Generating Useful Synthetic Text with Local LLMs and Data Valuation: A Study on Spanish Tourism Reviews

Carlos Minutti-Martinez[1,*], Boris Escalante-Ramirez[2] and Jimena Olveres[2]

[1]INFOTEC, Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Aguascalientes, Mexico
[2]Departamento de Procesamiento de Señales, División de Ingeniería Eléctrica, Universidad Nacional Autónoma de México, Mexico City, Mexico

Abstract

The 2026 edition of the REST-Mex challenge (Research on Synthetic Text Modeling for Mexican Magical Towns) introduces a novel shared task: generating a corpus of 3,000 synthetic Spanish-language tourism reviews that, when used to train a sentiment classifier, generalizes effectively to real user opinions.

We present the system developed by team AxoloTux. We explore two complementary generation strategies: few-shot prompting with locally deployed instruction-tuned Qwen3.5 models (9B and 27B parameters) and LoRA-based fine-tuning of `meta-llama/Llama-3.2-3B` on balanced review subsets, followed by a three-stage quality filtering pipeline that assesses label consistency, linguistic realism, and estimated predictive utility via KNN-Shapley data valuation.

Our best submission achieved a track score of 0.511 on the official evaluation, an 8.3% improvement over a real-data baseline of equal corpus size, placing our team second among competing teams and within 2% of the top score of 0.522, while ranking first in macro-averaged precision across all submissions. Beyond the competition results, this work was designed and tested as a privacy-preserving pipeline: all generation, scoring, and filtering steps run on local hardware without transmitting any real review to an external service. This makes the approach directly applicable to domains such as clinical NLP, legal document processing, or mental health support, where similar classification problems arise and data cannot be shared openly.

Keywords

Synthetic Data Generation, Privacy-Preserving NLP, Large Language Models, Spanish NLP, LLaMA, Qwen, Data Valuation, KNN-Shapley, Tourism Reviews, Text Generation

1. Introduction

Natural language processing has become indispensable across a broad spectrum of domains, from e-commerce and tourism to healthcare, legal services, and mental health support. Yet many of the most valuable textual datasets (clinical encounter notes, therapist-patient dialogues, legal proceedings, and financial disclosures) cannot be freely shared due to stringent privacy requirements. Even when data subjects provide informed consent, the risk of re-identification through future analytical methods, adversarial inference, or cross-dataset linkage remains a documented concern [1]. This creates a fundamental bottleneck for NLP researchers and practitioners, who increasingly need large annotated corpora to train competitive models.

The challenge is particularly acute in healthcare. Electronic health records and clinical narratives are among the richest sources of linguistic signal for tasks such as medical sentiment analysis, symptom extraction, or therapy outcome prediction. Despite ongoing de-identification efforts [2], access to these datasets is restricted even within research consortia, slowing scientific progress in exactly the domains where accurate models could have the greatest social impact. Similar constraints apply to attorney-client

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
carlos.minutti@infotec.mx (C. Minutti-Martinez); boris@lapi.unam.mx (B. Escalante-Ramirez); jolveres@lapi.unam.mx (J. Olveres)
https://cminuttim.github.io (C. Minutti-Martinez); https://lapi.fi-p.unam.mx/ (B. Escalante-Ramirez); https://lapi.fi-p.unam.mx/ (J. Olveres)
0000-0002-4002-0773 (C. Minutti-Martinez); 0000-0003-4936-8714 (B. Escalante-Ramirez); 0000-0002-1514-4520 (J. Olveres)

communications in legal NLP, tutoring transcripts in educational technology, and session notes in digital mental health platforms.

Synthetic data generation has emerged as a promising strategy to break this deadlock. The core premise is that a model trained on synthetically generated text should generalize as well, or nearly as well, as one trained on real data, provided the synthetic corpus faithfully captures the distributional properties that matter for the target task. If this holds, synthetic datasets can be openly shared and reused without exposing any real individual's information. Earlier approaches based on rule-based templates, variational autoencoders, and back-translation offered limited fluency and controllability [3, 4]. The advent of large instruction-tuned generative models has dramatically raised the ceiling of what is achievable: modern LLMs can produce contextually coherent, stylistically varied text across diverse domains with minimal task-specific engineering [5, 6].

However, generating fluent text is not equivalent to generating *useful* text. A synthetic corpus that reads naturally but does not reflect the real data's decision boundaries will produce models that fail to generalize. Measuring and maximizing *predictive utility*, meaning the extent to which synthetic examples contribute to training a classifier that performs well on real data, remains an open and largely unsolved problem. Data valuation methods, such as KNN-Shapley values [7], offer a principled framework for estimating each synthetic example's marginal contribution to downstream model accuracy, providing a data-driven basis for quality filtering.

An equally important but often overlooked concern is *where* generation happens. If synthetic data is produced by querying commercial cloud APIs using real examples as context (a common few-shot prompting pattern), the sensitive data must be transmitted to a third-party server. In privacy-critical settings this undermines the entire rationale. For this reason, our methodology is designed from the outset around *local inference*: all generation, scoring, and filtering steps run on private hardware, ensuring that no real review, patient record, or sensitive document leaves the organization's infrastructure.

The 2026 edition of the REST-Mex challenge (*Research on Synthetic Text Modeling for Mexican Magical Towns*) [8], part of IberLEF 2026 [9], provides a natural, controlled testbed for these questions. Departing from the traditional classification framing of prior editions [10], the 2026 task asks participants to generate a corpus of 3,000 original Spanish-language tourism reviews about Mexican *Pueblos Mágicos*. The evaluation is indirect: organizers train a polarity classifier exclusively on each team's submitted synthetic corpus and measure its performance on a hidden test set of real TripAdvisor reviews using a class-frequency-weighted macro-F1 that emphasizes minority polarity classes. Success therefore requires generating synthetic data that captures the distributional properties necessary for genuine generalization to real user opinions, precisely the property we care about in the broader privacy-sensitive data problem.

Although the application domain is tourism, we deliberately designed our system to address the more general problem. Tourism reviews are publicly available and allow rigorous ground-truth evaluation; they serve as a tractable proxy for the harder settings (clinical notes, legal records, therapy transcripts) where the same pipeline would be most valuable but where direct evaluation is difficult. The techniques we explore (local LLM generation with controllable metadata conditioning, label-consistency filtering, realism scoring, and KNN-Shapley utility estimation) are domain-agnostic and could be applied to any setting where one needs a privacy-safe synthetic substitute for real annotated text.

In this work we describe the system developed by team AxoloTux, which participated in REST-Mex 2026 using this perspective. Our main contributions are as follows. First, we explore two complementary **local-first generation** strategies using exclusively locally deployed models: few-shot prompting with instruction-tuned Qwen3.5 models (9B and 27B parameters) [11] and LoRA-based fine-tuning [12] of `meta-llama/Llama-3.2-3B` [13] on balanced review subsets, ensuring that no real data is transmitted to external services. Second, we propose a **multi-stage quality filtering** pipeline that screens synthetic candidates by (1) label consistency (agreement between two independent polarity classifiers), (2) linguistic realism (probability that the text is indistinguishable from a real review, as estimated by a real/synthetic discriminator), and (3) predictive utility via KNN-Shapley data valuation [7]. Third, we provide an **empirical analysis** of the trade-offs between generation strategy (few-shot

prompting vs. fine-tuning), polarity balancing, and quality filtering thresholds in terms of downstream classifier performance.

Our best submission, based on few-shot generation with a locally deployed 27B Qwen3.5 model followed by quality filtering, achieved a track score of 0.511 on the official evaluation, an 8.3% improvement over a real-data baseline of equal corpus size, and within 2% of the top score of 0.522. Team AxoloTux placed second among competing teams and ranked first in macro-averaged precision across all submissions.

The remainder of this paper is organized as follows. Section 2 describes the dataset, generation strategies, and filtering pipeline. Section 3 presents the results. Section 4 discusses the key findings and their implications. Section 5 concludes.

2. Methodology

2.1. Dataset

The REST-Mex 2026 training corpus consists of 208,051 TripAdvisor reviews of Mexican *Pueblos Mágicos*. Each record includes six fields: the review title, the full review text, a polarity score from 1 (Very Bad) to 5 (Very Good), the destination town, the Mexican state (*Region*), and the establishment type (Restaurant, Hotel, or Attractive). Encoding inconsistencies were corrected using the `ftfy` and `unicodedata` Python libraries to standardize all text to UTF-8.

2.1.1. Dataset Statistics

The polarity distribution is strongly skewed: 65.6% of reviews carry the highest sentiment score (class 5, Very Good), while the two negative classes (Very Bad and Bad) together account for only 5.3% of the data, reflecting the natural positivity bias of voluntary travel reviews. Reviews span three establishment types: restaurants (41.7%), tourist attractions (33.6%), and hotels (24.7%). Figure 1 shows that attractions and restaurants concentrate the highest proportions of very positive reviews, while hotels show a somewhat broader sentiment range. Coverage across the 40 destinations is highly concentrated: Tulum alone accounts for 21.8% of all reviews, and the top five towns contribute nearly half the corpus. Figure 2 shows the geographic distribution across Mexican states.

Review lengths, measured in subword tokens using the `xlm-roberta-base` tokenizer, follow a right-skewed distribution (Table 1): the median review is 64 tokens, and 97.9% of reviews fit within 256 tokens. Only 0.26% exceed 512 tokens, so truncation does not affect the vast majority of the corpus. Figure 3 shows the token count histogram.

Table 1
Summary statistics for the number of tokens per review (title + review, `xlm-roberta-base` tokenizer).

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Tokens	7	44	64	86.8	108	1,999

2.2. Evaluation Metric

In the REST-Mex 2026 generation task, each team submits a corpus of 3,000 synthetic reviews. The organizers train a polarity classifier exclusively on the submitted corpus and evaluate it on a hidden test set of real reviews. Performance is measured by a class-frequency-weighted macro-F1 score [8]:

$$\text{TrackScore}(k) = \frac{\sum_{i \in C} \left(1 - \frac{TC_i}{TC}\right) F_i(k)}{\sum_{i \in C} \left(1 - \frac{TC_i}{TC}\right)}, \quad C = \{1, 2, 3, 4, 5\} \tag{1}$$

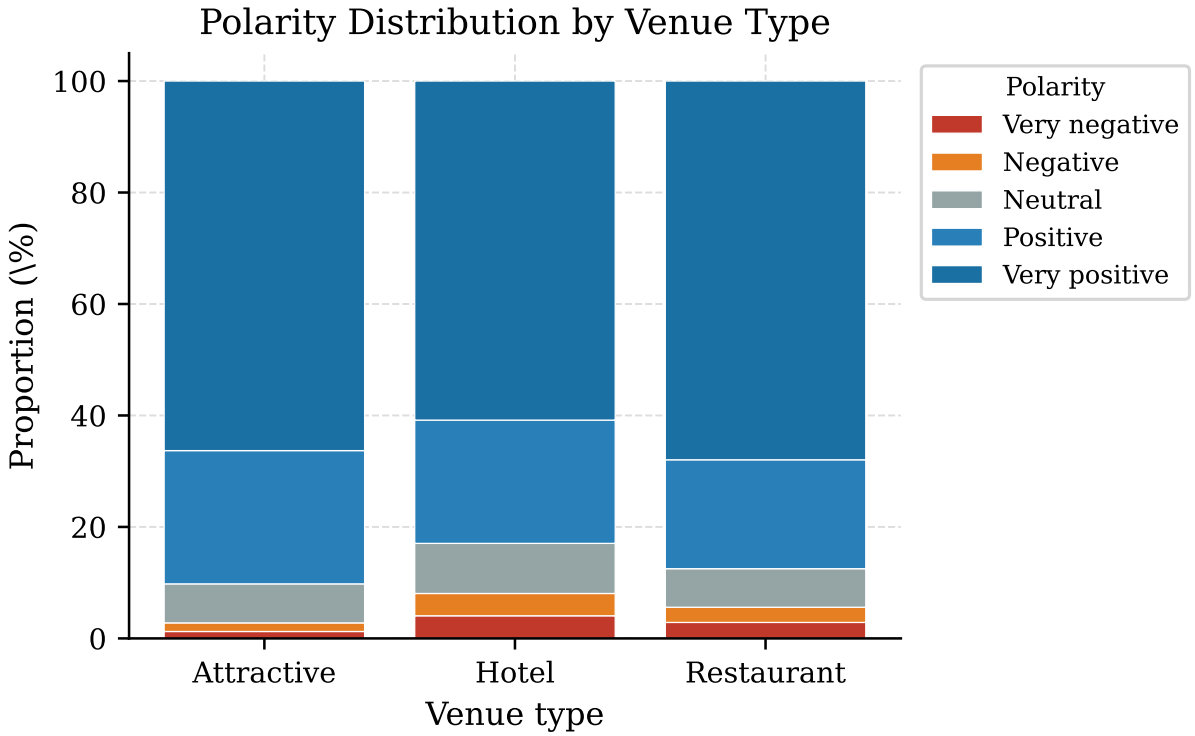

Figure 1: Polarity distribution by establishment type. Attractions and restaurants concentrate the highest proportion of very positive reviews; hotels show a broader sentiment range.

where TC_i is the number of test instances belonging to class i, TC is the total number of test instances, and $F_i(k)$ is the F1 score of model k on class i. The weight $(1 - TC_i/TC)$ is inversely proportional to class frequency, so minority polarity classes (Very Bad, Bad) receive near-unit weight while the dominant class (Very Good) receives the lowest weight. A higher score indicates that the synthetic corpus transfers its polarity signal more faithfully to the real data distribution, with particular emphasis on the hardest-to-predict minority classes.

2.3. Generation Strategies

Our approach explores two complementary local generation strategies. All generation, scoring, and filtering steps run on private hardware; no real review is transmitted to any external service. Each strategy produces 30,000 candidate reviews, for a total pool of up to 90,000 candidates in the mixed submissions.

2.4. Strategy 1: LoRA Fine-Tuning of Llama-3.2-3B

Following the approach of [14], we fine-tune `meta-llama/Llama-3.2-3B` [13] to generate tourism reviews conditioned on structured metadata (Polarity, Town, Region, Type).

2.4.1. Balanced Training Data

The strong class imbalance of the original dataset would bias a fine-tuned generator toward very positive reviews. We therefore construct a balanced training subset by using the median class count as the target per class (15,501 reviews). Classes larger than the target are randomly under-sampled; smaller classes are oversampled with replacement. Table 2 shows the resulting balanced distribution of 77,505 reviews.

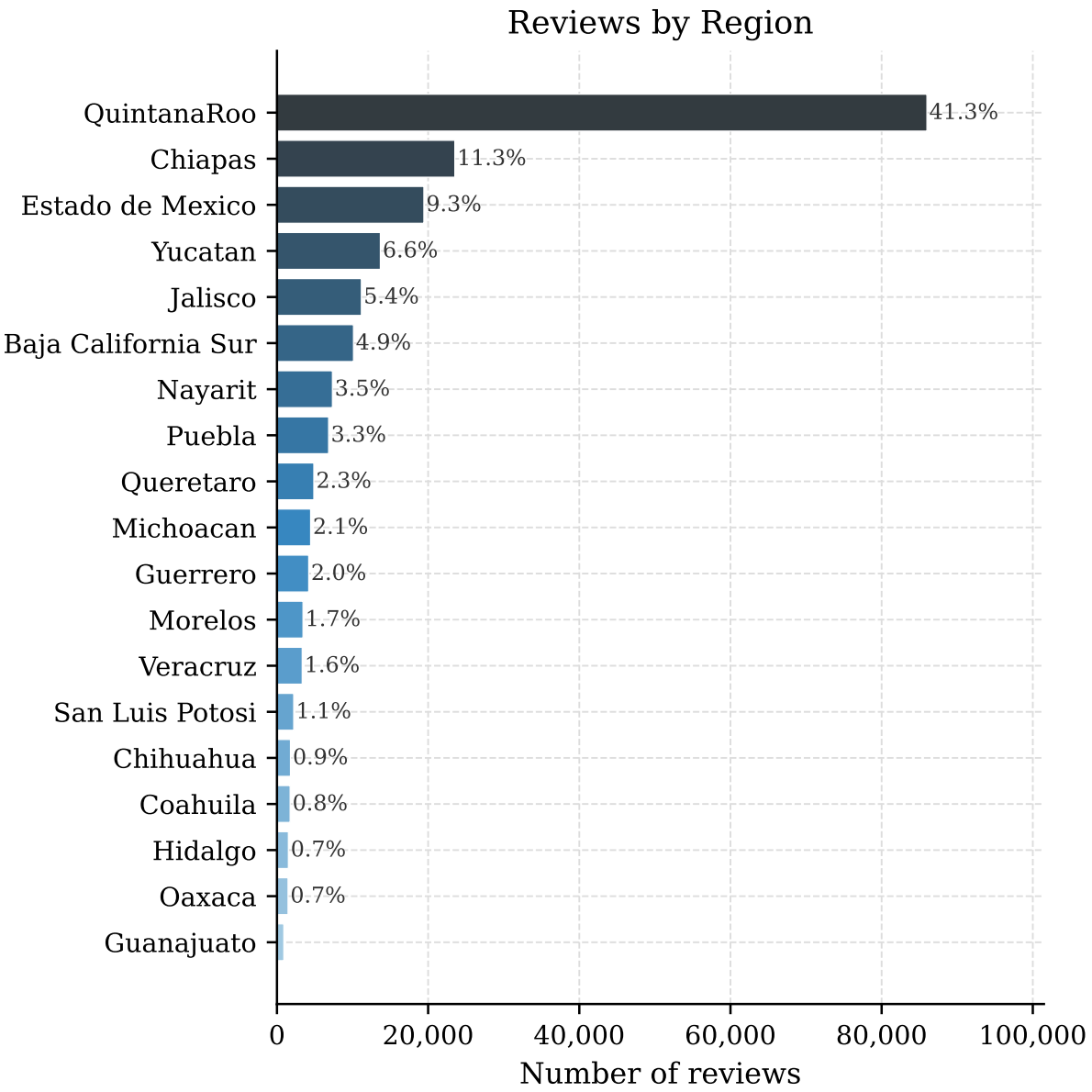

Figure 2: Number of reviews per Mexican state (*Region*). Coverage is concentrated in the Yucatán Peninsula (Quintana Roo leading with Tulum and Isla Mujeres) and a handful of other popular tourist states.

Table 2
Class distribution before and after balancing for the Llama-3.2-3B fine-tuning. Original counts are computed after removing duplicate records (207,872 unique reviews out of 208,051).

Polarity	Original count	Balanced count
1 (Very Bad)	5,438	15,501
2 (Bad)	5,492	15,501
3 (Neutral)	15,501	15,501
4 (Good)	44,982	15,501
5 (Very Good)	136,459	15,501
Total	**207,872**	**77,505**

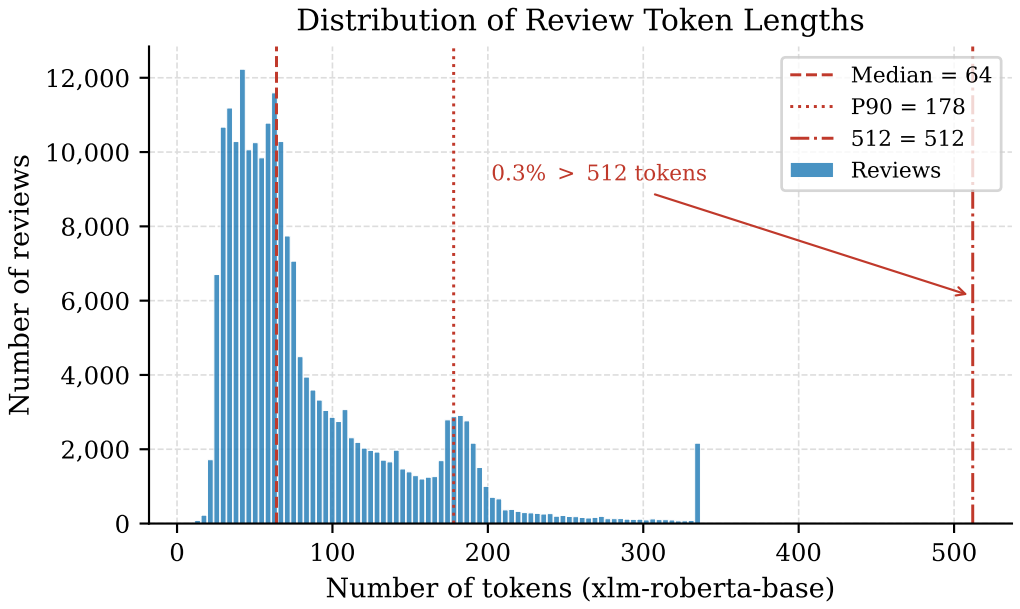

Figure 3: Distribution of token counts per review. The distribution is right-skewed; 97.9% of reviews contain at most 256 tokens.

2.4.2. Prompt Format and Completion Masking

Each training example is formatted as an instruction-following prompt providing the four conditioning attributes, followed by the target completion:

```
### Instruccion:
Genera una resena turistica realista en espanol para el
siguiente lugar.
Polaridad: {polarity}  Ciudad: {town}
Region: {region}       Tipo: {type}

### Respuesta:
Title: {title}
Review: {review}<|end_of_text|>
```

The model is trained only on the completion tokens (Title and Review); prompt tokens are masked with -100 so the loss is not back-propagated through the conditioning context. This ensures the model learns to generate reviews rather than to predict conditioning attributes.

2.4.3. Fine-Tuning Configuration

The base model is loaded in 4-bit NF4 precision using `BitsAndBytes` and fine-tuned with LoRA [12] via the PEFT library. Adapters are applied to all attention and feed-forward projection matrices: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. Table 3 summarizes the training hyperparameters.

After fine-tuning, 30,000 reviews are generated in batches of 64. Conditions (Polarity, Town, Region, Type) are sampled with replacement from the balanced training set, yielding approximately uniform coverage across the five polarity classes. Decoding uses temperature 0.85, top-p 0.9, and a repetition penalty of 1.1, with a maximum of 300 new tokens per review.

2.5. Strategy 2: Few-Shot Prompting with Qwen3.5

We deploy two sizes of the Qwen3.5 model [11] locally via Ollama (9B and 27B parameters) using abliterated variants [15, 16] to prevent the model from refusing to generate text with the informal writing

Table 3
Llama-3.2-3B generation fine-tuning hyperparameters.

Parameter	Value
Epochs	3
Per-device batch size	4
Gradient accumulation steps	4 (effective batch: 16)
Learning rate	2×10^{-4}
Weight decay	0.01
Max sequence length	512 tokens
Quantization	NF4 (4-bit)
LoRA rank (r)	16
LoRA α	32
LoRA dropout	0.05

patterns typical of real user-generated Spanish content (missing accents, inconsistent capitalization, informal connectors). No real review text is transmitted to any external service.

2.5.1. Sampling Strategy

For each generation call, a group of three real reviews is sampled to serve as in-context examples. A seed row is first drawn at random from the training data; two additional reviews are then retrieved by matching on (Polarity, Town, Region, Type). If the matched pool contains fewer than ten records, the constraint is relaxed progressively: first to (Polarity, Region, Type), then to (Polarity, Type) only. The three retrieved reviews are passed together to the model.

2.5.2. System Prompt

The system prompt instructs the model to synthesize a new review that blends details from all three reference reviews equally, without copying any single one, while preserving their shared sentiment class, establishment type, and geographic attributes. Crucially, the prompt also instructs the model to replicate the informal writing conventions found in the reference texts: missing accent marks, sparse punctuation, capitalized emphasis words, and informal discourse markers (e.g., *"y pues"*, *"la verdad"*). The model is asked to return a structured output with Title, Review, Polarity, Town, Region, and Type fields.

2.5.3. Polarity Balancing

By default, the seed row is drawn uniformly at random from the full training data, so generated polarities reflect the natural distribution (predominantly class 5). When the `--balanced-polarity` flag is enabled, the target polarity is sampled uniformly from $\{1, 2, 3, 4, 5\}$ before retrieving the seed row, producing approximately balanced outputs. Both Qwen3.5 submissions use the natural (unbalanced) mode, as the filtering pipeline can enforce polarity balance at selection time.

2.6. Quality Filtering Pipeline

After generation, each candidate review is scored along three quality dimensions and filtered sequentially. Only reviews passing all three stages are retained for the final corpus assembly.

2.6.1. Polarity Classifiers

Two independent polarity classifiers are used to assess label consistency.

Bag-of-words (BOW) classifier. A TF-IDF vectorizer with up to 30,000 unigram and bigram features (sublinear TF scaling, minimum document frequency of 5) is combined with a balanced-weight multinomial logistic regression model trained on the full real training corpus. Reviews are preprocessed by lowercasing, accent normalization, and domain-specific stop-word removal before vectorization. The BOW model provides a fast, interpretable polarity signal that is entirely independent of the generation backbone.

LLaMA-based classifier. A two-head MLP classifier is trained on top of the frozen Llama-3.2-3B + LoRA backbone from Section 2.4. The last hidden state of the backbone (extracted with left-padded inputs so the final real token is always at position -1) is projected through a shared Linear + GELU + Dropout layer ($d_{\text{proj}} = 512$) and then split into: (i) a five-way polarity head; and (ii) a binary real/synthetic discriminator head. The combined training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{pol,real}} + \lambda \cdot \mathcal{L}_{\text{pol,synth}} + \mathcal{L}_{\text{RS}} \tag{2}$$

where $\mathcal{L}_{\text{pol,real}}$ and $\mathcal{L}_{\text{pol,synth}}$ are cross-entropy losses on real and synthetic samples respectively ($\lambda = 0.2$), and $\mathcal{L}_{\text{RS}}$ is the binary cross-entropy on the real/synthetic label. Training data consists of all real reviews merged with all available synthetic candidates, split 90/10. Inverse-frequency class weights are applied to the polarity loss to compensate for the natural class imbalance.

2.6.2. Quality Metrics

Three scalar metrics are derived for each candidate review.

Consistency measures the agreement between the original polarity label assigned during generation and the two independent classifier predictions:

$$\text{consistency} = 1 - \frac{|\text{Polarity} - \hat{p}_{\text{LLM}}| + |\text{Polarity} - \hat{p}_{\text{BOW}}|}{8} \tag{3}$$

The denominator 8 is the maximum possible total error (two classifiers $\times$ maximum per-classifier error of 4 on the 1–5 scale), so consistency $\in [0, 1]$, with 1 indicating perfect agreement among all three polarity sources.

Realism is the estimated probability that the review is indistinguishable from a real one, as produced by the binary discriminator head:

$$\text{realistic} = 1 - \hat{p}_{\text{synthetic}} \tag{4}$$

Predictive utility is estimated via KNN-Shapley data valuation [7]. Review texts are embedded with `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2` [17]; cosine similarities are computed via normalized dot products. Algorithm 1 of Jia et al. [7] then analytically computes each synthetic review's Shapley value, defined as its expected marginal contribution to a k-NN classifier (k=5) evaluated on a balanced set of real reviews, without any model retraining. Scores are normalized by the maximum absolute value across the candidate set.

KNN-Shapley is particularly well suited here: it is deterministic, requires no model retraining per candidate, and directly quantifies how much each synthetic example is expected to improve KNN accuracy on real data, precisely the property we want to maximize.

2.6.3. Sequential Filtering and Submitted Systems

Candidates are filtered in three sequential stages:

1. **Consistency filter:** retain only reviews with consistency $\geq C$.

2. **Realism filter:** from the survivors, retain only reviews with realistic $\geq R$.
3. **Utility selection:** from the remaining reviews, select the final 3,000 either by taking the top-600 per polarity class (*n-per-class* mode) or by taking the global top-3,000 by utility score (*n-total* mode).

The per-class selection mode addresses a systematic bias: higher-polarity reviews (classes 4 and 5) tend to score higher on both consistency and realism because they are more fluent and unambiguous. Without a per-class constraint, the selected corpus would be dominated by positive reviews, reducing its discriminative value for training a classifier.

Table 4 summarizes the six submitted systems and their filtering configurations. The Llama submissions use stricter thresholds because the balanced training data produces more consistent outputs; the Qwen submissions use a lower realism threshold since the abliterated, instruction-tuned models tend to produce more structurally "clean" text that the real/synthetic discriminator detects more easily. The two mix submissions combine candidates from all three generation sources (90,000 reviews total) and apply per-class selection to enforce polarity balance.

Table 4
Submitted systems and their filtering configurations. C: consistency threshold; R: realism threshold. *n-per-class*: selects top-600 per polarity class (balanced); *n-total*: selects the global top-3,000 by utility.

Submission	Source	Selection	C	R
`Axolotux_qwen_9b`	Qwen3.5 9B	n-total	0.70	0.05
`Axolotux_qwen_27b`	Qwen3.5 27B	n-total	0.70	0.05
`Axolotux_llama_1`	Llama-3.2-3B (LoRA)	n-per-class	0.85	0.35
`Axolotux_llama_2`	Llama-3.2-3B (LoRA)	n-total	0.85	0.50
`Axolotux_mix`	Qwen 9B + 27B + Llama	n-per-class	0.85	0.35
`Axolotux_mix_2`	Qwen 9B + 27B + Llama	n-per-class	0.70	0.50

3. Results

3.1. Overview

The REST-Mex 2026 evaluation report 59 submissions from multiple competing teams (excluding post-hoc combination entries and including two baselines entries) [8, 9]. The official ranking metric is the class-frequency-weighted macro-F1 (Track Score), which assigns higher weight to minority polarity classes as described in Section 2. Team AxoloTux submitted six systems covering two generation strategies and a range of filtering configurations; their results are reported in Table 5 alongside the real-data baseline and the top-ranked submission for reference. Figure 4 shows the empirical Cumulative Distribution Function (CDF) of Track Scores across all 59 submissions.

Table 5
Track Score, per-class F1, and macro-averaged precision for the real-data baseline, all AxoloTux submissions, and the top-ranked system. Best AxoloTux value in each column is shown in bold.

System	Track Score	F1(1)	F1(2)	F1(3)	F1(4)	F1(5)	Avg Prec.
Baseline (real data)	0.4721	0.5571	0.3794	0.4296	0.3636	0.8544	0.5405
AxoloTux Qwen3.5-27B	**0.5114**	**0.5989**	**0.3916**	0.4738	0.4750	0.7862	**0.5975**
AxoloTux Qwen3.5-9B	0.4938	0.5704	0.3698	**0.4746**	0.4747	0.7230	0.5879
AxoloTux Mix	0.4865	0.5630	0.3547	0.4356	**0.4839**	**0.7857**	0.5955
AxoloTux LLaMA (1)	0.4693	0.5517	0.3502	0.4024	0.4584	0.7781	0.5893
AxoloTux Mix-2	0.4694	0.5444	0.3429	0.4096	0.4688	0.7775	0.5903
AxoloTux LLaMA (2)	0.4383	0.5048	0.3223	0.3686	0.4339	0.7749	0.5855

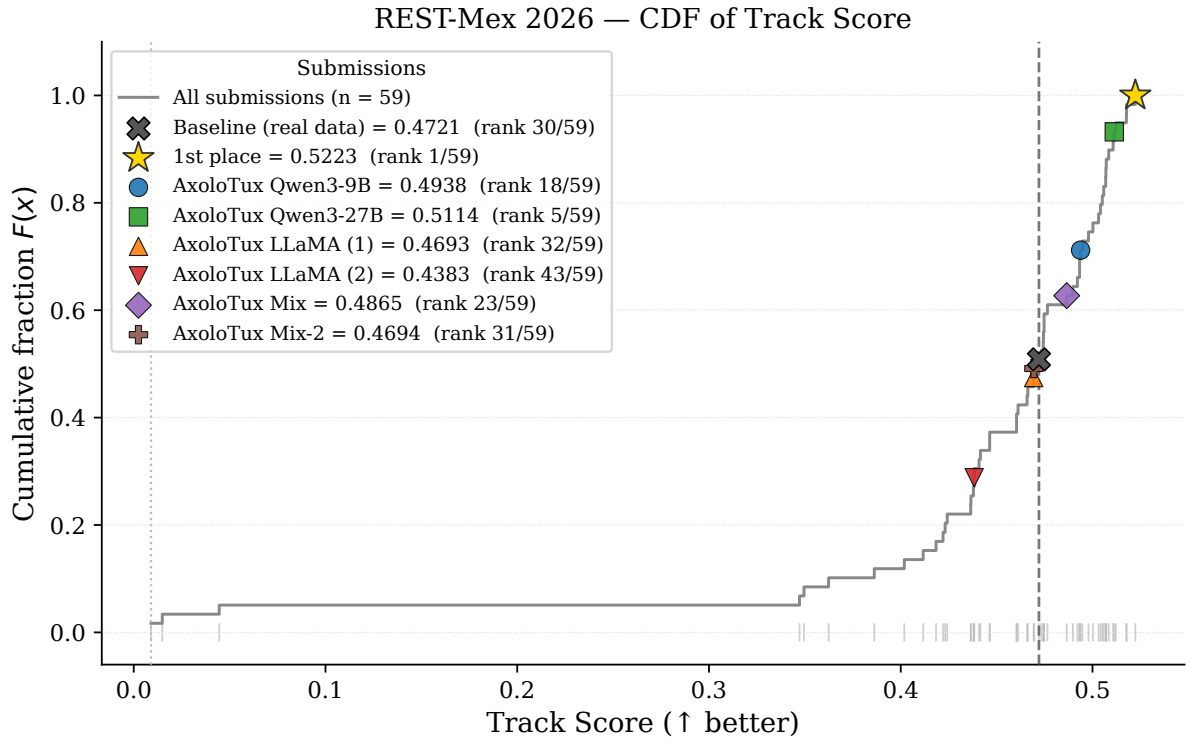

Figure 4: Empirical CDF of Track Score across all 59 REST-Mex 2026 submissions. The best AxoloTux submission (Qwen3.5-27B, rank 5) and the real-data baseline (rank 30) are annotated on the curve.

3.2. Comparison with the Real-Data Baseline

The real-data baseline trains a polarity classifier on 3,000 randomly sampled real reviews (the same corpus size as the task requirement) and achieves a Track Score of 0.4721. This establishes the practical bar: a well-formed synthetic corpus can outperform randomly sampled real data of equal size.

Five of our six submissions clear this bar. Our best submission, AxoloTux Qwen3.5-27B, achieves a Track Score of 0.5114, an absolute improvement of $+0.0393$ ($+8.3\%$ relative) over the baseline, and ranks 5th overall among all 59 submissions, placing team AxoloTux second among competing teams. Our worst-performing submission, AxoloTux LLaMA (2), falls below the baseline at 0.4383, the only submission that failed to improve upon random real-data sampling.

Beyond Track Score, all six AxoloTux submissions achieve higher macro-averaged precision than the baseline (0.5405), with Qwen3.5-27B reaching 0.5975, the highest average precision among all submissions in the competition, surpassing even the top-ranked system (0.5811). Figure 5 shows the CDF of macro-averaged precision across all submissions, illustrating that our best submission sits at the top of the distribution.

3.3. Per-Class Analysis

Figure 6 shows the per-class F1 scores for the baseline, our best submission (Qwen3.5-27B). The improvement over the baseline is concentrated in the intermediate and difficult classes: class 3 (Neutral, $+0.044$) and class 4 (Good, $+0.111$) show the largest gains, while class 1 (Very Bad, $+0.042$) also improves. Class 2 (Bad) sees only a modest gain ($+0.012$). Class 5 (Very Good) decreases from 0.854 to 0.786 (-0.068), reflecting the inherent trade-off: the real training corpus is dominated by class 5 (65.6% of reviews), so the baseline classifier is highly tuned to that class; our quality filtering pipeline selects for predictive utility across all classes, naturally shifting accuracy toward the underrepresented ones.

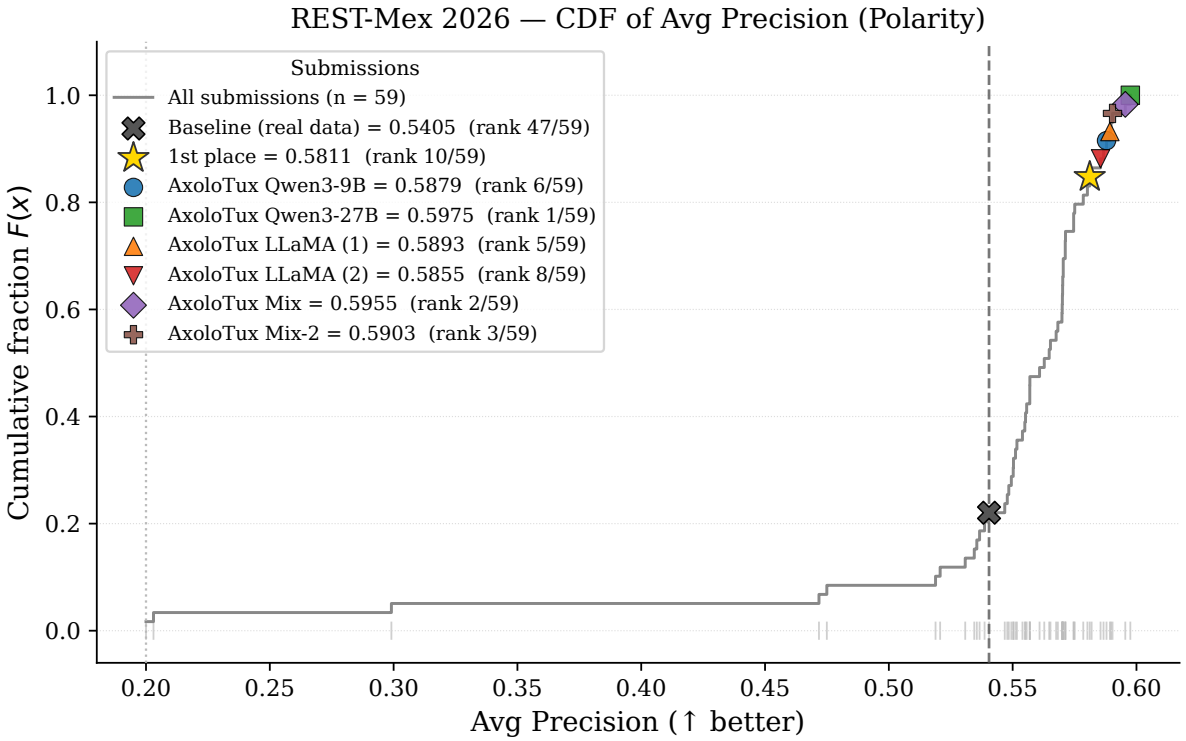

Figure 5: CDF of macro-averaged precision across all 59 REST-Mex 2026 submissions. AxoloTux Qwen3.5-27B achieves the highest average precision in the competition, surpassing the top-ranked system in Track Score.

3.4. Effect of Generation Strategy and Filtering

Among AxoloTux submissions, three patterns are clear. First, **model size is the dominant factor**: Qwen3.5-27B (Track Score 0.5114) outperforms Qwen3.5-9B (0.4938) by 0.018 points despite identical filtering configurations, confirming that larger models produce qualitatively superior few-shot outputs. Second, **few-shot prompting with large models outperforms LoRA fine-tuning**: both LLaMA submissions score below 0.47, while both Qwen submissions score above 0.49. The fine-tuned model's advantage of balanced polarity generation does not compensate for the reduced linguistic quality compared to the larger Qwen3.5 models. Third, **mixing generation sources does not improve over the best single source**: the Mix submission (0.4865) falls between the Qwen3.5-9B and Qwen3.5-27B results, suggesting that the lower-quality LLaMA candidates dilute the pool even when strict filtering is applied.

4. Discussion

4.1. Synthetic Data Can Outperform Real Data of Equal Size

The most striking result is that our best synthetic corpus outperforms a corpus of 3,000 randomly sampled real reviews by 8.3% in Track Score. This is not self-evident: random real reviews carry genuine linguistic diversity and correct polarity labels, so an 8% gain requires the synthetic corpus to capture something the random real sample does not. We attribute this advantage to the interplay of two design choices. First, few-shot prompting with a large model (Qwen3.5-27B) produces reviews that are internally consistent in their polarity signal, as the content, vocabulary, and sentiment markers all align with the assigned label by construction, whereas real reviews frequently contain label noise [10]. Second, the KNN-Shapley utility selection step actively screens for examples whose addition to the training set improves k-NN accuracy on real data, effectively choosing the 3,000 candidates that span the real data's decision regions most efficiently. The combination of clean labels and utility-aware

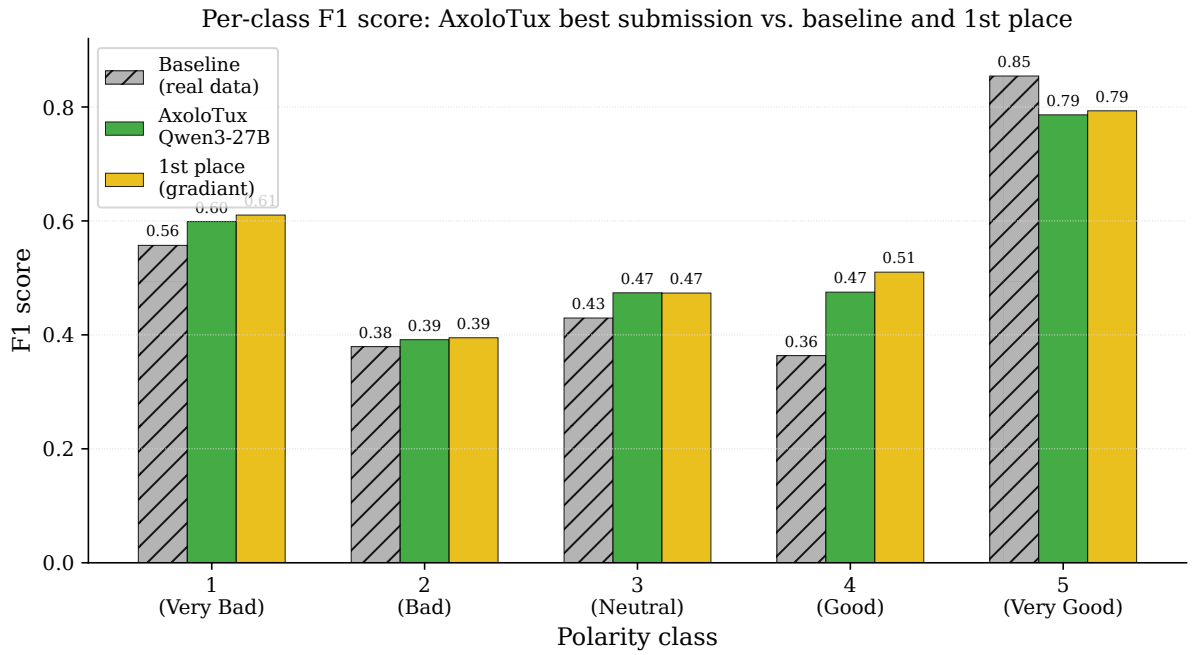

Figure 6: Per-class F1 score for the real-data baseline, our best submission (AxoloTux Qwen3.5-27B), and the top-ranked system (gradiant). Gains are concentrated in classes 3 and 4; class 5 slightly decreases due to reduced majority-class bias.

selection produces a corpus that is more informationally dense than an equal-sized random real sample.

4.2. What Precision and Track Score Each Reveal

Our best submission achieves the highest macro-averaged precision among all competing systems (0.5975), yet ranks 5th in Track Score (0.5114). These two facts are not contradictory: they reflect different aspects of the downstream classifier's behavior. Macro-averaged precision measures the rate of correct positive predictions across all classes uniformly. The Track Score, by contrast, is a minority-weighted macro-F1: classes 1 (Very Bad) and 2 (Bad), each representing roughly 2.6% of the data, receive near-unit weight, while the dominant class 5 (Very Good, 65.6%) receives the lowest weight. Our Qwen3.5-27B classifier recovers well on classes 1–3 and achieves the highest precision across all systems, but falls 0.035 F1 points behind the top-ranked system on class 4 (Good, 21.6% of data, weight ≈ 0.78). Although class 4 is not among the rarest classes and thus does not carry the highest weight, the large absolute gap there makes its weighted contribution the dominant term in the 0.011 Track Score difference. This suggests that the top system's synthetic corpus better covers the semantically ambiguous boundary between "Good" and "Very Good" reviews, a distinction that is notoriously hard to generate reliably. Closing the class 4 gap is therefore the clearest path to further improving Track Score.

4.3. Model Size as the Primary Driver

Across all our experiments, generator model size is the factor most consistently associated with downstream quality. Qwen3.5-27B outperforms Qwen3.5-9B by 0.018 Track Score points under identical filtering settings; both Qwen submissions surpass all three LLaMA-3B-based submissions despite the LLaMA approach benefiting from balanced training data. This pattern is consistent with the broader literature on scaling laws [11]: larger models produce more fluent, more coherent, and more style-consistent text, and these properties translate directly into more useful training signal for the downstream classifier.

The LoRA fine-tuning approach offers a different trade-off. Its advantage is controllability: the fine-tuned model reliably follows the structured prompt and produces polarity-consistent reviews even

for underrepresented classes. However, a 3B parameter model, even when fine-tuned on balanced data, cannot match the linguistic range of a 27B parameter model accessed through few-shot prompting. The balanced training distribution helps the generator avoid majority-class bias, but the resulting reviews lack the stylistic variety that makes the Qwen3.5 outputs useful across diverse test cases.

4.4. The Role of the Filtering Pipeline

The filtering pipeline proved critical for the Qwen submissions. With a very permissive realism threshold (0.05), the consistency and utility filters do the main work of separating informative from noisy candidates. The low realism threshold for Qwen is a deliberate choice: large instruction-tuned models produce structurally clean text that the real/synthetic discriminator easily identifies as synthetic; a high realism threshold would discard nearly all Qwen candidates regardless of their quality. This exposes a limitation of the discriminator-based realism score as a quality signal for large language models: it measures how much a review *looks synthetic*, not how *useful* it is. For the LoRA-based submissions, stricter thresholds (realism ≥ 0.35) are appropriate because the fine-tuned model occasionally produces malformed or repetitive text that genuinely resembles synthetic output in a problematic way.

The worst-performing submission, AxoloTux LLaMA (2), combined the strictest thresholds (consistency ≥ 0.85, realism ≥ 0.50) with global top-3,000 selection. The overly aggressive filtering left a pool too small and too homogeneous for effective utility selection, ultimately producing a corpus that underperformed the random real-data baseline. This confirms that threshold tuning is not monotone: beyond a certain point, stricter filtering hurts rather than helps.

4.5. Privacy-First Generation is Viable

All generation, scoring, and filtering ran on local hardware, yet the resulting system placed 2nd among teams. This shows that local inference with sufficiently large models (27B parameters) is a practical and competitive strategy for privacy-preserving synthetic data generation, relevant far beyond the tourism domain. For settings such as clinical NLP or legal document processing, where data cannot leave an organization's infrastructure, our pipeline offers a directly applicable template.

4.6. Limitations

Some limitations should be acknowledged. First, the evaluation is indirect: we cannot directly inspect whether the generated reviews are realistic, factually coherent, or free of harmful content, but only whether they produce a useful polarity classifier. Second, the design space exploration is narrow (six submissions, two strategies) and does not permit rigorous ablation. Third, the real/synthetic discriminator was trained on the same generation outputs it is intended to filter, creating a potential circularity. Finally, the Qwen3.5 submissions used the natural (imbalanced) polarity mode, leaving open the question of whether explicit polarity balancing at generation time would further improve results; however, the filtering process resulted in a nearly balanced dataset.

5. Conclusions

We presented the AxoloTux system for the REST-Mex 2026 synthetic data generation task, which challenges participants to produce 3,000 Spanish-language tourism reviews whose polarity signal transfers effectively to real data. Our approach combines two locally deployed generation strategies, namely few-shot prompting with Qwen3.5 (9B and 27B) and LoRA fine-tuning of Llama-3.2-3B on a class-balanced subset, combined with a three-stage quality filtering pipeline that sequentially screens candidates by label consistency, linguistic realism, and KNN-Shapley predictive utility.

Our best submission (Qwen3.5-27B, Track Score 0.511) placed 5th among 59 submissions and 2nd among competing teams, achieving an 8.3% improvement over a baseline that trains on 3,000 real reviews. This result demonstrates that carefully curated synthetic data can outperform an equal-sized real corpus,

not merely substitute for it. The system also achieved the highest macro-averaged precision across all submissions, indicating that the generated corpus produces classifiers with fewer false positives across polarity classes.

The main finding are:

- **Generator model size is the dominant factor.** Qwen3.5-27B substantially outperformed Qwen3.5-9B and Llama-3.2-3B under comparable filtering conditions. For practitioners constrained to local hardware, investing in the largest feasible model could be more impactful than complex fine-tuning strategies.
- **Utility-aware selection is necessary.** Filtering by consistency and realism alone is insufficient; the KNN-Shapley step selects the subset that is most informative for the downstream task. Fluency and label agreement are necessary but not sufficient conditions for a useful synthetic example.
- **Threshold calibration matters.** Overly strict filtering degrades performance by reducing the candidate pool below the diversity needed for effective utility selection. For large instruction-tuned models, the realism score is a poor quality proxy and should be applied with a very permissive threshold.
- **Local-first generation is competitive.** A fully local pipeline, with no real data transmitted to external services, achieves near-state-of-the-art results, validating the approach for privacy-sensitive applications in healthcare, legal services, and other restricted domains.

Therefore, this work could open the door to exploring the direct application of this process to privacy-sensitive corpora, where the generation of synthetic data has more significant consequences.

Acknowledgments

The authors thank the organizers of REST-Mex 2026 challenge for compiling the corpus and running the shared tasks.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) to assist in code prototyping and refine the language. After using these tools, the authors reviewed and edited the content as needed and assume full responsibility for the publication's content.

References

[1] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Foundations and Trends in Theoretical Computer Science 9 (2014) 211–407.

[2] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, M. H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Medical Research Methodology 10 (2010) 70.

[3] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, 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), 2019, pp. 6382–6388.

[4] M. Bayer, M.-A. Kaufhold, C. Reuter, A survey on data augmentation for text classification, ACM Computing Surveys 55 (2023) 1–39.

[5] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, et al., AugGPT: Leveraging ChatGPT for text data augmentation, arXiv preprint arXiv:2302.13007 (2023).

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[7] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, C. J. Spanos, Towards efficient data valuation based on the Shapley value, in: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR, 2019, pp. 1167–1176.

[8] M. Á. Álvarez-Carmona, R. Aranda, Á. Díaz-Pacheco, M. Reyes-Sierra, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Herrera-Semenets, Overview of rest-mex at iberlef 2026: Research on synthetic text modeling for mexican magical towns, Procesamiento del Lenguaje Natural 77 (2026).

[9] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[10] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Herrera-Semenets, Overview of rest-mex at iberlef 2025: Researching sentiment evaluation in text for mexican magical towns, volume 75, 2025.

[11] Qwen Team, Qwen3 technical report, https://qwenlm.github.io/blog/qwen3/, 2025. Accessed: 2026-06-04.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685.

[13] Meta, Llama-3.2-3b, https://huggingface.co/meta-llama/Llama-3.2-3B, 2024. Accessed: 2025-06-03.

[14] C. Minutti-Martinez, B. Escalante-Ramirez, J. Olveres-Montiel, Multitask analysis of spanish travel reviews: Sentiment, destination, and topic classification with roberta and llama ensembles, CEUR Workshop Proceedings (2025).

[15] Huihui-AI, Huihui-Qwen3.5-9B-abliterated, https://huggingface.co/huihui-ai/Huihui-Qwen3.5-9B-abliterated, 2026.

[16] Huihui-AI, Huihui-Qwen3.5-27B-abliterated, https://huggingface.co/huihui-ai/Huihui-Qwen3.5-27B-abliterated, 2026.

[17] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 4512–4525. URL: https://aclanthology.org/2020.emnlp-main.365.