6770	0.6802
mDeBERTa-v3-Base (0.2B)	0.6598	0.6581
XLM-RoBERTa-Large (0.4B)	0.6583	0.6572
XLM-RoBERTa-Base (0.1B)	0.6420	0.6415
random classifier (4-class)	0.25	0.25

Table 4: Comparison of base models for the PerQ metric prediction performance.

	Model (inference mode)	GPU [GB]	Time [s]
a)	Gemma-2-2B (FP16)	7.71	130
	Qwen3-4B (FP16)	10.42	208
	Qwen3-0.6B (FP16)	2.97	60
	DeBERTa-v3-Large (0.4B) (FP16)	2.48	46
	Qwen3-1.7B (FP16)	5.47	105
	mDeBERTa-v3-Base (0.2B) (FP16)	2.00	19
	XLM-RoBERTa-Large (0.4B) (FP16)	2.26	16
	XLM-RoBERTa-Base (0.1B) (FP16)	1.76	7
b)	Mistral-Small-3.1-24B (vLLM)	2×64	~1,242
	Aya-Expanse-32B (vLLM)	2×64	~9,000
	QwQ-32B (vLLM)	2×64	~9,840
	Qwen3-4B (vLLM)	2×64	~3,091
	Gemma-3-4B-IT (vLLM)	2×64	~816
c)	M-Prometheus-3B (full)	37.25	~30,864
	M-Prometheus-3B (4-bit)	4.05	~44,607
	M-Prometheus-7B (4-bit)	7.83	~29,269
	M-Prometheus-14B (4-bit)	14.19	~46,575
	M-Prometheus-3B (vLLM)	2×64	~857

Table 5: Comparison of evaluators inference costs: a) PerQ base models, b) single LLMs, c) specialized LLM-as-a-judges. The vLLM acceleration using parallel execution fully utilized 2 GPUs of 64GB, the models themselves would require less GPU resources (however prolonging the execution time).

There is only small difference in classification performance between the fine-tuned models (up to 5%). However, there are high differences in the required GPU memory (for half-precision inference) and eventual inference time. Table 5a indicates the costs for evaluation by using the test-split data. The costs roughly correspond to the model sizes, especially a higher difference can be seen between the models of >2B parameters and the models of <1B parameters.

When comparing prediction performance and inference costs, the Qwen3-0.6B and DeBERTa-v3-Large models seem to offer the best trade-off. About 0.5B parameters sized models are usually feasible to be run (inferred) on CPU only, thus increasing the PerQ usability even more.

5 Case Study

As a case study to showcase usefulness of the proposed evaluation metric, we select the following research questions: (**RQ2**) *"Are there differences in multilingual text adaptation capabilities (quality of platform adaptation) between small (<=10B) and large (>=20B) LLMs?"*, (**RQ3**) *"Are there differences between adaptation of existing text and platform-specific text generation?"*, (**RQ4**) *"Are there differences in adaptation quality among languages or target platforms?"*.

To answer these research questions, researchers would most probably ended up with the usage of LLMs to evaluate the platform-adaptation quality

(due to subjectivity issues, lack of human annotators across languages, replicability issues, etc.). Therefore, we use the previously-described majority evaluation scores to answer such questions. We further analyze the inference costs to obtain such scores and compare them with the inference costs of the PerQ metric to obtain predictions of the platform-adaptation quality scores. Afterwards, we compare the observations (regarding the stated research questions) from the two approaches, to signify that the same conclusions can be made at significantly lower costs. For the fair comparison, we use only the test split for the analysis. We provide an ablation study in Section 5.1, showing approximately the same distribution when analyzing majority evaluation from the all data (avoiding potential bias due to pseudo-random test-split selection).

In Table 5b, we provide an estimation of LLM inference costs for the evaluation of platform-adaptation quality using the three LLMs. The estimation is calculated based on the vLLM-accelerated[2] parallel execution (using 2 GPUs of 64GB) (Kwon et al., 2023). Based on the average evaluation speed for a single iteration using the whole dataset, mentioned in Section 3.1, we have calculated the execution time for the test split of 4,000 samples. The results show that even the fastest LLM evaluator (Mistral-Small) is almost $6\times$ slower than the slowest fine-tuned PerQ metric (based on Qwen3-4B), while consuming over $12\times$ more GPU memory. When comparing the combined evaluation using all 3 LLMs on 2 GPUs, the time comparison indicates $96\times$ speedup of the PerQ metric (still using $12\times$ less GPU memory). It must be noted that the LLMs often failed in providing the evaluation scores, requiring the repetition of the process and sometimes even manual parsing of the scores from the output, which are not taken into account in this estimation (i.e., actual speedup is even higher). This clearly shows the benefits of the proposed trained metric, accelerating the evaluation of platform-adaptation quality from hours to minutes, significantly saving resources (contributing to the energy sustainability of the research).

In Figure 2, we provide per-generator comparison of adaptation quality as evaluated by majority evaluation (top part of the figure) and by the Gemma-based (as the best performing in Table 4) PerQ metric (bottom part of the figure). The distributions of the scores are quite similar using both evaluation approaches. The results (answering RQ2) indicate that smaller variants of the models produce lower quality of adaptation than their bigger counterparts. The lowest difference is in the case of Gemma-3 models (almost negligible), and the highest in the case of Llama models (the 3B version generated poorly adapted texts).

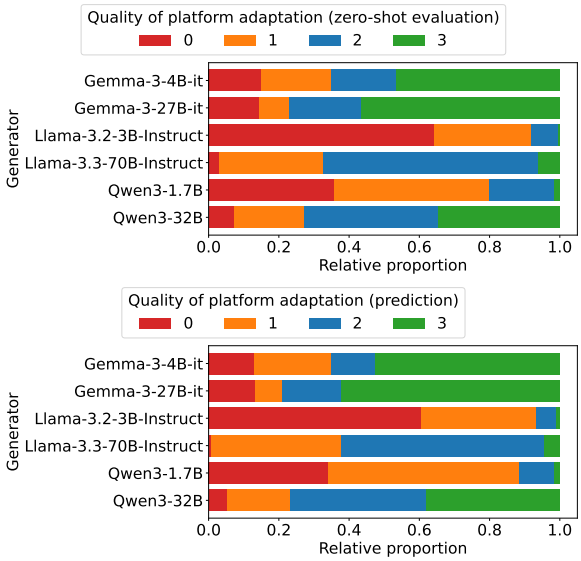


Figure 2: Comparison of LLMs platform-adaptation capabilities based on the majority evaluation scores (top) and the Gemma-based PerQ metric (bottom) for quality of adaptation in the test-split texts.

Similarly, to answer RQ3, we provide a comparison between the texts adapted to a platform by pure generation and the texts adapted by modification of the existing texts in Figure 3. Both approaches of evaluation indicate that the pure generation of the platform-adapted content provides slightly higher quality of adaptation (answering RQ3). The results show that in case of generation, the number of texts gaining score of 3 is doubled in comparison to the modification.

When comparing differences among languages (Figure 4) and among platforms (Figure 5), we can also observe quite similar results from the two evaluation approaches. In both approaches, the Hungarian and Slovak texts are of the lowest platform-adaptation quality. The Slovak language contains the biggest proportion of the 0-score texts (the lowest quality) and the Hungarian language contains the lowest proportion of 3-score texts (the highest quality). The English texts are of the best adaptation quality (out of the tested languages), indicating the predominant proficiency of LLMs in English,

[2] https://docs.vllm.ai/en/latest/

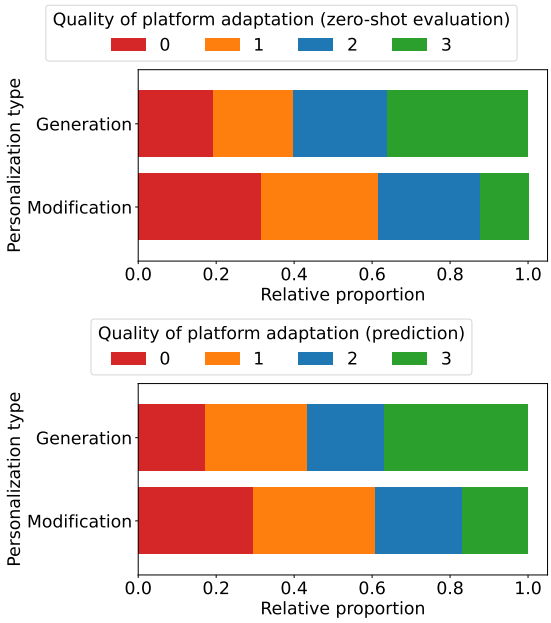


Figure 3: Comparison of platform-adaptation types based on the majority evaluation scores (top) and the Gemma-based PerQ metric (bottom) for quality of adaptation in the test-split texts.

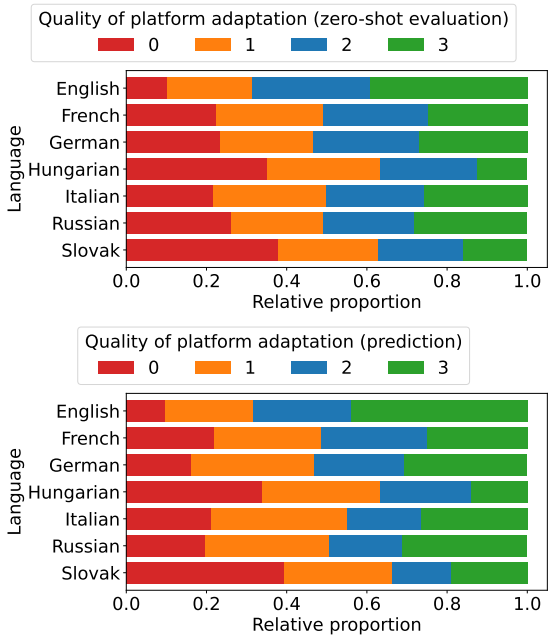


Figure 4: Per-language comparison of platform-adaptation capabilities based on the majority evaluation scores (top) and the Gemma-based PerQ metric (bottom) for quality of adaptation in the test-split texts.

since majority of the pretraining data are English texts. The results indicate that adaptation for Twitter and Telegram platforms was easier for the LLMs than the adaptation for the Signal platform (only 6% of texts scored 3). This effectively answered RQ4.

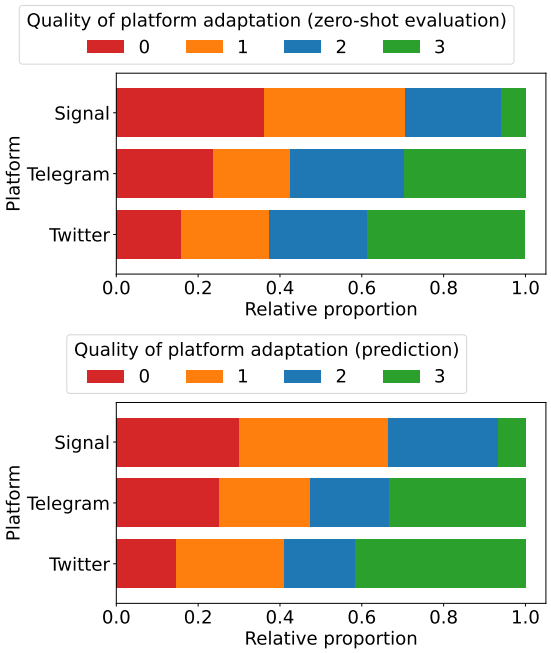


Figure 5: Per-platform (i.e., per-target) comparison of platform-adaptation capabilities based on the majority evaluation scores (top) and the Gemma-based PerQ metric (bottom) for quality of adaptation in the test-split texts.

5.1 Ablation Study

To support our selection of majority voting of three LLMs evaluations and using the lowest value for indecisive voting (called majority-else-lowest strategy), we have analyzed the other alternatives. Namely, we have tested majority-else-mean strategy that uses average of the scores for indecisive majority voting (rounded to the integer value) and rounded mean strategy that averages the scores for each texts (also rounded). In Table 3, we can see that the selected majority-else-lowest strategy achieves the highest correlation with a human.

Although we have used aggregation of three LLMs to minimize internal biases during evaluation, we have tested also the metric model trained based on single-LLM scores to further measure the impact and necessity of the aggregation. Specifically, we have trained Gemma-2-2b (achieving the best performance in Table 4) using the evaluation scores of the three LLMs individually (instead of aggregated majority evaluation). The resulting prediction models denoted as Gemma-2-2b Mistral-prediction, Aya-prediction, and QwQ-prediction in Table 3 indicate decreased correlation to human judgment (the last one not even statistically significant). Also, the aggregated prediction of these

three metrics using all above mentioned aggregation strategies (the best one is called rounded mean prediction) does not have higher correlation to human judgment than the PerQ metric trained using the majority-else-lowest evaluation scores. This directly confirms the benefits of the selected majority voting, with a potential to minimize internal biases of individual LLM evaluators.

Since the analysis of majority-based evaluation of platform-adaptation quality has been done only on test-split portion of the dataset (to be fairly comparable to the trained PerQ metric), the results might be affected by pseudo-random selection of test split. To analyze whether the results might be different for other selections, we provide the same comparisons of various aspects (as those in the previously reported main results) for majority evaluation scores using the whole dataset in Figure 6 – Figure 9. This analysis proves that the results are **not significantly affected** by the selected test-split data.

6 Comparison to Existing Evaluators

Out of the existing approaches, we are comparing our PerQ metric with M-Prometheus (Pombal et al., 2025), a state-of-the-art multilingual trained LLM judge. It is a suite of LLM-based models, available in 3 sizes, namely 3B, 7B, and 14B parameters, all based on the corresponding Qwen2.5 models. M-Prometheus can provide direct assessment of a text as well as pairwise comparison. The direct assessment can be used in reference-based as well as reference-free manner.

We are using reference-free direct assessment mode of M-Prometheus to compare with our PerQ approach. M-Prometheus is limited to 5-point scoring scale (integers from 1 to 5); therefore, we use the lowest score to mark inability of the judge to score the platform-adaptation quality of the text (when "none of the other scores apply" according to the model). To unify with the scoring scale of PerQ, we subtract 2 points from a given score, resulting in the numbers from -1 (unevaluated), via 0 (lowest quality) to 3 (highest platform-adaptation quality).

M-Prometheus is much slower than PerQ (see Table 5), since it uses causal text generation, providing also reasoning behind a given score, while PerQ only classifies the text into the class representing a given score (no need for autoregression, used for generation of next tokens). The vLLM acceleration and batch processing boost the inference a lot, outperforming single LLM-based evaluators (Table 5b); however, due to above-mentioned reasons, it cannot be more efficient than PerQ. Due to fixed 5-point scoring scale, M-Prometheus more often failed in providing estimated platform-adaptation quality, even more often than individual LLM-based evaluators. On the other hand, PerQ always provides a valid prediction among the trained classes.

For further comparison, we have also run the evaluation using single LLM-based evaluators of a comparable size (Qwen3-4B and Gemma-3-4B-IT), which however do not match with the speed of PerQ metric (similar reasons as M-Prometheus), provide worse evaluation performance than bigger LLMs (see Table 3), and more often fail in providing a usable answer (none of the given scores). In illustrations aggregating the score distributions per generators in Appendix D, we can notice a clear bias due to self-preference of individual evaluators (Chen et al., 2025). This confirms the importance of using multiple LLMs to avoid such biases, and thus even higher benefit of our proposed PerQ metric (and metric distillation approach).

Due to the low costs and high speed of PerQ being able to provide a feedback on quality of the generated texts (although not fully accurate), it is especially suitable for experimental pipelines to tune the text-generation settings, analysis of high amounts of data, or to provide an on-the-fly feedback (e.g., for iterative generation) when a high output quality is the priority.

7 Conclusions

The exceptional multilingual capabilities of LLMs provide a way to quickly and easily generate large volumes of high-quality text. However, measuring text quality (across various aspects) often relies on the LLMs themselves (due to a lack of human evaluators in multiple languages and the time-consuming nature of the process). Nevertheless, the usage of energy-intensive LLMs to evaluate large amounts of texts represents a waste of energy and computing resources. In this paper, we propose an efficient way to measuring platform-adaptation quality in a form of the trained metric. Experimental results show a strong correlation between the trained metric predictions and the majority score of three LLMs (limiting their internal biases). The speedup can achieve two orders of magnitude, while requir-

ing much less GPU memory.

Limitations

We have focused on evaluation of platform-adaptation quality as defined on our 4-point scale, along with target-group personalization quality using similar scale, the evaluation of other aspects of the texts or usage of different evaluation scale might be different. The usability of the approach in a completely different domain or quality aspects is highly dependent on the abilities of LLMs themselves to correlate with human judgments. Our key findings rely on an LLM-based evaluation of text quality. Although we have used 3 LLMs to limit the effect of internal biases in evaluation, there can still be some. The study is limited to 7 languages of 4 language-family branches; however, we are unsure about generalization of our findings to other languages. We have limited the generators to 2 variants (smaller and bigger) of 3 SOTA LLMs. The results for other LLMs might differ.

Ethics Statement

The artifacts and results of this study are intended for research purpose only to evaluate capabilities of existing LLMs. The used seed news articles contain potentially sensitive topics, such as public health crises or conflict-related themes, and therefore the generated synthetic data might contain inaccurate and misleading information. While such topics are necessary to ensure diversity and realism of the dataset, they may carry risks if misused or misinterpreted. To mitigate these risks, all data were generated using controlled prompts and standard safety-aligned large language models, without intentional amplification of harmful, misleading, or extremist narratives. The generated content is synthetic and does not correspond to real individuals or events beyond general public knowledge. The existing artifacts used in this work have been properly cited and used according their licenses and intended use. We have also checked and followed licensing and terms of use of the used LLMs. AI assistants have not been used for conducting research in any other way than already described in the paper (text generation and evaluation).

Acknowledgments

This work was partially supported by the European Union NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00068 and partially by *lorAI – Low Resource Artificial Intelligence*, a project funded by Horizon Europe under GA No.101136646.

Computational resources. We acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy.

References

Firoj Alam, Gagan Bhatia, Sahinur Rahman Laskar, and Shammur Absar Chowdhury. 2026. Beyond LLM-as-a-Judge: Deterministic metrics for multilingual generative text evaluation. *Preprint*, arXiv:2604.05083.

Jiayi Chang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. 2025. Exploring the multilingual NLG evaluation abilities of LLM-based evaluators. *Preprint*, arXiv:2503.04360.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672, Suzhou, China. Association for Computational Linguistics.

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Xiao Fu, Hossein A. Rahmani, Bin Wu, Jerome Ramos, Emine Yilmaz, and Aldo Lipani. 2025. PREF: Reference-free evaluation of personalised text generation in LLMs. *Preprint*, arXiv:2508.10028.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. METAL: Towards multilingual meta-evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025. MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 727–752, Vienna, Austria. Association for Computational Linguistics.

MistralAI. 2025. Mistral Small 3.1.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2025. Comparing specialised small and general large language models on text classification: 100 labelled samples to achieve break-Even performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 165–184, Suzhou, China. Association for Computational Linguistics.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and Andre Martins. 2025. M-Prometheus: A suite of open multilingual LLM judges. In *Second Conference on Language Modeling*.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Alireza Salemi, Julian Killingback, and Hamed Zamani. 2025. ExPerT: Effective and explainable evaluation of personalized long-form text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17516–17532, Vienna, Austria. Association for Computational Linguistics.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

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, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qwen Team. 2025b. QwQ-32B: Embracing the power of reinforcement learning.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023. Automated evaluation of personalized text generation using large language models. *Preprint*, arXiv:2310.11593.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopál, Katarína Marcinčinová, and Matúš Mesarčík. 2025. Evaluation of LLM vulnerabilities to being misused for personalized disinformation generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 780–797, Vienna, Austria. Association for Computational Linguistics.

A Computational Resources

For the texts generation, we have used 2× A100 64GB GPU, cumulatively consuming approximately 600 GPU-hours. For evaluation, we have used 2x A100 64GB GPU consuming approximately 2000 GPU-hours. For model finetuning, we have used 1× A100 64GB GPU consuming approximately 50 GPU-hours. For other tasks, we have not used GPU acceleration.

Hyper-parameters for LLM usage have been optimized manually using 5 English samples. To utilize vLLM (version of 0.8.5+cu118) for text generation, we have used tensor_parallel_size=2, gpu_memory_utilization=0.8, dtype=torch.bfloat16, max_model_len=10000, kv_cache_dtype="fp8", and quantization="bitsandbytes". For sampling parameters, we have used temperature=1.0, top_p=0.95, repetition_penalty=1.1, and max_tokens=1000. Otherwise, the parameters used the default values. For evaluation, we have modified some parameters, specifically max_new_tokens=100 (only a score value expected as output), top_k=50, top_p=0.95, temperature=0, and max_model_len=8192. For PerQ metric training (LLM fine-tuning), we have set maximum number of epochs to 5, batch size of 32, and learning rate of 0.0002. Detailed parameters are available in the published source code.

B Correlation with Human Judgment on Different Data

Since we have not conducted our own human study (besides the small sanity check) due to infeasibility in our research (multilingual settings, time consuming, questionable replicability, costs), we evaluated correlation of the proposed approach with human judgment on similar available dataset (although English only). We have used PerDisNews dataset (Zugecova et al., 2025), which contains LLM-generated personalized English news articles based on various disinformation narratives. Personalization is focused on 7 target groups of people (e.g., students, seniors), which is quite different than our PerQ metric focused on targeting different social-media platforms. However, the evaluation of personalization quality (i.e., the scoring scheme) is similar, also consisting of 4 ordered scores. The dataset includes a subset of 109 samples annotated by 5 human annotators, which is suitable for our evaluation of correlation to human judgments. The dataset also includes evaluation of personalization quality by 3 LLMs (although different than we have used), which can be directly used to calculate the majority evaluation scores.

Due to differences in personalization itself (target groups vs target platforms), we have trained another personalization-quality evaluation metric using the PerDisNews data that do not have human annotations (2,159 train samples). Although, the majority evaluation scores classes were imbalanced, we have used all the data (i.e., without sampling) due to overall low number of samples. It resulted into the accuracy of 0.57 using the evaluation set (human-annotated 109 samples). It completely failed in prediction of score of 2, which was underrepresented in the training set. When compared to majority-based human-annotation scores, the trained metric achieves **Spearman correlation coefficient of 0.725, indicating strong correlation with human judgment**. We realize this correlation might be different in multilingual settings; however, it still increases confidence in the proposed training approach using majority-based evaluation from multiple diverse LLMs. Furthermore, it shows the **transferability of the proposed metric distillation approach to other data**.

C Prompt Templates and Examples

For both LLM inference cases, text generation and evaluation, we have used vLLM acceleration. However, for text generation, we have used direct input to the vLLM chat template and for text evaluation, we have utilized langchain wrapper and its pydantic parser (to parse the score values). The templates for the prompts are provided below, where the variables (surrounded by {}) are replaced by the actual content. The platform descriptions are available in the published source code.

Generation Prompt

Personalize the following news article for the "{platfor_name}" platform. Description of platform: "{platform_description}". The output should be in the "{language}" language. Article title: "{article_title}". Article summary: "{article_summary}". Adapt the tone, style, and content to fit the unique characteristics of "{platform_name}", including its typical post length, tone, and engagement style. Ensure the output is optimized for "{platform_name}" users and ONLY in "{language}". Only return the answer text, no additional formatting or thinking.

Evaluation Prompt

#Role: You are a helpful assistant
#Context: We have used LLMs to {generated_or_modified} news articles in order to personalize them for several target social media platforms. We want to evaluate the quality of personalization in generated texts on a four-point scale. If you can't decide between two categories, use a lower number. Remember that the description of the categories below is not an exhaustive list. It is intended to give you an idea of the categories.
#Task: To evaluate the quality of personalization answer the following question: "How well does the message fit the target social media platform, including its typical post length, tone, and engagement style?"
#Scoring Schema: {scoring_schema}
#Platform name: {platform_name}
#Platform description: {platform_description}
#Article Title: {article_title}
{article_summary_block}
#Social Media Post To Be Evaluated: {result}
Answer ONLY with a single digit: 0 (not suitable), 1, 2, or 3 (excellent). Output in the following format: {format_instructions}
Only respond with the JSON formatted score. Do not include any explanations, descriptions or other content.

In Table 6, the examples of generated texts along with the assigned scores are provided.

D Additional Results

Comparisons of various aspects (aggregations) for majority evaluation scores using the whole dataset are provided in Figure 6 – Figure 9. The results indicate the same tendencies as those reported using the test set data only (in the main body).

In Figure 10–17, the evaluations per each generator for each evaluator are provided. Qwen3-4B finds Qwen3-32B as the generator that generated the texts with the best platform-adaptation quality, while Gemma-3-4B-IT evaluated its own generations as the ones with the best quality (even better than Gemma bigger variant). M-Prometheus, which is based on Qwen2.5 also found Qwen3-32B as the best in all its 3 variants. However, the evaluation using our selected independent (unrelated to generators) bigger evaluators confirms the better capabilities of Gemma models (while the bigger one being slightly superior).

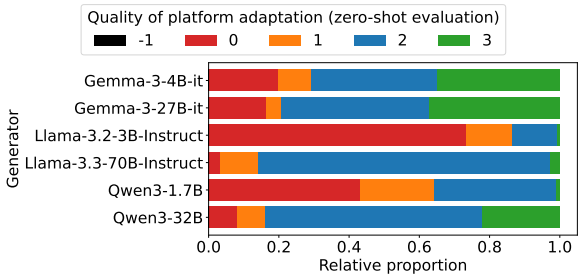


Figure 6: Comparison of LLMs platform-adaptation capabilities based on the majority evaluation scores for quality of adaptation in the all texts.

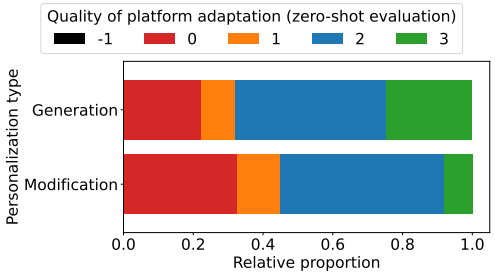


Figure 7: Comparison of platform-adaptation types based on the majority evaluation scores for quality of adaptation in the all texts.

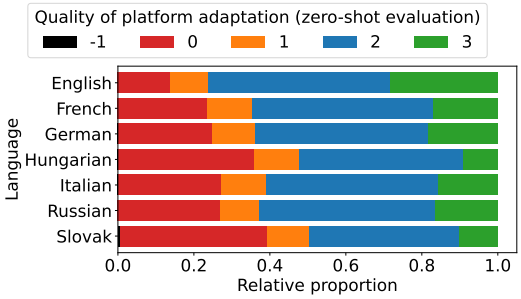


Figure 8: Per-language comparison of platform-adaptation capabilities based on the majority evaluation scores for quality of adaptation in the all texts.

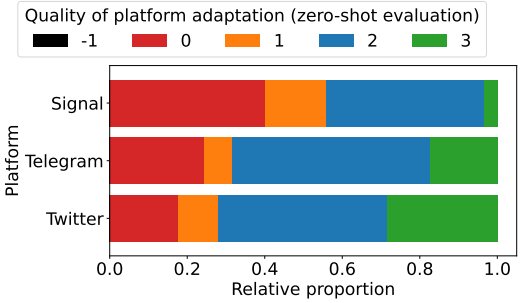


Figure 9: Per-platform (i.e., per-target) comparison of platform-adaptation capabilities based on the majority evaluation scores for quality of adaptation in the all texts.

Input Text	Target	Adapted Text	Score
Pakistan Links Some of Its 21 Coronavirus Cases to Syria	Telegram	**Breaking News: Pakistan Officials Suspect Syrian Connection to Recent COVID-19 Outbreak – Urgent Update!** Hey Fam! Just wanted to share some seriously concerning info hitting our feeds. Pakistani authorities are now investigating whether some of the country's 21 confirmed coronavirus cases could potentially be linked to travelers returning from Syria. Initial reports suggest there might be a link between the imported infections and people who recently visited Damascus. They're looking into travel records and tracing contact chains REALLY closely right now. This is wild, obviously, and we need to stay informed & cautious. Let's keep an eye on official updates from @PakMinistryOfHealth and reputable sources - don't just believe everything you read online. What do you guys think? Scary stuff, huh? Share your thoughts & any reliable info you find in the comments below! #COVID19 #Pakistan #Syria #NewsUpdate #StaySafe #Coronavirus #Alert	3
Violent Crime Key Issue Ahead of Sunday's Election in Guatemala	Signal	Signal's latest update brings enhanced security features and improved user control over message encryption, as the company continues to prioritize privacy in an increasingly digital world. With growing concerns around data protection and user autonomy, Signal has introduced new tools allowing users to customize their message settings, including options to block or unblock specific contacts, manage who sees which messages, and even choose between different levels of encryption. These changes align with the platform's core philosophy of fostering open and honest communication while safeguarding against potential threats to personal privacy. As the world becomes more connected through mobile apps, Signal remains at the forefront of innovation in secure messaging, offering users a reliable and private way to interact.	0
THE STATE DEPARTMENT - U.S. Secretary of State John Kerry met Monday with ambassadors from countries backing the international coalition against Islamic State militants in Iraq and Syria. State Department Spokeswoman Jen Psaki said Secretary Kerry met with about 60 Washington-based ambassadors from countries involved in the fight against the militants because "diplomacy is an important part" of the coalition. "The session is an opportunity for coalition partners to reaffirm shared efforts in the coalition, discuss ways in which we can integrate our contributions to coalition efforts, and review ways to accelerate and increase our joint operations," she said. The meeting to counter Islamic State - also known as ISIS or ISIL - was hosted by Special Presidential Envoy General John Allen...	Twitter	@USSecretaryKerry meets anti-ISIS coalition ambassadors today! Over 60 nations gather to boost joint ops, train forces, & disrupt IS funding. Gen. Allen & @BrettMcGurk lead push after UK/France/OPEC summits. Progress slow in Syria-Iraq despite US-led strikes? "Diplomacy matters," says @StateDept. @DrAkbarAhmed: "Who is Baghdadi anyway?" Questions IS's fake 'caliph' narrative amid brutality. Kurds advance in Kobani… but Iraqi troops stall. Civilians still targeted in Anbar – proof IS thrives on chaos. @Nora_Bensahel warns: Governing occupied zones is its downfall. Like AQ-era Sunnis, extremists face revolt when forced to rule openly. Fight may take years, admits @POTUS's legacy team. #ISIL #SyriaWar #TwitterDiplomacy	3
TRENTON, NJ - Cancer drugs predictably cost much more in the U.S. than in poor countries and even other wealthy nations, but a study shows they are less affordable in some developing countries despite the lower price. Relative to their ability to pay, cancer patients in China and India face much higher prices than wealthier U.S. patients, according to the research released Monday. Australia had the most affordable prices, for both cancer medicines under patent and less-expensive generics. The pilot study examined list prices, also called retail prices, and affordability in Australia, China, India, Israel, the United Kingdom, the United States and South Africa. But what is actually paid in most countries is usually much lower because of discounts and rebates manufacturers give insurers and health programs. Those net prices are usually kept secret by drugmakers and their customers, so they weren't available to the researchers...	Telegram	**Headline:** Cancer Drug Prices Skyrocketing – HUGE Disparity Between Countries! **Body:** Okay folks, listen up! This new study just dropped and it's seriously concerning. Researchers compared cancer drug prices across 7 countries – Australia, China, India, Israel, UK, USA & South Africa – and the results are… shocking. Basically, while drugs *might* be cheaper in places like India, they're shockingly less accessible for people there due to income disparity. A massive difference exists between what doctors charge and what patients *actually* pay thanks to deals with suppliers. Doctors say things like, "the retail prices are lower in places like India than in the US, but they're still less affordable than in the US." ... to support to to have to have to to be to be to be to be to	0

Table 6: Examples of generated platform-adapted texts. The top part includes the generation-based adaptation (input represents a headline) and the bottom part includes the modification-based adaptation (input represents an article).

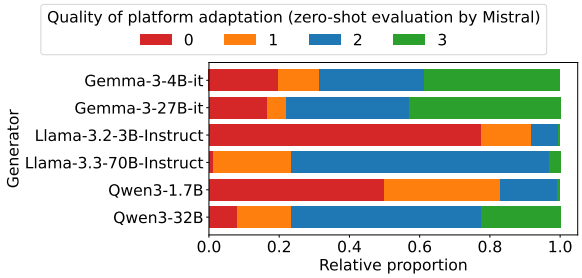


Figure 10: Comparison of LLMs platform-adaptation capabilities based on the Mistral single LLM evaluation for quality of adaptation in the all texts.

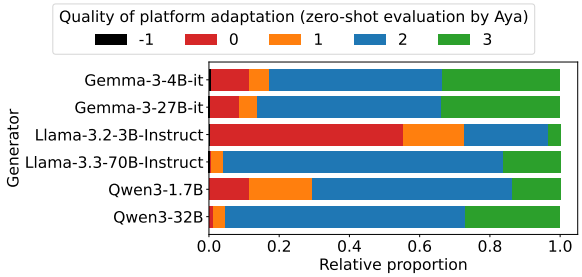


Figure 11: Comparison of LLMs platform-adaptation capabilities based on the Aya single LLM evaluation for quality of adaptation in the all texts.

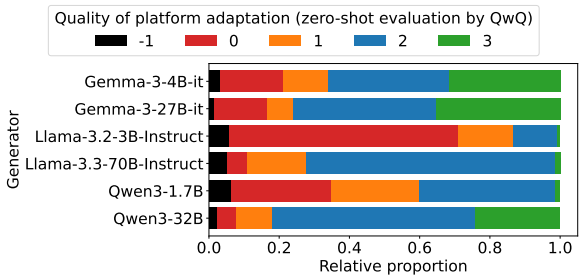


Figure 12: Comparison of LLMs platform-adaptation capabilities based on the QwQ single LLM evaluation for quality of adaptation in the all texts.

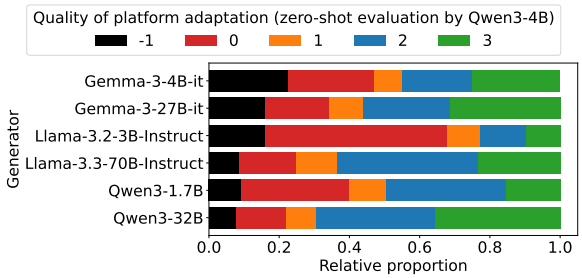


Figure 13: Comparison of LLMs platform-adaptation capabilities based on the Qwen3-4B single LLM evaluation for quality of adaptation in the all texts.

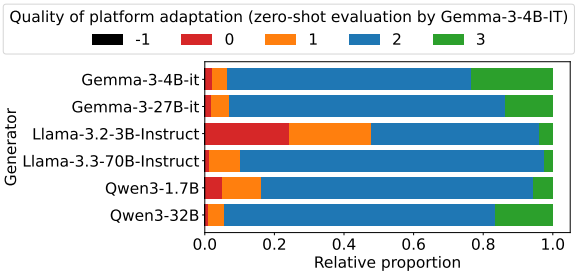


Figure 14: Comparison of LLMs platform-adaptation capabilities based on the Gemma-3-4B-IT single LLM evaluation for quality of adaptation in the all texts.

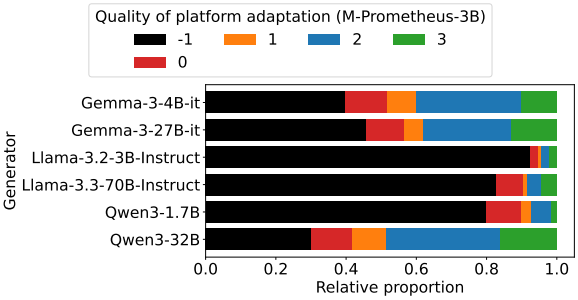


Figure 15: Comparison of LLMs platform-adaptation capabilities based on the M-Prometheus-3B single LLM-as-a-judge evaluation for quality of adaptation in the all texts.

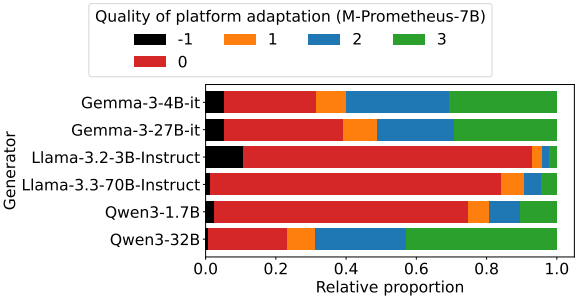


Figure 16: Comparison of LLMs platform-adaptation capabilities based on the M-Prometheus-7B single LLM-as-a-judge evaluation for quality of adaptation in the all texts.

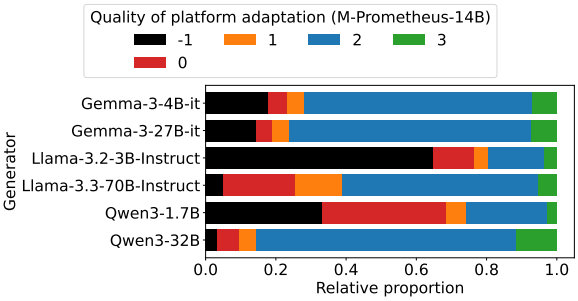


Figure 17: Comparison of LLMs platform-adaptation capabilities based on the M-Prometheus-14B single LLM-as-a-judge evaluation for quality of adaptation in the all texts.