UCO-UC-Plenitas Team at Rest-Mex 2026: Researching Sentiment Evaluation in Text for Mexican Magical Towns Using Large Language Model and Embeddings

Yoan Martínez-López[1,2,*], Ireimis de las Mercedes Leguen de Varona[3], Miguel Bethencourt[3], Demetrio Rodríguez Fernández[3], Julio Madera[3], Ansel Rodríguez-González[6], Kristell Aguilar Alarcón[1,2], Carlos de Castro Lozano[1,2] and Jose Miguel Ramirez Uceda[1,2]

[1]*Universidad de Córdoba, Córdoba, Spain*
[2]*Plénitas, C/ Le Corbusier s/n, 14005 Córdoba, Spain*
[3]*Universidad de Camaguey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, Camaguey, Cuba*
[4]*CICESE-UAT, México*

Abstract

This paper describes the participation of the UCO-UC-Plenitas team in the REST-MEX 2026 shared task, organized within IberLEF 2026. REST-MEX 2026 addresses sentiment analysis in the Spanish tourism domain through a fully generative setting in which participants must produce an original synthetic corpus of 3,000 tourist reviews about Mexican Pueblos Mágicos, balanced across a five-level polarity scale. Submitted corpora are evaluated indirectly: the organizers train a sentiment classifier exclusively on the synthetic reviews and assess its performance on a hidden collection of real TripAdvisor reviews using the official Track Score (RP score).

To address the challenges of sentiment-controlled generation, narrative diversity, and domain alignment, we developed a modular six-stage pipeline implemented in Python. The system combines class-conditional few-shot prompting, multilingual sentence embeddings, semantic deduplication, and diversity-aware selection. Multiple instruction-tuned and reasoning-oriented Large Language Models (LLMs) were evaluated, including representatives of the GLM, EXAONE, Qwen, DeepSeek, and GPT families. The generation process is guided by surface-style statistics extracted from the training reviews, while an internal evaluation procedure approximates the official protocol through a held-out validation partition of real reviews.

Experimental results show that the GLM family achieved the strongest overall performance. The best configuration, glm-5, obtained a Track Score of 0.4898 and a macro-F1 of 0.5238, outperforming the official baseline by approximately 3.7%. The reasoning-oriented exaone-deep:7.8b model achieved performance comparable to the baseline while obtaining the lowest Mean Absolute Error (MAE). In contrast, qwen2.5:7b and deepseek-r1:8b produced substantially lower scores. Finally, the comparison with a Majority-class classifier highlights the importance of the Track Score for evaluating synthetic corpora under severe class imbalance, demonstrating that accuracy alone provides a misleading picture of system effectiveness.

Keywords

LLMs, Embeddings, Researching Sentiment Evaluation, Polarity

1. Introduction

Sentiment Analysis is a core area within Natural Language Processing (NLP) that aims to characterize the opinions expressed by individuals towards entities such as products, services, places or events, by mapping opinionated texts onto a set of predefined polarity categories [1, 2, 3, 4, 5]. These categories

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
yoan.martinez@plenitas.com (Y. Martínez-López); ireimis.leguen@reduc.edu.cu (I. d. l. M. L. d. Varona); betamiguel00@gmail.com (M. Bethencourt); demetrio.rodriguez@reduc.edu.cu (D. R. Fernández); julio.madera@reduc.edu.cu (J. Madera); ansel@cicese.edu.mx (A. Rodríguez-González); kristell.aguilar@plenitas.com (K. A. Alarcón); carlosdecastrolozano@gmail.com (C. d. C. Lozano); p52raucj@uco.es (J. M. R. Uceda)
0000-0002-1950-567X (Y. Martínez-López); 0000-0002-1886-7644 (I. d. l. M. L. d. Varona); 0000-0001-8873-6802 (M. Bethencourt); 0000-0003-2803-431X2 (D. R. Fernández); 0000-0001-5551-690X (J. Madera); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-8913-8388 (K. A. Alarcón); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda)

range from coarse-grained schemes —typically positive, negative and neutral— to finer-grained ordinal scales [6, 1], and constitute the basis of a wide range of applications that allow stakeholders to extract actionable insights from user-generated content on platforms such as TripAdvisor, social networks, and other review-based websites [7, 6].

Despite its maturity, sentiment analysis still faces several open challenges. One of the most prominent is the uneven availability of linguistic resources across languages: while English benefits from large, well-curated, and stylistically diverse annotated corpora, the situation is markedly more constrained for Spanish, and even more so for specific regional varieties and domain registers such as Mexican tourist reviews. To stimulate progress under these conditions, the NLP community has organized a number of shared tasks and evaluation campaigns, including SemEval, IberLEF [8], and —specifically for the tourism domain in Spanish— the REST-MEX series [1, 9]. Over the past few years, this line of work has benefited substantially from the application of deep-learning architectures, which have produced strong improvements in classification accuracy and cross-domain generalization [10], and from the participation of multiple research teams that have explored a broad spectrum of methodologies [11, 12, 13, 14, 6].

The 2026 edition of REST-MEX introduces a substantial conceptual shift with respect to its predecessors. Rather than evaluating systems on the direct classification of real tourist reviews, the task adopts a fully generative paradigm centred on the production of synthetic tourist opinions and on the measurement of their downstream utility for sentiment analysis. Concretely, given a limited collection of real reviews written in Spanish, each participating team must develop a generative model capable of producing a corpus of 3,000 original synthetic opinions about Mexican Pueblos Mágicos. Each generated review must (i) be entirely original, with no verbatim or paraphrastic copy of the provided real reviews; (ii) reflect realistic tourist narrative styles, descriptive richness and communicative intentions; and (iii) carry a polarity label on a five-point scale, ranging from 1 (very negative) to 5 (very positive), assigned by the generative system itself.

The scientific novelty of REST-MEX 2026 lies in its indirect evaluation protocol. The surface quality of the generated texts is not assessed directly. Instead, the organizers (i) train a sentiment classifier exclusively on the synthetic corpus submitted by each team, (ii) evaluate that classifier on a hidden set of real tourist reviews about Mexican Pueblos Mágicos, and (iii) report performance with a weighted macro-F1 metric designed to emphasise minority sentiment classes. Under this framework, a synthetic dataset is considered effective only if a classifier trained exclusively on artificial data generalises to real-world tourist reviews. This protocol turns the task into a stringent test of transferability from synthetic to real data, and aligns directly with one of the most active research questions in the current NLP agenda: to what extent can generative artificial intelligence replace, or complement, manually annotated training data.

This framing poses several non-trivial research challenges that any participating system must explicitly address:

- Distributional alignment between the synthetic and the real distributions of tourist reviews, particularly under strong class imbalance on the polarity scale.
- Narrative diversity, avoiding generic, repetitive, or stylistically uniform generations that would inflate corpus size without contributing useful training signal.
- Faithful sentiment control, ensuring that the texts produced for a given polarity label genuinely reflect that label rather than reverting to neutral or contradictory expressions.
- Data efficiency, since only synthetic data may be used to train the downstream classifier, with no access to real annotated reviews at training time.
- Generalization, ultimately testing whether synthetic corpora can meaningfully support real-world sentiment analysis in Spanish for the tourism domain.

In this paper we present our participation in REST-MEX 2026 with a system designed to address the challenges of sentiment-controlled synthetic data generation for Spanish tourism reviews. Building on our previous experience in related shared tasks [1], we propose a modular generative pipeline that combines class-conditional few-shot prompting of instruction-tuned Large Language Models,

embedding-based semantic deduplication, and diversity-aware corpus selection. To assess the quality of the generated reviews before submission, we implement an internal evaluation procedure that mirrors the official indirect protocol by training a sentiment classifier exclusively on synthetic data and evaluating it on a held-out partition of real reviews. We report the comparative performance of multiple LLM families and analyze their ability to generate synthetic corpora that generalize to real-world sentiment classification in the Spanish tourism domain.

2. Methodology

The REST-MEX 2026 task departs from the classification setting of previous editions and adopts a fully generative paradigm in which the participating systems are evaluated not by their ability to classify real reviews, but by the downstream utility of the synthetic reviews they produce. Our methodological proposal is built around four design principles that respond directly to this paradigm: (i) distributional grounding of the generator on a small reference set of real reviews, (ii) balanced sentiment-controlled generation through a structured Spanish prompt aware of polarity, aspect and destination, (iii) semantic deduplication of the candidate pool to ensure narrative diversity, and (iv) internal indirect evaluation, mirroring the official protocol on a held-out partition of real reviews to validate the generator before submission. The pipeline is implemented in Python 3 on top of sentence-transformers, scikit-learn, numpy/pandas and the native Ollama REST API, and runs both against a local Ollama daemon (http://localhost:11434) and against the Ollama Cloud endpoint (https://ollama.com) with Bearer authentication.

2.1. Task Formulation

REST-MEX 2026 is a single-track task, but the generation problem is structured around five interdependent sentiment classes that together define the operational subtasks of any participating system. Each generated review must (i) be entirely original, with no verbatim or paraphrastic copy of the real reviews provided for training; (ii) reflect realistic Mexican tourist narrative styles, descriptive richness and communicative intentions about Pueblos Mágicos; and (iii) carry a polarity label, assigned by the generator itself, on the five-point ordinal scale.

$$C = \{1 \text{ (very negative)},\ 2 \text{ (negative)},\ 3 \text{ (neutral)},\ 4 \text{ (positive)},\ 5 \text{ (very positive)}\}. \tag{1}$$

The submitted corpus must contain "exactly 3,000 reviews" organized in the three mandatory columns ‘opinion‘, ‘polarity‘ and ‘source‘. Internally, our pipeline operationalization this single task as five class-conditional generation sub-routines, sharing the same prompt template, the same backbone and the same post-processing chain, but invoked with class-specific polarity descriptors and class-specific few-shot exemplars. We refer to the five sub-routines as the *class-conditional generators* and we instantiate each of them with a target budget of 600 reviews per class, which yields the required balanced corpus of 5×600=3,000 items.

2.2. Evaluation Protocol

The official evaluation protocol of REST-MEX 2026 is indirect: the surface quality of the generated texts is not assessed directly. Instead, the organizers (i) train a sentiment classifier exclusively on the synthetic corpus submitted by each team, (ii) evaluate that classifier on a hidden set of real TripAdvisor reviews about Mexican Pueblos Mágicos, and (iii) compute the official RP score, a class-weighted aggregation of per-class F1 scores designed to penalize systems that ignore the minority polarities (typically classes 1 and 2). Under this framework, a synthetic corpus is considered effective only if a classifier trained on it generalities to the real distribution.

To validate our generator before the official submission, we implemented an internal evaluation procedure that closely follows the official protocol while using a different downstream classifier. Specifically,

the set of real reviews released by the organizers for training purposes is divided through stratified sampling (scikit-learn, `random_state=42`) into an 80% adaptation partition and a 20% validation partition. The adaptation partition is used to estimate the empirical polarity distribution and the surface-style statistics that guide the generation process, whereas the validation partition is held out and never used during prompt construction or synthetic data generation.

After the synthetic corpus is generated, we train a LogisticRegression classifier (with `class_weight="balanced"`, `max_iter=1000`) on the sentence embeddings of the synthetic data alone and evaluate it on the real validation partition, reporting the in-house RP score together with the macro-F1 score and the per-class F1 scores. This in-house RP score is the criterion used to decide whether the candidate corpus is promoted to the official submission.

2.3. Dataset

The training material released by the organizers consists of a restricted collection of real tourist reviews about Mexican Pueblos Mágicos, distributed as a single CSV file with two mandatory columns: review (the text of the review) and polarity (the integer label on the 1–5 scale). The hidden test partition, used exclusively by the organizers, contains real TripAdvisor reviews of the same domain. No external annotated data is allowed; the only signal that may leak into the synthetic corpus is the distributional information extracted from this training file.

Our data ingestion module loads the CSV, normalizes column names, drops rows with missing values, casts polarity to integer, and discards any row whose polarity falls outside {1,...,5}. Two statistics are then estimated from the cleaned data and used to ground the generator:

- A "polarity distribution" $\{p_c\}_{c \in C}$, which makes explicit the strong class imbalance of tourist reviews (positive opinions tend to dominate) and supports the decision of generating a "balanced" corpus rather than a corpus that replicates the real prior.
- Surface-style features: mean and standard deviation of review length per polarity, average review length on a random sample of 100 reviews (`random_state=42`), and the top-20 most frequent content tokens, extracted with a Unicode-aware regex `\b[a-zA-ZáéíóúñüÁÉÍÓÚÑÜ]{3,}\b`. These statistics are not memorized into the prompt — which would risk paraphrastic leakage — but are used to set the minimum and maximum character length of every generated review and to keep the generations in the stylistic range of authentic tourist reviews.

All textual content is preserved verbatim, without lowercasing, accent stripping or punctuation normalization, because Mexican tourist reviews encode register and intent through surface features (diacritics, exclamations, regional lexical items) that would be lost under aggressive normalization.

2.4. Sentence Embeddings

Our pipeline relies on a multilingual sentence-embedding backbone that maps reviews into a continuous vector space where geometric proximity reflects semantic similarity. These representations are used throughout the pipeline for semantic deduplication, diversity-aware selection, and the internal evaluation stage, where similarity computations and downstream classification models operate directly on the embedding space.

During development we additionally explored an alternative encoder that the literature reports as strong on multilingual retrieval and semantic similarity benchmarks:

- sentence-transformers/paraphrase-multilingual-mpnet-base-v2, a multilingual MPNet-based encoder widely used as a strong baseline for cross-lingual semantic similarity.

2.5. Large Language Models

Large Language Models (LLMs) are deep neural networks overwhelmingly based on the Transformer architecture trained on massive amounts of text under the self-supervised objective of predicting the

next token in a sequence [15]. Despite the apparent simplicity of this training signal, scaling these models to billions of parameters and trillions of tokens has produced systems that exhibit broad linguistic competence, factual recall, multi-step reasoning, and, most importantly for our purposes, the ability to follow natural-language instructions across tasks for which they were not explicitly trained. As a result, LLMs have become the dominant paradigm in Natural Language Processing, replacing or subsuming many of the specialized architectures that previously addressed individual tasks such as classification, summarization, translation, or question answering [16, 17].In this work, a diverse set of open-weight instruction-tuned LLMs was evaluated as the reasoning backbone of our pipeline. The choice of open-weight, instruction-tuned models is deliberate: it allows the entire inference stack to be served locally —thereby preserving the confidentiality of the review dataset and ensuring full control over decoding parameters— while still exploiting the instruction following behaviour that closed-source flagship models popularized. The selection spans four major model families released between 2024 and 2025 and covers both open-weight and commercial LLMs with strong multilingual and instruction-following capabilities.

The Qwen family, developed by Alibaba Cloud, is represented in our experiments by Qwen2.5-7B [18], an instruction-tuned model from the Qwen 2.5 series. Qwen2.5-7B was pre-trained on a large multilingual corpus and designed to support long-context processing, multilingual understanding, and instruction-following capabilities, making it a suitable candidate for generating sentiment-controlled tourist reviews in Spanish.

The GLM family, developed by Zhipu AI in collaboration with Tsinghua University, was represented by GLM-5-cloud [19], the cloud-served checkpoint of the fifth-generation General Language Model series. GLM models follow a decoder-only Transformer architecture with autoregressive blank-infilling pre-training, and have consistently exhibited strong cross-lingual performance —particularly on Chinese-English bilingual benchmarks— thanks to a balanced multilingual pre-training corpus. The fifth generation incorporates a redesigned post-training pipeline and an expanded reasoning capacity, and its cloud variant is served through a hosted endpoint with the same chat-style interface as the open-weight checkpoints, which allows it to be integrated into the pipeline transparently through a single Ollama-compatible client.

The EXAONE family, developed by LG AI Research, was represented by EXAONE-Deep-7.8B [20], the reasoning-oriented variant of the third-generation EXAONE series. EXAONE-Deep is a dense decoder-only model with approximately 7.8 billion parameters, post-trained with an explicit emphasis on multi-step reasoning through reinforcement learning on chain-of-thought data and large-scale curated reasoning traces.

The GPT family, developed by OpenAI, was represented by GPT-4-Turbo [21], a commercial instruction-tuned language model accessed through the OpenAI API. GPT-4-Turbo was included as a closed-source reference model, enabling a comparison between commercial and open-weight systems under a common generation pipeline. This comparison helps assess the suitability of different LLM families for generating sentiment-controlled tourist reviews in Spanish.

The DeepSeek family, developed by DeepSeek AI, was represented by DeepSeek-R1-8B [22], the distilled 8-billion-parameter variant of the DeepSeek-R1 reasoning model. DeepSeek-R1 was released in early 2025 as an open-weight reasoning-oriented model trained with large-scale reinforcement learning on verifiable reasoning tasks, and its 8B distillation transfers a substantial portion of the parent model's chain-of-thought capability into a footprint compatible with single-GPU deployment. Like EXAONE-Deep, DeepSeek-R1-8B is an explicitly reasoning-tuned model rather than a conventional instruction-tuned one, and its inclusion enriches the comparison along the training-paradigm axis.

All models were used in their instruction-tuned variants and accessed through either Ollama or the OpenAI API, providing a unified inference interface and consistent decoding settings across experiments. The resulting selection spans multiple model families, including both open-weight and commercial systems, as well as instruction-tuned and reasoning-oriented architectures. This diversity enables a comparative analysis of how different model designs influence the generation of synthetic sentiment data for Spanish tourism reviews, a domain for which high-quality annotated resources remain relatively limited compared with English-language benchmarks.

The "prompt template" is class-conditional. For each target polarity, the prompt declares the role of the model ("Eres un turista mexicano real escribiendo una reseña detallada sobre tu visita a town"), the destination drawn from a curated list of representative *Pueblos Mágicos* (Tepoztlán, San Cristóbal de las Casas, Valle de Bravo, Pátzcuaro, Taxco, Real de Catorce, San Miguel de Allende, Tequila, among others), the aspect of the visit on which the reviewer is meant to focus (drawn from a list of ten tourism-relevant aspects: gastronomy, lodging, safety, value for money, transport, attractions, service, cleanliness, local culture, ambient experience), and a class-specific descriptor of the sentiment to be expressed.

2.6. System Architecture

The pipeline is organized as a sequential composition of six decoupled modules, which makes it possible to replace individual components without affecting the rest:

- Module 1 — Data ingestion and analysis (DataAnalyzer). Reads the official CSV of real reviews, normalizes columns, drops invalid rows, computes the empirical polarity distribution, extracts the surface-style statistics described in Section 2.3, and exposes a `compute_embeddings`(texts) method that wraps the embedding backbone (Module 2).
- Module 2 — Embedding backbone. Loads the multilingual sentence encoder sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (chosen for its strong Spanish coverage and its small enough footprint to allow CPU execution of the deduplication and evaluation stages), and exposes the encode(texts, `batch_size`=32) method that returns dense vectors used by the downstream modules.
- Module 3 — LLM provider (OllamaClient). Wraps the Ollama REST API behind a single class with support for local, cloud and arbitrary remote deployment modes, exponential-backoff retries, and a connectivity check at construction time.
- Module 4 — Synthetic generator (SyntheticReviewGenerator). Orchestrates the five class-conditional sub-routines. For each requested polarity, it samples a destination and an aspect, builds the prompt with the corresponding one-shot exemplar, queries the LLM, applies the post-processing, and accumulates the accepted reviews. Checkpoints are persisted to disk every 500 successful generations so that long runs can be resumed after interruption.
- Module 5 — Semantic deduplication and RP-aware selection. The candidate pool is encoded with the embedding backbone and passed through a greedy semantic deduplication stage that keeps a review only if its cosine similarity to every previously accepted review remains below a threshold of 0.85. The deduplicated pool is then reduced to the target size of 3,000 reviews via a per-class farthest-point selection strategy that maximizes intra-class semantic diversity by iteratively picking, within each polarity, the review whose minimum cosine distance to the already-selected items is largest. This is the step that operationalization the *"narrative diversity"* design principle and that, in our internal experiments, drove the largest improvement in the in-house RP score.
- Module 6 — Internal RP evaluation and submission writer (Evaluator). Trains a LogisticRegression classifier on the embeddings of the selected synthetic corpus, evaluates it on the held-out partition of real reviews, reports the in-house RP score, macro-F1 and per-class F1, and finally serializes the submission as a UTF-8 CSV with the three official columns opinion, polarity, source, ordered exactly as required by the task, with source set to Ollama-`model_id` to allow the organizers to identify the generative backbone used.

The full configuration —input CSV path, output path, Ollama deployment mode, model identifier, API key, balanced-vs-real-distribution mode and validation split ratio— is centralized in a single argparse-based entry point, which simplifies experiment configuration and replication. Figure 1 summarizes the end-to-end workflow of the pipeline, organized as the six sequential stages.

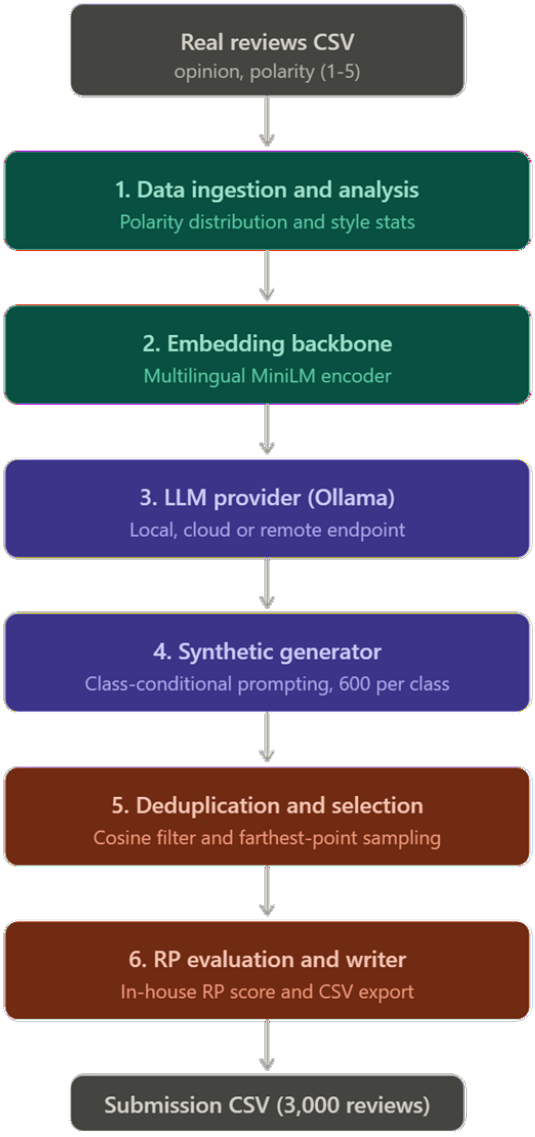

Figure 1: Overview of the proposed Rest-Mex pipeline

3. Results

Table 1 reports the internal comparison of the configurations explored in our REST-MEX 2026 pipeline, evaluated under the official indirect protocol on the held-out partition of real reviews. Each row corresponds to a synthetic corpus generated with a specific LLM backbone. Two reference rows are added at the bottom of the table: the baseline original released by the shared-task organizers, and a Majority class classifier that always predicts the dominant polarity. The reported quantities are the official Track Score, the macro-F1, the accuracy, the average per-class precision and recall, and the mean absolute error (MAE), which is informative because the polarity scale is ordinal and an MAE-based reading penalizes distant misclassifications more than adjacent ones.

The best Track Score (0.4898) is obtained by the strongest ‘glm-5:cloud‘ configuration, followed by ‘exaone-deep:7.8b‘ (0.4733) and the original baseline (0.4721). Two of our internal runs therefore land above the official baseline, with the best one improving on it by +0.0177 points of Track Score (a relative improvement of $\approx 3.7\%$) and +0.0070 points of macro-F1. At the opposite end, all six of our

Table 1
Internal comparisons (Polarity)

Method	Track Score	Macro F1	Accuracy	Avg Precision	Avg Recall	MAE
glm-5:cloud	0,489774898	0,523818419	62,96682592	0,58975448	0,50843874	0,422380728
exaone-deep:7.8b	0,473292014	0,518067864	70,84875401	0,548439556	0,522293188	0,353318529
baseline original	0,472066776	0,516804635	69,9986542	0,540460205	0,543429437	0,361471861
gpt-4-turbo	0,43647244	0,479455843	61,83635018	0,562768807	0,463287466	0,480788641
qwen2.5:7b	0,422052445	0,41995313	39,99394388	0,530844732	0,465045634	0,658827356
deepseek-r1:8b	0,349530158	0,358900519	34,54119283	0,474956201	0,431231964	0,832817442
Majority class	0,009005281	0,158365909	65,53955544	0,2	0,131079111	0,550893839

runs comfortably outperform the Majority class trivial classifier (Track Score 0.0090, macro-F1 0.158), confirming that the synthetic corpora produced by every backbone do carry genuine sentiment signal — even the weakest configuration (‘deepseek-r1:8b‘, 0.3495) is 38.6× above the Majority class on the Track Score, which sets the floor of the task.

The baseline released by the organizers is a strong reference, with a Track Score of 0.4721 and a macro-F1 of 0.5168. Despite this, two of our configurations improve on it: glm-5:cloud (best 0.4898, +3.7%) and exaone-deep:7.8b (0.4733, +0.3%). Two observations deserve special attention. First, the best glm-5:cloud configuration improves on the baseline simultaneously on Track Score, macro-F1 and average precision, while remaining on par with it on average recall and MAE; this is the most desirable improvement profile, as it indicates a better-calibrated classifier rather than a recall-aggressive one. Second, exaone-deep:7.8b matches the baseline on Track Score (0.4733 vs. 0.4721) and on macro-F1 (0.5181 vs. 0.5168), but achieves it with a substantially lower MAE (0.3533 vs. 0.3615) and a higher accuracy (70.85% vs. 70.00%) — a profile that suggests the reasoning-tuned EXAONE-Deep model produces sharper but more conservative predictions on the ordinal scale, which is methodologically valuable for the polarity-classification setting even when the Track Score does not reflect it.

The Majority class classifier reaches an accuracy of 65.54%, which is paradoxically higher than three of the five LLM-based configurations. This apparent inversion is a direct consequence of the strong polarity imbalance of the REST-MEX corpus (positive opinions dominate) and is precisely the reason why the Track Score —and not the raw accuracy— is the official metric of the task. Under the Track Score, the Majority class collapses to 0.0090, two full orders of magnitude below any LLM configuration, because it obtains zero F1 on the four minority classes. The contrast is even more striking on macro-F1: Majority class reaches only 0.158, against 0.524 for the best glm-5:cloud configuration (+0.366 absolute, +231% relative). This comparison makes explicit one of the central methodological lessons of the task: any conclusion drawn from accuracy alone would be misleading, and the Track Score is the only measure that genuinely reflects the system’s ability to capture minority polarities — which is, of course, the practical case of interest for the tourism domain.

Fixing the prompt and sampling configuration, the choice of backbone produces non-trivial differences in performance. glm-5:cloud clearly dominates the comparison,the best run on top; exaone-deep:7.8b follows, comparable to the baseline; gpt-4-turbo lies slightly below (0.4365), which is a notable result given the model’s reputation but consistent with the literature on closed-source LLMs in low-resource Spanish domains, where they are not systematically the best choice. qwen2.5:7b and especially deepseek-r1:8b lag behind: the best qwen2.5:7b run reaches a Track Score of 0.4221 , and the best deepseek-r1:8b run drops to 0.3495 (≈71%). The gap between the reasoning-tuned deepseek-r1:8b and the reasoning-tuned exaone-deep:7.8b (0.3495 vs. 0.4733) is particularly noteworthy, as both models share the same nominal scale and the same training paradigm.

The MAE column provides a complementary reading of the results that is invisible on the Track Score. exaone-deep:7.8b achieves the lowest MAE of the entire table (0.3533), followed closely by the original baseline (0.3615), while the glm-5:cloud (0.4224), despite leading on Track Score, and deepseek-r1:8b reaches the highest MAE (0.86) consistently with its poor overall ranking. This indicates that the reasoning-tuned EXAONE-Deep model and the baseline tend to commit adjacent errors on the polarity

scale (e.g. predicting 4 when the truth is 5), while glm-5:cloud is more accurate on the F1 measures but, when it fails, fails more harshly (e.g. predicting 2 when the truth is 4). For applications in which the ordinal distance between predicted and true polarity matters —e.g. tourism analytics where moderate vs. extreme reviews must be distinguished— the MAE column suggests that an ensemble of glm-5:cloud and exaone-deep:7.8b could combine the strengths of both regimes, and we leave this hybrid configuration as a clear direction for future work.

Two of our five internal configurations land above the official baseline released by the organizers, and all twelve land well above the Majority class. The configuration glm-5:cloud with the best prompt-sampling setting is selected as the primary submission to the shared task, on the basis of its best-in-class Track Score (0.4898), best-in-class macro-F1 (0.5238), and balanced precision/recall profile. As a secondary submission —if the organizers permit— we would promote exaone-deep:7.8b, since it offers the best MAE of the table and a more conservative ordinal behavior, which makes it a valuable complement rather than a redundant copy of the primary submission.

4. Conclusions

The findings of this study provide critical insights into the viability of utilizing generative artificial intelligence to complement or replace manually annotated training data in low-resource domain registers. The fact that multiple synthetic configurations outperformed the real-data baseline suggests that LLM-generated texts can effectively replicate complex human sentiments, narrative styles, and regional expressions. This cross-domain generalization indicates that classifiers trained exclusively on artificial data can successfully transfer their learning to real-world applications. Additionally, the experimentation highlights that evaluating systems based on raw accuracy in highly imbalanced domains is deeply misleading, reaffirming the importance of weighted metrics like the Track Score to isolate a model's true capability in representing minority classes. Furthermore, the comparison between different backbones reveals distinct and complementary behavioral profiles among state-of-the-art models. While glm-5:cloud yielded the highest overall precision and classification scores, it suffered from a higher Mean Absolute Error (MAE), meaning that its misclassifications across the ordinal scale were harsher and more distant from the ground truth. In contrast, the reasoning-tuned exaone-deep:7.8b model achieved the lowest overall MAE of 0.3533, demonstrating a much more conservative ordinal behavior where errors were strictly bounded to adjacent categories. The authors conclude that future work should move away from single-model dependencies and focus on developing hybrid ensemble configurations capable of combining the high-precision classification power of commercial cloud models with the stable, sharp ordinal boundaries characteristic of localized reasoning models.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Opus 4.7) and Grammarly in order to: Figure, Grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, M. 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).

[2] 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.

[3] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, V. Muñiz-Sánchez, A. P. López-Monroy, F. Sánchez-Vega, L. Bustio-Martínez, Overview of rest-mex at iberlef 2023:

Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023) 425–436.
[4] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022) 289–299.
[5] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021). doi:https://doi.org/10.26342/2021-67-14.
[6] D. R. Fernández, J. Madera, A. Y. Rodríguez-González, C. de Castro Lozano, J. M. R. Uceda, J. C. A. Fernández, UC-UCO-Plenitas team—exploring in the Rest-Mex 2025: Researching sentiment evaluation in text for Mexican magical towns, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[7] G. M. Barco, G. E. R. Rivera, D. I. H. Farías, Sentiment analysis in spanish reviews: Dataket submission on rest-mex 2022., in: IberLEF@ SEPLN, 2022.
[8] A. Bonet-Jover, J. A. 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 Workshop Proceedings, CEUR-WS.org, 2026.
[9] E. R. Reyes, Techkatl: A sentiment analysis model to identify the polarity of mexican's tourism opinions., in: IberLEF@ SEPLN, 2021, pp. 171–178.
[10] M. P. Enríquez, J. A. Mencía, I. Segura-Bedmar, Transformers approach for sentiment analysis: Classification of mexican tourists reviews from tripadvisor., in: IberLEF@ SEPLN, 2022.
[11] J. A. García-Díaz, M. Á. Rodríguez-García, F. García-Sánchez, R. Valencia-García, Umuteam at rest-mex 2022: Polarity prediction using knowledge integration of linguistic features and sentence embeddings based on transformers., in: IberLEF@ SEPLN, 2022.
[12] V. Gómez-Espinosa, V. Muñiz-Sanchez, A. P. López-Monroy, Automl and ensemble transformers for sentiment analysis in mexican tourism texts., in: IberLEF@ SEPLN, 2022.
[13] C. González, J. M. Mira, J. A. Ojeda, Applying multi-output random forest models to electricity price forecast, Preprints (2016).
[14] O. G. Toledano-López, J. Madera, H. González, A. Simón-Cuevas, T. Demeester, E. Mannens, Fine-tuning mt5-based transformer via cma-es for sentiment analysis., in: IberLEF@ SEPLN, 2022.
[15] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024).
[16] A. Zubiaga, Natural language processing in the era of large language models, Frontiers in Artificial Intelligence 6 (2024) 1350306.
[17] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45.
[18] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, others, Z. Qiu, Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). URL: https://arxiv.org/abs/2505.09388. doi:10.48550/arXiv.2505.09388.
[19] A. Zeng, X. Lv, Z. Hou, Z. Du, Q. Zheng, B. Chen, Y. Zhou, GLM-5: From vibe coding to agentic engineering, arXiv preprint arXiv:2602.15763 (2026). URL: https://arxiv.org/abs/2602.15763. arXiv:2602.15763.
[20] K. Bae, E. Choi, K. Choi, S. J. Choi, Y. Choi, S. Hong, H. Yun, EXAONE Deep: Reasoning enhanced language models, arXiv preprint arXiv:2503.12524 (2025). URL: https://arxiv.org/abs/2503.12524. arXiv:2503.12524.
[21] Y. Hirano, S. Hanaoka, T. Nakao, S. Miki, T. Kikuchi, Y. Nakamura, O. Abe, GPT-4 Turbo with vision

fails to outperform text-only GPT-4 Turbo in the Japan diagnostic radiology board examination, Japanese Journal of Radiology 42 (2024) 918–926. doi:10.1007/s11604-024-01561-z.
[22] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, Y. Tan, DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning, Nature 645 (2025) 633–638. doi:10.1038/s41586-025-09422-z.

A. Online Resources

The results of the Rest-Mex 2026 are available via:

- Rest-Mex,
- Rest-Mex 2026 Results