# RoSE: Round-robin Synthetic Data Evaluation for Selecting LLM Generators without Human Test Sets

**Jan Cegin**[♠†], **Branislav Pecher**[†], **Ivan Srba**[†], **Jakub Simko**[†]
[♠] Faculty of Information Technology, Brno University of Technology, Brno, Czechia
[†] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
{jan.cegin, branislav.pecher, jakub.simko, ivan.srba}@kinit.sk

## Abstract

LLMs are powerful generators of synthetic data, which are used for training smaller, specific models. This is especially valuable for low-resource languages, where human-labelled data is scarce but LLMs can still produce high-quality text. However, LLMs differ in how useful their outputs are for training. Selecting the best LLM as a generator is challenging because extrinsic evaluation requires costly human annotations (which are often unavailable for low-resource languages), while intrinsic metrics correlate poorly with downstream performance. We introduce Round-robin Synthetic data Evaluation (RoSE), a proxy metric for selecting the best LLM generator without human test sets. RoSE trains a small model on the outputs of a candidate generator (LLM) and then evaluates it on generated synthetic examples from all other candidate LLMs. The final RoSE score is the mean performance of this small model. Across six LLMs, eleven languages, and three tasks (sentiment, topic, intent), RoSE identifies the optimal generator more often than any other intrinsic heuristics. RoSE outperforms intrinsic heuristics and comes within 0.76 percentage points of the optimal generator baseline. This result is measured in terms of downstream performance, obtained by training a small model on the chosen generator's outputs (optimal vs. proxy-metric–selected) and evaluating it on human-labelled test data. Additionally, RoSE is the only metric to achieve a positive correlation with performance on human test data.

## 1 Introduction

Current large language models (LLMs), such as GPT4, Llama, and others, show impressive performance in generating well-formed texts (Cegin et al., 2023). Such synthetic texts are commonly leveraged to train smaller, more efficient downstream models, enhancing their performance (Wang et al., 2025; Cegin et al., 2025). While most of the previous research has focused on improving the LLMs' ability to generate representative synthetic data in English (Piedboeuf and Langlais, 2023; Cegin et al., 2024; Wang et al., 2025), some research has also focused on generative strategies and LLM evaluation for low-resource languages (Anikina et al., 2025). Even for low-resource languages, LLMs can produce high-quality samples (Son et al., 2025; Chung et al., 2025; Anikina et al., 2025). However, previous research has identified that differences exist between LLMs when generating texts in low-resource languages, as their quality for training smaller models can vary significantly (Anikina et al., 2025). It is therefore important to have measures for evaluating the LLM's performance for generating textual data.

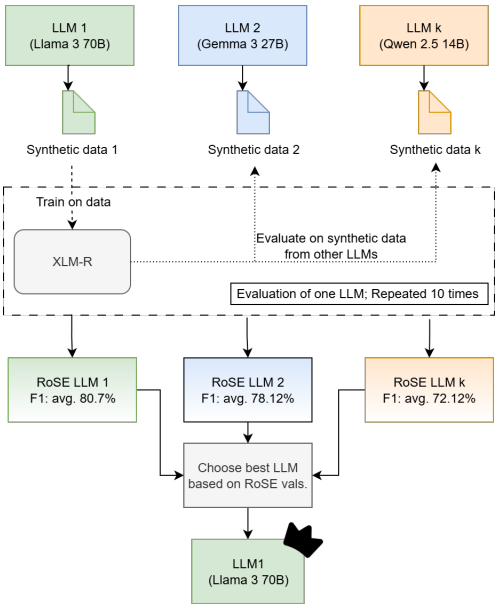



Figure 1: Overview of RoSE proxy metric calculations. For each candidate generator (LLM), we first generate synthetic data. Then, we train a smaller model on each generator's synthetic dataset and evaluate it on the synthetic data generated by all other generators. The mean F1 score across these test sets is the final RoSE score for that LLM. The highest RoSE score LLM is considered the best generator.

Evaluation methods for the purpose of identifying the best LLM generator generally fall into two categories: intrinsic and extrinsic metrics (Li et al., 2024). Intrinsic evaluation measures properties of the generated text itself, such as vocabulary size, type–token ratio, or diversity, without training a downstream model (and without much dependence on the task). Extrinsic evaluation, in contrast, assesses the usefulness of synthetic data indirectly, by training a downstream model on it and evaluating performance on human-labelled test data. As this relies on human annotations, extrinsic evaluation is treated as an oracle, the gold standard for measuring the true utility of generated data (Cegin et al., 2023; Li et al., 2024), though it is rarely available in low-resource settings (Anikina et al., 2025).

The challenge is that for many low-resource language-task combinations, human test sets do not exist (Anikina et al., 2025), or, at most, only a very small number of human samples ($\approx$10 per label) exist. In these settings, evaluation of LLMs as synthetic data generators and their selection relies only on intrinsic metrics or simple heuristics. Yet prior work shows that intrinsic metrics do not reliably correlate with extrinsic outcomes even for English data: correlations may be weak (Wang et al., 2025) or inconsistent (Cegin et al., 2025). As the quality of generated data in low-resource languages can vary widely across various LLMs used as generators (Anikina et al., 2025), **reliable proxy metrics are needed to identify the best LLM for synthetic data generation when human evaluation data is unavailable**.

For such cases, when human evaluation data is unavailable, we propose our proxy metric, Round-robin Synthetic data Evaluation (RoSE), visualised in Figure 1, for evaluating LLMs as generators and identifying the best generator, where human test data is not available. The intuition is that synthetic data carries the representational signature of its generator, and since different LLMs produce data of varying quality and coverage (Anikina et al., 2025), cross-evaluating on each other's outputs could reveal which generator provides the most generalizable training signal. First, given a set of evaluated LLMs as generators from which we aim to find the best generator, we generate synthetic text for each task–language combination per each LLM via the current state-of-the-art method, which leverages a small (10 per label) amount of human examples. Second, a smaller model is trained on one generator's data and evaluated on synthetic test sets from all other generators. The performance of that LLM as a generator is represented as the mean over all the synthetic test sets. This process is repeated for all candidate LLMs 10 times, where the best LLM is chosen based on its highest mean performance.

To evaluate RoSE, we conduct a comparative analysis of proxy metrics, including intrinsic metrics, simple heuristics, and our proposed approach (RoSE). Experiments span 11 typologically diverse languages with varied scripts, covering several very low-resource cases such as Welsh, Romanian, Azerbaijani, etc., across three classification tasks (sentiment, topic, and intent). We evaluate six open-weight LLMs of different sizes and families on each task–language combination. To test whether a proxy metric can replace human-labelled data for identifying the best generator LLM, we treat the generator selected using human validation sets as an optimal generator. For each proxy metric, we measure the F1 performance gap between (a) a small model trained on data from the proxy-selected generator and (b) a model trained on data from the optimally-selected generator and evaluated on human test data. We also report Pearson correlations between each metric's scores and the corresponding LLMs' human test performance, as well as ranking-based correlations.

Our findings show that:

- Across all languages and tasks, RoSE identifies the optimal LLM generator more often than any other proxy metric.
- LLM generators selected by RoSE achieve an average gap of only 0.76% F1 compared to the optimal human-performance-based generator selection, versus 2.52% for the second-best proxy metric.
- RoSE is the only proxy metric that consistently yields a positive correlation between classifier performance and human evaluation.
- RoSE ranks best across 9 of 11 languages, and second-best in the remaining 2.

Additional ablations further demonstrate that RoSE remains effective even when comparing as few as three LLMs. RoSE also performs strongly when comparing LLMs of similar parameter size. We also further analyse how the number of LLMs used for computing RoSE affects selection quality and how the generation setup influences RoSE's

performance. We release all our code, data and results in https://github.com/kinit-sk/RoSE.

## 2 Related Work

Since their introduction, large language models (LLMs) such as GPT-4 and Llama have been increasingly adopted as tools for data augmentation and generation. The synthetic data they produce is commonly used to train smaller downstream models, improving efficiency while maintaining strong performance. This approach has been applied across a wide range of tasks, including automated scoring (Fang et al., 2023), intent classification (Sahu et al., 2022), sentiment analysis (Piedboeuf and Langlais, 2023; Onan, 2023; Yoo et al., 2021), hate speech detection (Sen et al., 2023), news classification (Piedboeuf and Langlais, 2023), content recommendation (Liu et al., 2024), and health symptom classification (Dai et al., 2023).

While most of the generation has been done in English, recent research has also focused on the usage of LLM for generating synthetic texts in low-resource languages. Multilingual synthetic generation using LLMs has been leveraged for various tasks like QA (Kramchaninova and Defauw, 2022; Namboori et al., 2023; Putri et al., 2024), fact-checking (Chung et al., 2025), reasoning (Pranida et al., 2025), NER (Liu et al., 2021), sentiment stance detection (Zotova et al., 2021) and classification (Glenn et al., 2023; Anikina et al., 2025).

Recent work has introduced efforts to benchmark LLMs as synthetic data generators, most notably with AgoraBench (Kim et al., 2025). While this benchmark is effective at identifying top-performing LLMs for generation tasks, its primary focus lies in post-training decoder-based models on synthetic data rather than training downstream encoder-based models. Currently, there is no established benchmarking framework for selecting the most suitable LLM generator for downstream classification tasks. Such selection is often performed on a case-by-case basis using extrinsic evaluation on human-annotated test sets corresponding to the target task (Glenn et al., 2023; Anikina et al., 2025). However, human-labelled test sets may be unavailable for many tasks and languages, and their creation can be prohibitively expensive (Putri et al., 2024; Pranida et al., 2025; Gurgurov et al., 2025). Thus, identifying optimal LLM generators through reliable proxy metrics becomes crucial, as it avoids the high cost of human evaluation.

## 3 Methodology and Experiments

### 3.1 Round-robin Synthetic Data Evaluation (RoSE)

We propose **Round-robin Synthetic data Evaluation (RoSE)** as a proxy metric for identifying the best large language model (LLM) for synthetic data generation in the absence of human test sets. The metric calculations are visualised in Figure 1.

Synthetic data produced by LLMs carries the representational signature of its generator, with different models exhibiting substantial variation in coverage and quality when generating texts in low-resource languages (Anikina et al., 2025). A robust generator should produce data that transfers well across distributions, such that a classifier trained on its outputs generalises effectively to data generated by other LLMs. This intuition motivates RoSE as a proxy for extrinsic evaluation, where human-labelled test sets act as an oracle.

To compute the method, we start with a given set of candidate LLMs, and we proceed in three steps in a round-robin manner. First, for each task–language combination, every LLM generates synthetic training and test data using a state-of-the-art generation procedure. Second, a small classifier is trained on the synthetic data produced by one LLM and evaluated on the synthetic test sets generated by all other LLMs. Third, the performance of the LLM is defined as the mean score of this classifier across all cross-evaluations. The process is repeated for all candidate LLMs 10 times (to mitigate randomness as per (Pecher et al., 2024)), and the generator achieving the highest mean performance is selected as the best model.

### 3.2 Intrinsic Metrics

We use a wide variety of 8 different intrinsic metrics to evaluate their effectiveness against our proposed RoSE metric. The tokenisation is done via the *XLMRoberta-base* tokeniser.

**Average Pairwise Cosine Distance:** Embeddings are computed for all samples within each label and measure the average cosine distance between sample pairs. This captures the semantic diversity (Reimers and Gurevych, 2019) of generations conditioned on the same label[1].

**Bigram Diversity:** This measures (Li et al., 2016) the proportion of distinct bigram tokens rela-

[1]The model used for this is: *paraphrase-multilingual-MiniLM-L12-v2*

tive to the total number of bigrams in the dataset, reflecting diversity at the phrase level.

**Number of Valid Samples:** We count the proportion of syntactically or semantically valid generations that remain after applying a revision step (as described in Section 3.5), reflecting robustness and generation quality.

**Silhouette Score:** Using embeddings, we compute the silhouette coefficient to assess clustering quality of the generated data, measuring how well samples group by their intended label compared to other labels (Rousseeuw, 1987).

**Type–Token Ratio (TTR):** This metric (Johnson, 1944) quantifies lexical diversity by normalising the number of unique tokens with respect to the total number of tokens.

**Token Entropy:** We measure the entropy of the token distribution in the generated data, where higher entropy indicates greater lexical variety and unpredictability (Rosillo-Rodes et al., 2024)

**LLM Parameter Size:** A simple heuristic that always chooses the LLM with the largest number of parameters, motivated by the observation that larger models often perform better.

**Random:** As a naive baseline, we select the best LLM uniformly at random across ten runs and average the resulting scores.

### 3.3 Data and Tasks

We evaluate RoSE and other proxy metrics on three classification tasks: intent recognition, topic classification, and sentiment analysis. For each task, we generate synthetic data in 11 typologically diverse languages. The selection includes two high-resource languages (English, German), four mid-resource languages (Thai, Hebrew, Indonesian, Swahili), and five low-resource languages (Romanian, Azerbaijani, Slovenian, Telugu, Welsh). This choice reflects both linguistic diversity and the availability of datasets for the target tasks.

For intent recognition, we use the MASSIVE dataset (FitzGerald et al., 2023), a multilingual benchmark for virtual assistant evaluation covering 51 languages and 60 intents. To simplify the task, we restrict the label set to the ten most common intents.

For topic classification, we rely on SIB-200 (Adelani et al., 2024), which provides seven topic labels derived from the FLORES-200 machine translation corpus, annotated at the sentence level.

For sentiment classification, no single multilingual dataset spanning both high- and low-resource languages currently exists. We therefore combine ten datasets for low-resource languages from (Gurgurov et al., 2025, 2024) with two additional datasets for English and German from (Mollanorozy et al., 2023). As noted in (Anikina et al., 2025), these datasets vary in coverage and text domain: for instance, the German set focuses on transportation and infrastructure, while Romanian data primarily consists of product reviews.

### 3.4 Models

We use up to 6 LLMs of different sizes: Gemma-3 (Team, 2025) with 4 and 27 billion parameters, Magistral *Small* with 24 billion parameters, Qwen 2.5 (Team, 2024) with 14 billion parameters, and Llama-3 (AI@Meta, 2024) with 8 and 70 billion parameters. These models were chosen for their open-weight nature and support for multiple languages. For finetuning a smaller model, we used the XLM-R *Base* (Conneau et al., 2019) model.

### 3.5 Experimental Setup

We evaluate our proxy metric, RoSE, alongside the selected intrinsic metrics from Section 3.2 across all task–language combinations (33 in total). For each combination, we first generate 100 samples per label from each LLM under comparison. To ensure strong baselines, we adopt the best-performing generation setup identified by (Anikina et al., 2025), which uses prompts with 10 in-context randomly selected human examples from the train set and applies self-revision as a filtering mechanism. We also test a different, worse-performing setup of excluding the in-context human examples. Results for this ablation can be found in Section 5. Details about generation setup can be found in Appendix A.4.

Next, we fine-tune XLM-R (Conneau et al., 2019) ten times for each task–language combination. The trained classifiers are then evaluated both on synthetic data generated by other LLMs (to compute RoSE) and on human-annotated data (to identify the best-performing LLM). Finally, after fine-tuning, we compute the remaining intrinsic metrics and report the corresponding F1 scores for each case. Details regarding the downstream model fine-tuning can be found in A.3.

To identify the best proxy metric for LLM selection, we proceed as follows. For each task–language case, we train classifiers on synthetic data generated by each LLM and evaluate them on human-annotated data. The resulting av-

| Proxy Metric | Top-1 Match | Top-3 Match |
|---|---|---|
| Avg. cos. dist. | 7 (21.21%) | 8 (24.24%) |
| Bigram Div. | 5 (15.15%) | 2 (6.06%) |
| No. Samples | 7 (21.21%) | 0 (0.00%) |
| Silhouette Score | 4 (12.12%) | 2 (6.06%) |
| Type Token Ratio | 3 (9.09%) | 4 (12.12%) |
| Token Entropy | 4 (12.12%) | 2 (6.06%) |
| LLM param. size | 12 (36.36%) | 0 (0.00%) |
| Random | 5 (15.15%) | 2 (6.06%) |
| RoSE (ours) | **20 (60.61%)** | **15 (45.45%)** |

Table 1: Number of times each proxy metric correctly identified the optimal LLM generator (Top-1) or correctly identified the top 3 generators (Top-3, irrespective of order) across all tasks and languages. RoSE achieves the highest performance in both cases.

erage F1 scores provide a ranking of the LLMs as generators, from worst to best, and enable us to identify the optimal LLM generator. For each proxy metric, we then compute an alternative ranking of the LLMs based on their values. To assess the quality of these rankings, we measure three outcomes. First, the correct identification: how many times was the optimal LLM generator identified by the proxy metric, and how many times were the best 3 LLM generators (irrespective of order) identified by the proxy metric. Second, the performance gap: the difference in F1 score between the downstream model trained with the LLM chosen by the proxy and the downstream model trained with the optimally-selected LLM (ideal outcome = 0%). Third, the correlations: the Pearson correlation between the proxy metric values and the classifiers' performance on human data. We also look at rank correlations via Kendal $\tau$ to compare order rankings for each proxy metric.

## 4 Results and Discussion

When comparing our RoSE metric with the proxy metrics using Top-1 and Top-3 evaluation in Table 1, Top-1 is achieved in 20 out of 33 cases (60.60%), outperforming other proxy metrics (next best at 36.36%). For Top-3, it matches the top 3 generators in 15 (45.45%) cases compared to the next best proxy at 8 cases (24.24%).

Figure 2 summarises the aggregated results across all tasks and languages when comparing six LLMs to identify the best generator. The reported values represent the performance gap in F1 score between downstream models trained on data from the LLM selected by a given metric and those trained on data from the optimal LLM generator.

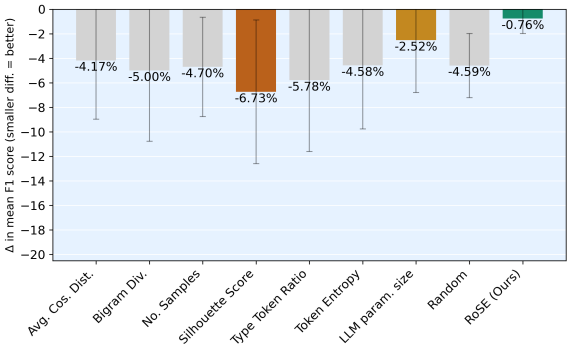



Figure 2: Comparison of proxy metrics for selecting the best LLM generator. Bars show the average gap in mean F1 score for models trained on the best generator selected by metrics vs. the optimal generator (smaller is better). The models are evaluated on human test data. The best metric is green, the second best is orange, and the worst is red.

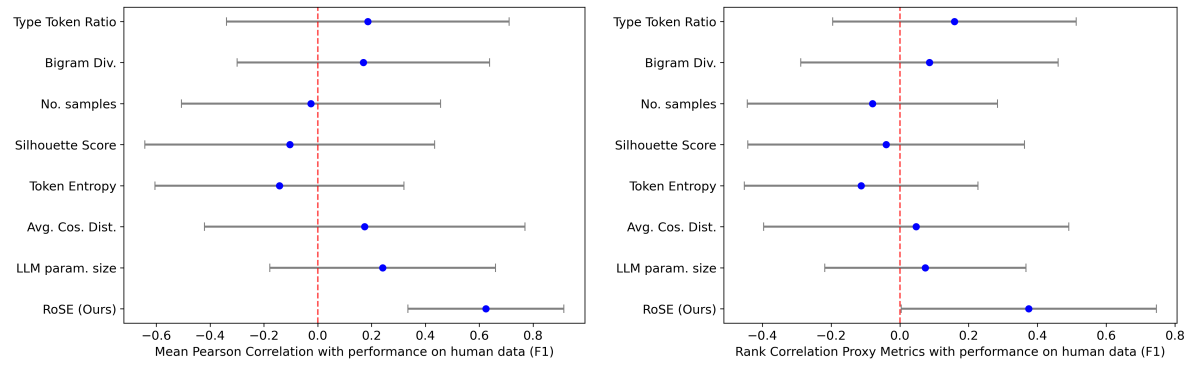
| Proxy Metric | Intent | Topic | Sentiment |
|---|---|---|---|
| Avg. cos. dist. | -2.84% | -2.58% | -7.10% |
| Bigram Div. | -3.05% | -5.27% | -6.67% |
| No. Samples | -3.26% | -4.82% | -6.02% |
| Silhouette Score | -2.70% | -8.02% | -9.47% |
| Type Token Ratio | -3.40% | -5.59% | -8.35% |
| Token Entropy | -3.48% | -6.09% | -4.17% |
| LLM param. size | -1.73% | **-0.31%** | -5.52% |
| Random | -3.05% | -4.65% | -6.07% |
| RoSE (Ours) | **-0.64%** | -0.77% | **-0.86%** |

Table 2: Comparison of proxy metrics for selecting the best LLM generator per task. The table shows the gap in the mean F1 score for models trained on the best generator selected by metrics vs. the optimal generator (smaller is better). The best metric is bold, the second best is underlined. RoSE performs well across all tasks as either the best or the second-best proxy metric.

In the cases where the optimal LLM is not selected, RoSE's selected LLM yields a downstream model whose performance is very close to a model trained on data from the optimal generator, resulting in the lowest average performance gap among all evaluated metrics (0.76%). We also note that on average, only 3 intrinsic metrics achieve statistically significantly (*p=0.05* for Mann-Whitney U test) better generator selection than choosing a generator at random. Considering that the second-best metric is the LLM parameter size, which always selects the largest LLM as the generator based on its parameter size, we can also conclude that RoSE successfully identifies cases where the largest LLM is not the optimal generator, resulting in efficient solutions.

| Proxy Metric | az | cy | he | th | sw | sl | en | de | id | ro | te |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Type Token Ratio | -5.07 | -8.62 | -14.82 | -4.05 | -6.56 | -7.95 | -3.80 | -0.95 | -5.69 | -2.65 | -3.41 |
| Bigram Div. | -3.30 | -6.56 | -14.82 | -1.36 | -6.56 | -7.85 | -3.80 | -1.65 | -3.00 | -2.65 | -3.41 |
| Token Entropy | -5.68 | -6.56 | -14.82 | -2.39 | -6.56 | -2.43 | -1.22 | -1.65 | -3.00 | -2.65 | -3.41 |
| Silhouette Score | -4.71 | -12.76 | -5.60 | -9.83 | -4.84 | -3.55 | -7.57 | -4.95 | -4.09 | -3.52 | -12.62 |
| Avg. cos. dist. | -4.58 | -8.48 | -6.55 | **0.00** | -2.51 | -7.95 | -3.80 | -3.91 | **-1.09** | -1.86 | -5.17 |
| No. Samples | -5.41 | -2.44 | -1.13 | -6.47 | -4.98 | -0.71 | -11.51 | -5.39 | -3.82 | -6.11 | -3.73 |
| LLM param. size | -2.03 | **-0.77** | -2.45 | -2.91 | -2.33 | -1.34 | -2.22 | -4.31 | -1.14 | -0.78 | -7.43 |
| Random | -3.61 | -5.96 | -6.35 | -4.28 | -5.34 | -2.48 | -4.90 | -3.01 | -3.87 | -4.36 | -6.29 |
| RoSE (ours) | **-1.77** | -0.90 | **0.00** | **0.00** | **-1.82** | **-0.43** | **-0.68** | **0.00** | -1.11 | **-0.43** | **-1.19** |

Table 3: Mean gap between F1 performance for downstream models trained on the LLM generator selected by proxy metric vs. optimal LLM generator. The lower the gap, the better the proxy metric. The best proxy for language is bolded, and the second-best is underlined. The best proxy metric for each language is bolded.




(a) Pearson correlation.

(b) Rank (Kendal $\tau$) correlation.

Figure 3: Forest plot showing mean Pearson and Rank (Kendal $\tau$) correlations between proxy metrics (except for random, as it has no values) and downstream performance of models on human-labelled data (F1). Error bars denote confidence intervals. RoSE is the most reliable proxy metric; its average Pearson and rank correlations with human F1 are clearly higher than all intrinsic metrics.

### 4.1 Task Dimension

Table 2 presents the per-task comparison for generator selection. Visualisation can be found in Appendix A.7. The LLM parameter size heuristic proves effective for topic classification but performs poorly on sentiment analysis. This discrepancy arises because the heuristic identifies Llama 3 70B as the optimal LLM: while this model captures the topic task well, it struggles to generate high-quality data for sentiment analysis, the most diverse task (see Section 3.3).

By contrast, RoSE delivers consistently strong performance across tasks. It achieves the best results in both intent and sentiment tasks, ranking second in topic classification.

### 4.2 Language Dimension

Table 3 presents the per-language comparison of proxy metric performance for selecting an LLM generator. Visualisation can be found in the Appendix A.7. In two cases, RoSE is the second-best metric for selecting an optimal LLM generator (Indonesian and Welsh). However, in both cases, the gap in F1 performance is relatively small.

For all the other cases, RoSE is the best proxy metric for selecting the optimal generator. RoSE is the best proxy metric for languages with various transcripts, working well for Telugu, Thai or Hebrew. We also note that no other proxy metrics work consistently well: the second-best proxy of LLM parameter size fails on languages such as German, Hebrew, Thai or Telugu.

### 4.3 Proxy Metric Correlations with Downstream Model Performance

Figure 3a presents a forest plot of mean Pearson correlations between each proxy metric and downstream classifier performance on human-labelled test data, with error bars indicating confidence intervals. Our results highlight the limitations of existing intrinsic metrics. Measures such as silhouette score and token entropy often exhibit weak or even negative correlations with human-based performance, while type–token ratio and bigram

diversity yield only small and unstable positive correlations. Heuristic proxies, such as dataset size and LLM parameter size, also fail to provide reliable guidance, as their correlations are low and the confidence intervals cross zero. In contrast, RoSE achieves a consistently strong correlation with downstream performance, with a mean Pearson correlation of around 0.6 and confidence intervals entirely above zero.

We also provide Rank correlations (via Kendal $\tau$) in Figure 3b to investigate if we can rank the performance of LLMs as generators by using proxy metrics. RoSE is the only metric to achieve mild positive correlations, indicating that it consistently points in the right direction for LLM ranking as generators. However, the correlations in both cases are not near-perfect, indicating that the gap to optimal LLM selection and ranking remains. Our findings confirm prior observations that intrinsic metrics are not reliable indicators of synthetic data utility (Cegin et al., 2024).

### 4.4 Excluding Largest LLM From Comparison

The second-best metric in our main results is the LLM parameter size heuristic, which consistently selects Llama 3 70B. To test the viability of this metric and our own RoSE, we conducted an ablation study excluding Llama 3 70B and comparing the remaining five LLMs listed in Section 3.4, where the next largest model is Gemma 3 27B. The results, shown in Figure 7, demonstrate that RoSE remains a strong proxy, with only a -1.26% F1 gap relative to the optimal LLM generator. In contrast, the next-best metric, average cosine distance, shows a larger gap of -3.62%, while the parameter size heuristic drops to -5.78%, performing worse than random selection. These findings highlight that RoSE continues to perform reliably even when competing LLMs are closer in parameter scale.

### 4.5 Performance on Varying Number of Candidate LLMs

We evaluate how RoSE compares to other proxy metrics in selecting LLM generators across different numbers of models. Specifically, we consider all unique combinations of 2 to 6 LLMs. For the case of 6 LLMs, the results coincide with those shown in Figure 2. The outcomes for each number of LLMs are illustrated in Figure 4, highlighting RoSE alongside the strongest-performing alternative metrics. The visualisation breakdown per task

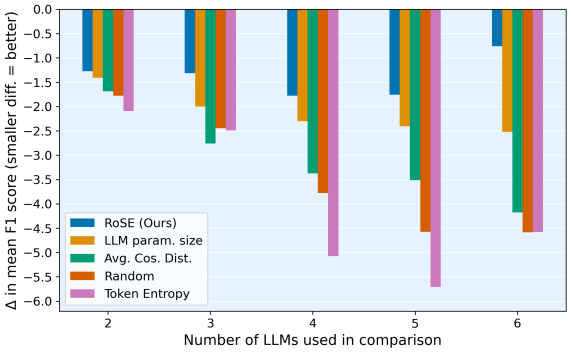



Figure 4: Comparison of a selection of proxy metrics for selecting the best LLM generator when comparing various combinations of LLMs (all combinations of up to 6 LLMs are considered). RoSE is the best proxy metric for a varying number of LLMs compared.

can be found in Appendix A.5.

Across all settings, RoSE consistently emerges as the best proxy metric for identifying the optimal LLM generator, with the smallest performance gap observed when comparing two models. Moreover, RoSE exhibits the most stable mean gap across varying numbers of LLMs, underscoring its predictive reliability across different model combinations.

### 4.6 Cost Effectiveness of RoSE

Although RoSE demonstrates strong performance in selecting the optimal LLM generator, it is also computationally expensive. For each LLM under comparison, RoSE requires repeatedly training a smaller model and evaluating it on data generated by the other LLMs. In contrast, embedding-based approaches or simple heuristics, such as LLM parameter size, are far less costly to compute. An overview Table 8 can be found in Appendix A.7 with relative computational costs and performance.

To explore how reducing the cost of RoSE affects its performance, we consider a simplified variant. Instead of evaluating the trained smaller model on test sets from all other LLMs, we randomly select a single alternative LLM to provide the evaluation set. This process is repeated for every LLM under comparison 10 times. We then assess whether this random selection strategy leads to a significant drop in RoSE's ability to identify the optimal LLM generator. We consider selecting randomly 1 or a combination of various LLMs for up to all 6. The results are shown in Figure 5. For per-task breakdown, see Appendix A.6.

Our results show that RoSE outperforms all other proxy metrics with as few as three LLMs used for

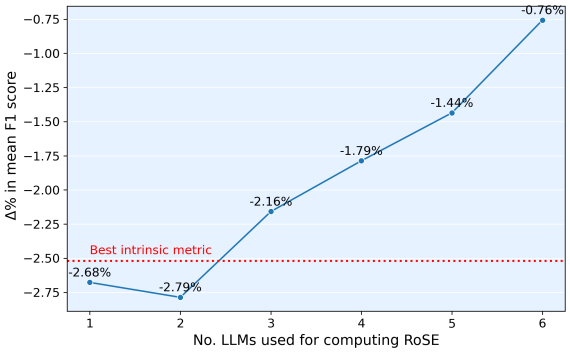



Figure 5: Number of randomly chosen LLMs used for computing RoSE and its effect on F1 score difference to optimal LLM generator selection. RoSE benefits from more LLMs being used during the evaluation of the downstream classifier during its computation.

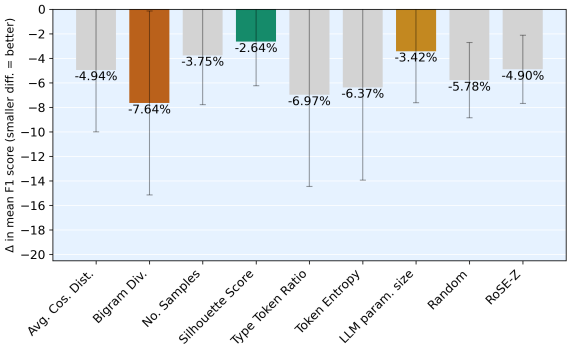



Figure 6: Comparison of proxy metrics for selecting the best LLM generator when generation setup without examples is used (zero-shot). RoSE-Z in this setup struggles to identify the optimal LLM generator, underlining the importance of including in-context examples in the data generation process.

evaluation, surpassing the next best proxy of LLM parameter size. With the exception of the two-LLM case, the performance gap when using RoSE to select the optimal LLM generator decreases as more models are included. The combination of two randomly chosen LLMs for computing RoSE can lead to significant variability of the computed values by providing proxy performance that can be skewed. This suggests that RoSE benefits from a larger pool of LLMs during evaluation, providing a closer approximation to human performance.

We further examine whether data from certain LLMs consistently degraded RoSE's effectiveness. While some models performed poorly in specific languages (e.g., Qwen 2.5 on Hebrew), we did not find any single LLM whose outputs were consistently detrimental. Thus, excluding specific LLMs from RoSE computation is not necessary.

## 5 The Role of Prompt Examples in RoSE

As described in Section 3.5, we follow the strategy of (Anikina et al., 2025), prompting LLMs with 10 human examples during data generation. This in-context approach has been shown to yield the best downstream performance.

To examine how these examples affect RoSE's predictive strength, we also evaluate a zero-shot variant, where prompts contain no examples and only LLM self-revision is applied, similar to the revision-only strategy of (Anikina et al., 2025). We denote our metric in this setting as **RoSE-Z**.

**We note that classifiers trained on zero-shot data perform substantially worse on human test sets, with an average drop of 8.45% in F1**, consistent with prior findings (Piedboeuf and Langlais, 2023; Anikina et al., 2025). Moreover, RoSE-Z is less effective at identifying the optimal LLM generator: while better than random selection, three alternative metrics outperform it (see Figure 6).

These results suggest that RoSE's strong predictive power depends on the inclusion of human examples in prompts. RoSE's effectiveness as a proxy metric seems to arise from its ability to approximate human-like data generation, making it a reliable proxy for extrinsic evaluation. This is in line with previous research that identified including examples during generation as beneficial for the downstream model performance and that such data resembles human data more closely (Anikina et al., 2025; Cegin et al., 2024).

## 6 Conclusion

In this work, we proposed RoSE (Round-robin Synthetic data Evaluation) as a principled proxy metric for selecting the most effective LLM generator in scenarios where human-labelled test sets are unavailable. Our extensive experiments across 11 languages and three diverse classification tasks demonstrate that RoSE consistently outperforms intrinsic metrics and heuristic baselines, identifying the optimal LLM most consistently and achieving an average F1 gap of just 0.76% for chosen LLM generators by this proxy. Beyond its predictive strength, RoSE proves reliable across tasks, languages, and even when the number of candidate LLMs is small, making it a practical tool for low-resource settings. Overall, RoSE represents a step toward the reliable and cost-effective selection of LLM generators in data-scarce scenarios.

## Limitations

Due to the scope of this study, we limited ourselves to 6 different LLMs used. We tried to mitigate this limitation by including LLMs of various sizes and from various LLM families. We additionally used only 11 languages and 3 tasks due to resource constraints and the availability of evaluation data for the low-resource languages.

We limited the number of used labels for the MASSIVE dataset to the 10 most common intents to avoid many conflating factors that can be potentially caused by the semantic overlaps in label descriptions.

We acknowledge that the extent to which data contamination affects RoSE's performance is unknown. The data on which LLMs were trained is not fully disclosed, which might be reflected in the disparities between the LLMs' performance when generating data in different languages and domains.

A limitation of the RoSE method is its requirement of needing a few (10 per label) human examples for the generation process. As we theorise in the paper, the generated data is then more similar to the human data we are trying to replace during evaluation (and computation of RoSE). Without it, RoSE performs significantly worse, but so do downstream models trained on data generated without human examples included. As such, a small number of human examples is still essential, both for RoSE performance and the general good performance of LLMs as generators.

## Ethical Considerations

Based on a thorough ethical assessment performed on the basis of intra-institutional ethical guidelines and checklists tailored to the use of data and algorithms, we see no ethical concerns pertaining directly to the conduct of this research. Although the production of new data through LLMs bears several risks, such as the introduction of biases, the small size of the produced dataset, sufficient for experimentation, is, at the same time, insufficient for any major machine learning endeavours where such biases could be transferred.

We follow the license terms for all the models and datasets we used (such as the one required for the use of the Llama-3 model) – all models and datasets allow their use as part of the research.

## References


David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Tatiana Anikina, Jan Cegin, Jakub Simko, and Simon Ostermann. 2025. A rigorous evaluation of llm data generation strategies for low-resource languages. *Preprint*, arXiv:2506.12158.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024. Effects of diversity incentives on sample diversity and downstream model performance in LLM-based text augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13148–13171, Bangkok, Thailand. Association for Computational Linguistics.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1889–1905, Singapore. Association for Computational Linguistics.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2025. LLMs vs established text augmentation techniques for classification: When do the benefits outweight the costs? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10476–10496, Albuquerque, New Mexico. Association for Computational Linguistics.

Yi-Ling Chung, Aurora Cobo, and Pablo Serna. 2025. Beyond translation: Llm-based data generation for multilingual fact-checking. *arXiv preprint arXiv:2502.15419*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *Preprint*, arXiv:2302.13007.

Luyang Fang, Gyeong-Geon Lee, and Xiaoming Zhai. 2023. Using gpt-4 to augment unbalanced data for automatic scoring. *Preprint*, arXiv:2310.18365.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using synthetic data and transfer learning for multilingual ESG issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139, Macao. -.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 63–74, Bangkok, Thailand. Association for Computational Linguistics.

Daniil Gurgurov, Rishu Kumar, and Simon Ostermann. 2025. GrEmLIn: A repository of green baseline embeddings for 87 low-resource languages injected with multilingual graph knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1204–1221, Albuquerque, New Mexico. Association for Computational Linguistics.

William H. Johnson. 1944. Type–token ratio: A measure of linguistic diversity. *Journal of Applied Linguistics*, 1:27–34.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2025. Evaluating language models as synthetic data generators. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6385–6403, Vienna, Austria. Association for Computational Linguistics.

Alina Kramchaninova and Arne Defauw. 2022. Synthetic data generation for multilingual domain-adaptable question answering systems. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 151–160, Ghent, Belgium. European Association for Machine Translation.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2024. LLatrieval: LLM-verified retrieval for verifiable generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5453–5471, Mexico City, Mexico. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Once: Boosting content-based recommendation with both open- and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 452–461, New York, NY, USA. Association for Computing Machinery.

Sepideh Mollanorozy, Marc Tanti, and Malvina Nissim. 2023. Cross-lingual transfer learning with Persian. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 89–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Amani Namboori, Shivam Sadashiv Mangale, Andy Rosenbaum, and Saleh Soltan. 2023. Gemquad: Generating multilingual question answering datasets from large language models using few shot learning. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*.

Aytuğ Onan. 2023. Srl-aco: A text augmentation framework based on semantic role labeling and ant colony optimization. *Journal of King Saud University - Computer and Information Sciences*, 35(7):101611.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. A survey on stability of learning with limited labelled data and its sensitivity to the effects of randomness. *ACM Computing Surveys*, 57(1).

Frédéric Piedboeuf and Philippe Langlais. 2023. Is ChatGPT the ultimate data augmentation algorithm? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore. Association for Computational Linguistics.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, and Fajri Koto. 2025. Synthetic data generation for culturally nuanced commonsense reasoning in low-resource languages. *Preprint*, arXiv:2502.12932.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? case study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Pablo Rosillo-Rodes, Maxi San Miguel, and David Sanchez. 2024. Entropy and type-token ratio in gigaword corpora. *Physical Review Research*, 6(4):043236.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Tran Ngoc Son, Nguyen Anh Tu, and Nguyen Minh Tri. 2025. An efficient approach for machine translation on low-resource languages: A case study in vietnamese-chinese. *arXiv preprint arXiv:2501.19314*.

Gemma Team. 2025. Gemma 3.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, Pengfei Wang, and Yuanchun Zhou. 2025. Diversity-oriented data augmentation with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22265–22283, Vienna, Austria. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elena Zotova, Rodrigo Agerri, and German Rigau. 2021. Semi-automatic generation of multilingual datasets for stance detection in twitter. *Expert Systems with Applications*, 170:114547.

# A Appendix

## A.1 Language Abbreviations

| Code | Language |
|---|---|
| az | Azerbaijani |
| cy | Welsh |
| de | German |
| en | English |
| he | Hebrew |
| id | Indonesian |
| ro | Romanian |
| sl | Slovenian |
| sw | Swahili |
| te | Telugu |
| th | Thai |

Table 4: Language abbreviations.

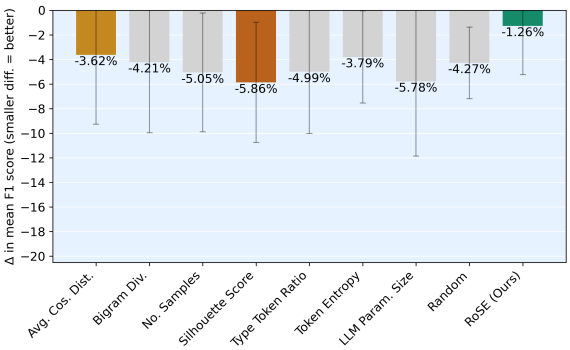



Figure 7: Comparison of proxy metrics for selecting the best LLM generator without considering Llama 3 70B in the comparison. Bars show the average gap in mean F1 score to the optimal human-based evaluation (smaller is better). The best metric is green, the second best is orange, and the worst is red.

## A.2 Computational Resources

All generation experiments were done using vllm (Kwon et al., 2023) for efficient inference. For a single language-model combination, it takes approximately 2 hours to generate the data for intent recognition, 1 hour 20 minutes for topic classification and 23 minutes for sentiment classification. We use a single H100 GPU to generate the samples. Thus, for 11 languages, 6 LLMs and 3 different tasks, it takes around 180 GPU hours to generate all the data.

The fine-tuning experiments were also performed on an H100 GPU. All the finetuning experiments took approximately an additional 200 GPU hours.

## A.3 Downstream Fine-tuning of XLM-R

For the downstream evaluation, we fine-tune the XLM-R (Conneau et al., 2019) *FacebookAI/xlm-roberta-base* model for 50 epochs with a batch size of 16 and employ early stopping with a patience of 5 epochs to prevent overfitting. We perform hyperparameter optimisation to determine the optimal learning rate and set it to 1e-5. *AdamW* is used as an optimiser. We balance the generated datasets to have the same number of samples per label. We normalise all inputs by converting them to lowercase and removing punctuation.

## A.4 Prompt Templates and Generation Details

For generating synthetic data, we used this general prompt template with examples included:

*"Please create 6 different {task_text_type} in the {language} language, separated by the sep_token token. The {task_text_type} should be about the {label}. Note that some examples from the dataset look as follows: Examples: {examples}. Output only the text in {language} and nothing else! Do not number the texts!"*

The *task_text_type* placeholder had values based on tasks, either semantic analysis, intent recognition or topic classification. The *sep_token* was represented as "—". The *label* placeholder was replaced for each task and label with an LLM-generated explanation of that label based on randomly preselected human examples. The *examples* placeholder contained the human examples for that given label being generated. For the RoSE-Z generation, this was excluded, and no examples were given to the LLM generator.

Sampling parameters used for vllm generation were: *temperature=0.7, top_p=0.9, max_tokens=4096, repetition_penalty=1.2*. We collected 10 generated samples per inference run for increased efficiency and performed cleaning of the generated data. We excluded duplicates and collected until 100 unique texts were generated.

Specific versions of LLMs used for generations were: Llama-3-70b-instruct [2], Llama-3-8b-instruct [3], Qwen2.5-14B-Instruct [4], Magistral-

[2] https://huggingface.co/TechxGenus/Meta-Llama-3-70B-Instruct-GPTQ

[3] https://huggingface.co/TechxGenus/Meta-Llama-3-8B-Instruct-GPTQ

[4] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

Small-2509 [5], gemma-3-27b-it [6], gemma-3-4b-it [7].

### A.5 Performance on Varying Number of Candidate LLMs: Breakdown Per Task

We provide a breakdown per task of the selected best-performing proxy metrics when using a varying number of LLMs in comparisons. Our results in Figure 8 show that RoSE remains the best performing metric except for Topic classification, where LLM parameter size's performance can be explained due to Llama-3 70B's impressive data generation capabilities for this task.

### A.6 Cost Effectiveness of RoSE: Breakdown Per Task

We provide a cost-effective breakdown of RoSE per task in Figure 9. Similar to findings in Section 4.6, the performance of RoSE in identifying a good LLM generator increases with more LLMs' synthetic data being used during the computation of RoSE. The only exception is when using 2 randomly chosen LLMs for the intent recognition task, where a sudden drop is present. The combination of 2 randomly chosen LLMs for computing RoSE can lead to significant variability of the computed values by providing proxy performance that can be skewed. Including 3 or more LLMs for computing RoSE is thus essential for the performance of these proxy metrics.

### A.7 Additional Visualisations and Tables

We provide an evaluation of proxy metrics, excluding Llama 3 70B, in Figure 7. We provide additional visualisation for proxy metric performance per task in Figure 10 and per language in Figure 11. We provide the mean F1 scores on human test data for finetuned XLM-R per language-task in Tables 5, 6 and 7. We provide an overview of the performance and costs of each proxy metric used in Table 8.

[5] https://huggingface.co/mistralai/Magistral-Small-2509

[6] https://huggingface.co/google/gemma-3-27b-it

[7] https://huggingface.co/google/gemma-3-4b-it

| LLM | az | cy | he | th | sw | sl | en | de | id | ro | te |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Magistral-Small | 66.90 | 47.19 | 70.62 | 69.78 | 48.48 | 79.53 | 77.24 | 66.20 | 83.26 | 82.06 | 68.88 |
| Meta-Llama-3-70B | 65.98 | 57.95 | 69.81 | 76.24 | 55.63 | 75.77 | 72.97 | 59.58 | 87.82 | 85.79 | 53.32 |
| Meta-Llama-3-8B | 70.01 | 45.28 | 69.98 | 65.15 | 60.48 | 76.92 | 79.26 | 69.36 | 81.63 | 80.23 | 70.26 |
| Qwen2.5-14B | 65.43 | 51.46 | 64.03 | 77.27 | 55.48 | 61.04 | 68.01 | 68.77 | 86.36 | 81.49 | 73.02 |
| gemma-3-27b | 64.44 | 36.72 | 77.17 | 69.48 | 47.69 | 77.65 | 63.41 | 69.36 | 88.71 | 77.04 | 75.61 |
| gemma-3-4b | 69.53 | 52.95 | 73.78 | 62.40 | 46.80 | 77.30 | 64.31 | 63.31 | 89.18 | 76.02 | 75.52 |

Table 5: Mean F1 scores of XLM-R finetuned on data generated from each LLM via RoSE ICL generation setup per language for the sentiment analysis task on human test data. Higher is better.

| LLM | az | cy | he | th | sw | sl | en | de | id | ro | te |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Magistral-Small | 67.89 | 65.40 | 68.83 | 78.70 | 62.26 | 76.24 | 80.24 | 78.80 | 75.85 | 74.33 | 67.45 |
| Meta-Llama-3-70B | 76.56 | 65.47 | 72.86 | 76.86 | 65.86 | 80.33 | 80.15 | 81.07 | 81.06 | 78.38 | 69.97 |
| Meta-Llama-3-8B | 72.14 | 64.68 | 71.40 | 74.62 | 65.24 | 79.27 | 80.39 | 78.88 | 78.19 | 79.67 | 54.39 |
| Qwen2.5-14B | 73.22 | 65.09 | 46.57 | 73.33 | 61.66 | 76.55 | 78.39 | 76.70 | 73.74 | 76.02 | 62.63 |
| gemma-3-27b | 70.05 | 60.83 | 63.42 | 69.95 | 57.52 | 72.17 | 73.23 | 74.11 | 72.22 | 70.14 | 62.62 |
| gemma-3-4b | 66.06 | 52.17 | 65.65 | 71.21 | 58.89 | 69.92 | 70.59 | 74.53 | 71.44 | 71.12 | 62.42 |

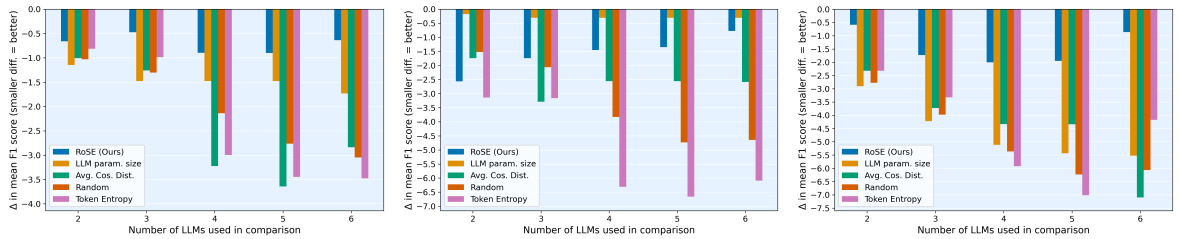
Table 6: Mean F1 scores of XLM-R finetuned on data generated from each LLM via RoSE ICL generation setup per language for the topic classification task on human test data. Higher is better.

| LLM | az | cy | he | th | sw | sl | en | de | id | ro | te |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Magistral-Small | 86.97 | 78.09 | 84.74 | 90.98 | 75.78 | 91.29 | 90.90 | 84.42 | 88.59 | 89.63 | 83.11 |
| Meta-Llama-3-70B | 84.93 | 75.77 | 87.21 | 85.10 | 73.64 | 91.12 | 92.58 | 85.99 | 90.82 | 89.58 | 84.07 |
| Meta-Llama-3-8B | 84.00 | 76.43 | 84.00 | 90.39 | 73.27 | 91.37 | 92.71 | 87.31 | 92.89 | 88.78 | 82.85 |
| Qwen2.5-14B | 85.01 | 65.29 | 82.16 | 87.90 | 65.31 | 90.09 | 91.05 | 89.14 | 91.93 | 90.61 | 82.09 |
| gemma-3-27b | 82.82 | 65.70 | 83.63 | 86.86 | 73.26 | 84.67 | 80.31 | 87.13 | 90.28 | 87.20 | 81.41 |
| gemma-3-4b | 82.14 | 67.23 | 85.13 | 87.29 | 71.05 | 88.03 | 91.15 | 80.27 | 92.43 | 86.92 | 80.44 |

Table 7: Mean F1 scores of XLM-R finetuned on data generated from each LLM via RoSE ICL generation setup per language for the intent recognition task on human test data. Higher is better.

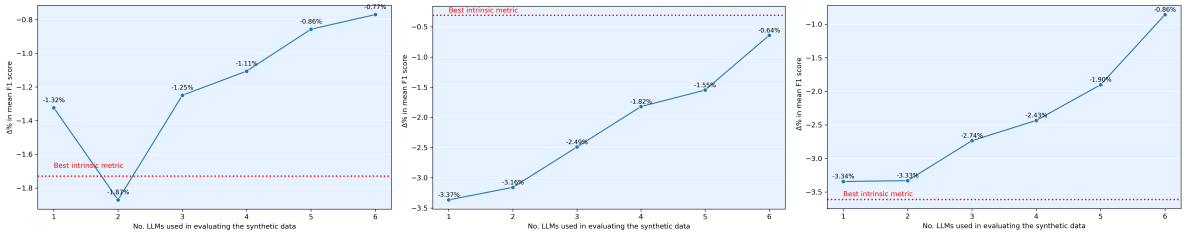
| Proxy Metric | Computational Cost | Performance |
|---|---|---|
| Type–Token Ratio (TTR) | Low | Weak, low/unstable correlations with human performance |
| Average Cosine Distance | Medium | Moderate, sometimes effective but inconsistent |
| Bigram Diversity | Low | Weak, small or negative correlations |
| Number of Valid Samples | Low–Medium | Weak to moderate, unstable across tasks/languages |
| Silhouette Score | Medium | Weak, often a negative correlation with human data |
| Token Entropy | Low | Weak, small and inconsistent effects |
| LLM Parameter Size | Very Low | Second-best, strong for some tasks (topic), poor for others (sentiment) |
| Random Selection | Negligible | Baseline, consistently poor |
| RoSE | Highest | Best overall; smallest F1 gap to optimal, only metric with positive correlations |

Table 8: Comparison of proxy metrics for selecting the best LLM generator. Costs are relative; performance is summarised based on overall results across languages and tasks.



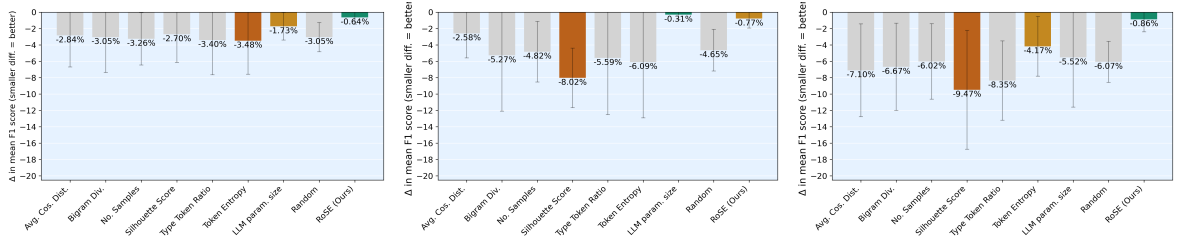

(a) Intent recognition. (b) Topic classification. (c) Sentiment analysis.

Figure 8: Comparison per-task of a selection of proxy metrics for selecting the best LLM generator when comparing various combinations of LLMs.



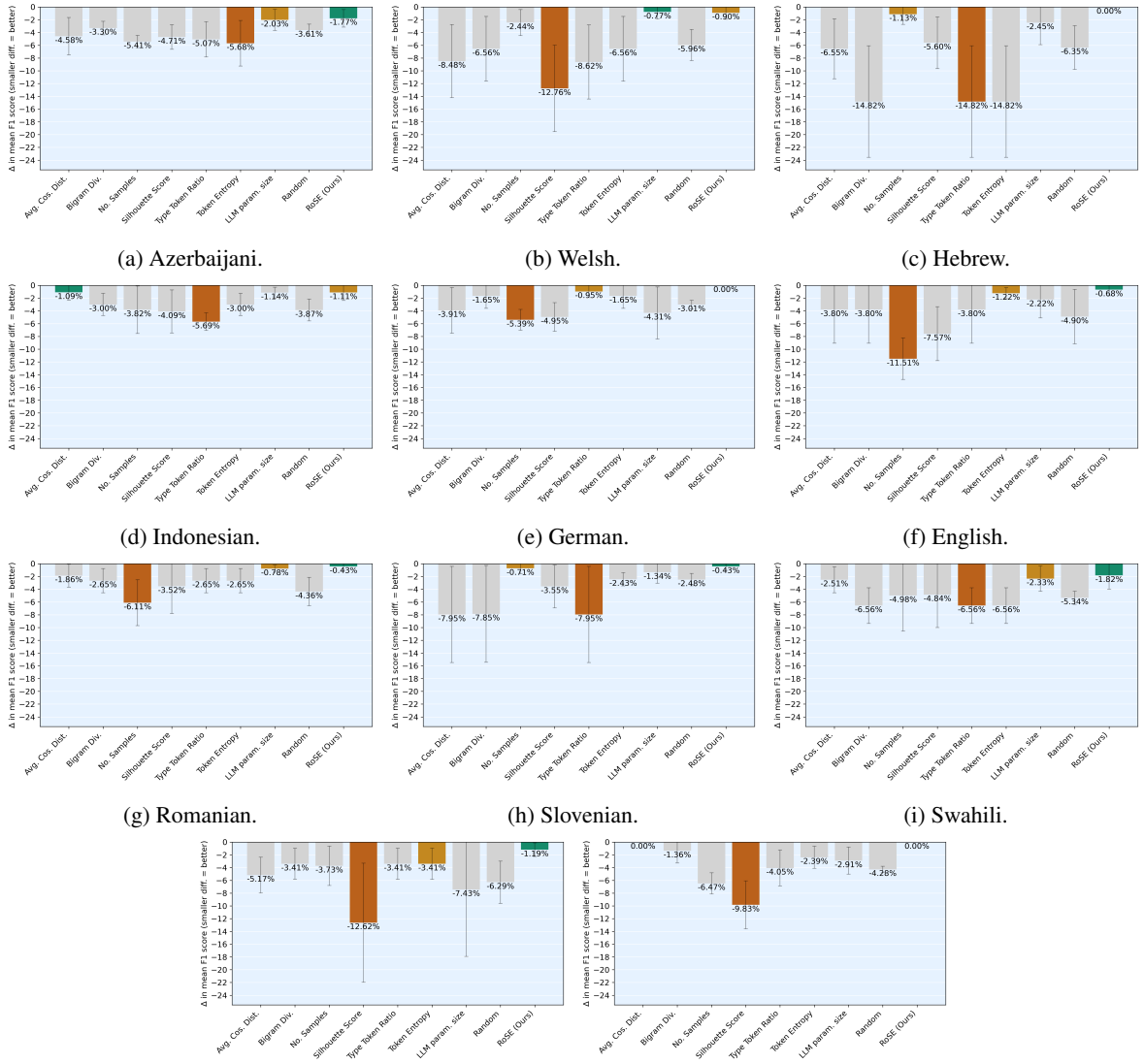

(a) Intent recognition. (b) Topic classification. (c) Sentiment analysis.

Figure 9: Number of randomly chosen LLMs used for computing RoSE and its effect on F1 score difference to optimal LLM generator selection. RoSE benefits from more LLMs being used during the evaluation of the downstream classifier during its computation. For all tasks, the performance increases from 3 or more LLMs used for computing RoSE.




(a) Intent recognition. (b) Topic classification. (c) Sentiment analysis.

Figure 10: Comparison of proxy metrics for selecting the best LLM generator per task. Bars show the average gap in mean F1 score for models trained on the best generator selected by metrics vs. the optimal generator (smaller is better). The best metric is green, the second best is orange, and the worst is red. RoSE performs well across all tasks as either the best or the second-best proxy metric.




(a) Azerbaijani. (b) Welsh. (c) Hebrew.

(d) Indonesian. (e) German. (f) English.

(g) Romanian. (h) Slovenian. (i) Swahili.

(j) Telugu. (k) Thai.

Figure 11: Comparison of proxy metrics for selecting the best LLM generator per language. Bars show the average gap in mean F1 score for models trained on the best generator selected by metrics vs. the optimal generator (smaller is better). The best metric is green, the second best is orange, and the worst is red. RoSE is the best proxy metric in 9 of 11 languages, and 2nd best for the other 2 cases. RoSE performs well across all languages regardless of the amount of resources per language.