Synthetic Text Generation for Mexican Magical Towns via Supervised Fine-Tuning with QLoRA

Oscar Oviedo-Tapia[1,*], Jairo Serrano[1], Juan Martinez-Santos[1] and Edwin Puertas[1]

[1]*Universidad Tecnológica de Bolívar, Cartagena, Colombia*

Abstract

This working notes paper presents our approach for the REST-MEX 2026 shared task at IberLEF 2026, which focuses on Synthetic Text Modeling for Mexican Magical Towns. Generating artificial text that accurately reflects the tourism domain requires models to deeply understand local context, cultural nuances, vocabulary, and specific sentiment expressions. Instead of relying on zero-shot inference, our methodology is centered on the Supervised Fine-Tuning (SFT) of a Large Language Model using the official training corpus provided by the organizers. Driven by hardware constraints, the model was fine-tuned using Quantized Low-Rank Adaptation (QLoRA) combined with the Unsloth framework on a local machine equipped with a consumer-grade NVIDIA RTX 3060 GPU. Our approach included an exhaustive exploratory data analysis phase to identify and mitigate extreme class imbalances, followed by targeted fine-tuning and strict prompt engineering. The qualitative and quantitative validations of the generated synthetic reviews demonstrate the model's capability to natively produce coherent, domain-specific text that faithfully captures the requested polarities, towns, and attraction types, highlighting the viability of local LLM deployment for specialized NLP tasks.

Keywords

Natural Language Processing, Synthetic Text Generation, Large Language Models, Supervised Fine-Tuning, QLoRA, REST-MEX 2026

1. Introduction

The contemporary tourism sector operates under a fundamentally digital paradigm, where market dynamics, destination reputation, and consumer decisions are intrinsically linked to user-generated content (UGC). Online reviews published massively on platforms like TripAdvisor, Google Reviews, and Booking.com not only act as feedback mechanisms for service providers but constitute the primary source of asymmetric information potential travelers rely upon to evaluate accommodations, gastronomic experiences, and cultural attractions. Within the geography and tourism policy of Mexico, the "Pueblos Mágicos" (Magical Towns) initiative represents a top-tier governmental and commercial strategy designed to highlight and protect the historical, architectural, gastronomic, and symbolic richness of specific localities distributed throughout the national territory. The digital discourse surrounding these Magical Towns possesses a profoundly particular linguistic identity, characterized by an amalgamation of regional idioms, endemic gastronomy references, descriptions of colonial or pre-Hispanic architecture, and a highly specific sociocultural framework that differs substantially from traditional sun-and-beach tourism.

In this context of linguistic and cultural richness, natural language processing (NLP) applied to tourism faces substantial analytical challenges. To drive the development of computational tools capable of processing this volume of information, the IberLEF 2026 evaluation forum has established the REST-MEX 2026 (Research on Synthetic Text Modeling for Mexican Magical Towns) challenge [1, 2]. Unlike previous editions of the forum, which prioritized discriminative tasks such as sentiment analysis, polarity classification, recommendation systems, and epidemiological traffic light prediction

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

ooviedo@utb.edu.co (O. Oviedo-Tapia); jserrano@utb.edu.co (J. Serrano); jcmartinezs@utb.edu.co (J. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

0009-0009-0201-8179 (O. Oviedo-Tapia); 0000-0001-8165-7343 (J. Serrano); 0000-0003-2755-0718 (J. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

from tourist texts, the 2026 edition introduces a fundamental track focused on synthetic text generation.

The generation of synthetic data has consolidated as an imperative need in the current artificial intelligence ecosystem. As deep learning architectures require increasingly larger volumes of training data to generalize correctly, researchers and developers frequently collide with domains where real-world data is scarce, protected by strict privacy regulations, or exhibits severely imbalanced statistical distributions [3]. The tourism domain is a paradigmatic example of this last problem. Consumer psychology dictates that individuals are strongly motivated to invest the cognitive effort of writing an online review primarily when their experience has far exceeded their expectations, or, to a lesser extent, when they have suffered a catastrophic incident. This generates a massive statistical void: neutral, moderately negative, or highly analytical reviews are statistically anomalous, while maximum satisfaction ratings flood data repositories. Therefore, the technical capability to artificially synthesize high-fidelity reviews conditioned to specific sentiment polarities represents an indispensable solution for data augmentation and bias mitigation [4].

Addressing generative tasks in the Spanish language requires deploying architectures capable of modeling highly complex linguistic distributions. During recent years, Large Language Models (LLMs) have demonstrated exceptional capabilities in zero-shot and few-shot learning environments. However, generic inference using foundational LLMs frequently produces aseptic texts, overly formalized, and lacking the cultural authenticity and specific colloquialisms demanded by highly localized tasks like REST-MEX. To overcome this limitation, this research proposes a methodology based on the Supervised Fine-Tuning (SFT) of the Meta Llama 3 8-billion parameter model [5].

2. Related Work and Methodological Foundation

The advancement in natural language processing architectures and the adaptation of large language models to highly specialized tasks does not emerge from a conceptual vacuum, but from the iterative accumulation of algorithmic strategies proven in competitive evaluation environments. The architecture and design decisions presented for the REST-MEX 2026 challenge build directly upon our group's extensive trajectory of research oriented toward deep semantic feature extraction, handling imbalanced data distributions, and computational resource optimization.

The capability to identify, classify, and model subtle variations in the tone, polarity, and lexical features of digital discourse has been rigorously developed in tasks related to mental health and social interaction dynamics. In the MentalRiskES 2025 challenge, we explored the early detection of risks associated with gambling disorders using data from messaging platforms [6]. The methodological approach implemented by the VerbaNexAI Lab demonstrated that combining contextual embeddings generated via Spanish BERT models with traditional machine learning algorithms produced exceptionally robust results for real-time risk identification. In this research, severe class imbalance was mitigated using advanced random oversampling in the 768-dimensional embedding space [6]. The experience acquired in handling niche vocabularies and rigorously mitigating class imbalance in MentalRiskES was fundamental in designing the undersampling and corpus curation strategy for the tourist reviews in the present project, ensuring the REST-MEX generative model does not suffer from mode collapse toward the majority five-star class.

Similarly, deep modeling of semantic relationships and multilayer feature extraction were validated during the SemEval-2024 competition (Task 1: Semantic Relationship Detection) [7]. Our system proposed a multilayer artificial intelligence model utilizing Long Short-Term Memory (LSTM) networks to extract sentence features from phoneme embeddings and sentence transformers, successfully capturing semantic nuances that traditional lexical models overlooked [7]. This level of rigor in processing syntactic and contextual features directly guided the strict XML-based prompt engineering strategy used in REST-MEX, forcing the Llama model to explicitly integrate attraction types and polarities structurally.

Furthermore, research on the early detection of depression markers presented in the eRisk 2025 forum by our COTECMAR-UTB team provided critical perspectives on the spatial location of semantic concepts [8]. By mapping the emotional intensity of texts using semantic-centroid symptom ranking

and adaptive decision rules [8], we established the logical foundation for the qualitative validation of synthetic texts in REST-MEX: ensuring that reviews generated for a 1-star polarity occupy a semantic space diametrically opposed to 5-star reviews.

Regarding the direct integration of LLMs, methodologies implemented in the ELOQUENT Lab 2025 established the computational protocols for human preference prediction and explanation generation using Retrieval-Augmented Generation (RAG), Few-Shot Learning, and Auto-CoT [9]. Finally, the development of a hybrid retrieval system driven by LLMs and the PubMed API for biomedical question answering consolidated the viability of deploying lightweight architectures [10]. Balancing high accuracy with computational feasibility in benchmark tests like BioASQ proved that efficient fine-tuning techniques can rival brute-force computing systems [10]. All this methodological scaffolding culminates in our selection and implementation of Llama 3 with QLoRA to model the Mexican tourist narrative.

3. Data: Exploration, Curation, and Structural Formatting

The effectiveness and logical coherence of any autoregressive generative model are absolutely determined by the statistical distribution, semantic diversity, and cleanliness of the corpus on which it is trained.

3.1. Exploratory Analysis and Severe Asymmetry

The official training dataset provided by the organization is statistically massive, comprising over 208,051 individual records of tourist opinions. Each instance contains metadata attributes—such as Polarity (1 to 5), Magical Town, and Attraction Type (*Hotel*, *Restaurant*, *Attractive*)—coupled with the unstructured text of the review.

The Exploratory Data Analysis (EDA) immediately exposed a critical structural vulnerability for generative modeling: the corpus presented an extreme class imbalance in the distribution of polarities. This phenomenon reflects the self-selection bias of tourists in online feedback platforms, where positive motivations overwhelmingly surpass punitive ones. Table 1 and Figure 1 present the real distribution extracted from the analysis.

Table 1
Polarity Distribution in the Original Training Dataset

Polarity (Satisfaction Level)	Real Records	Corpus Percentage
1.0 (Very Poor)	5,441	2.61%
2.0 (Poor)	5,496	2.64%
3.0 (Neutral)	15,519	7.46%
4.0 (Good)	45,034	21.65%
5.0 (Excellent)	136,561	65.64%

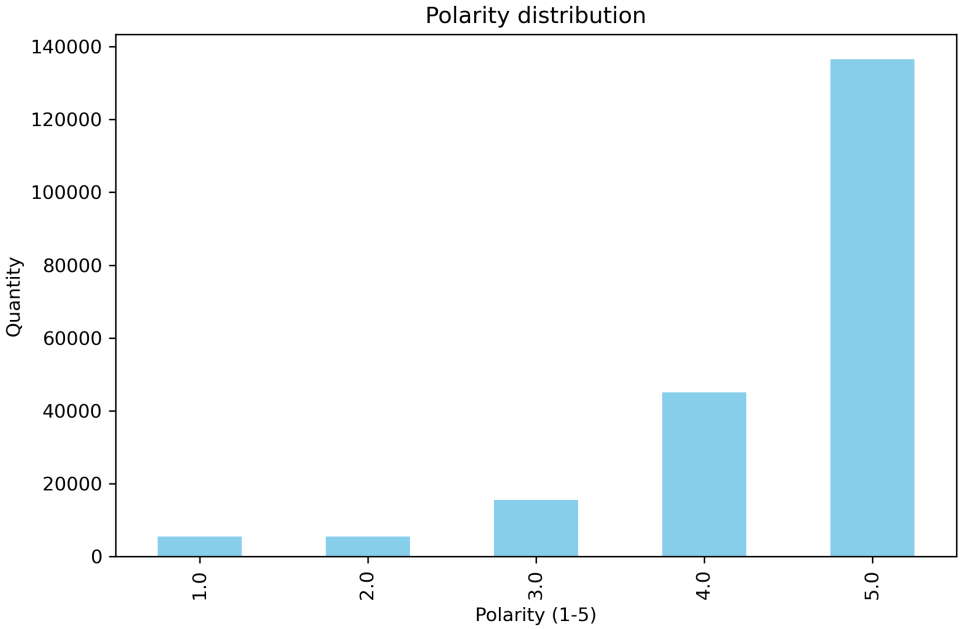

Figure 1: Distribution of Review Polarities illustrating extreme class imbalance in the training corpus.

As is evident visually in Figure 1 and Table 1, 5-star opinions constitute approximately 65.64% of the total data volume, while critical opinions (1 and 2 stars) sum up to barely 5.25%. This disproportion poses an unacceptable algorithmic risk.

3.2. Mode Collapse Mitigation via Rigorous Undersampling

Training a generative architecture to optimize maximum likelihood estimation on this asymmetrical corpus leads inexorably to mode collapse [3]. The network learns that the safest probabilistic route is to over-represent the linguistic patterns of the dominant class, generating strings of superlative adjectives and completely losing the ability to structure logical complaints.

While traditional classification scenarios often mitigate this via oversampling (as done in our MentalRiskES approach [6]), direct oversampling in LLM fine-tuning can cause severe overfitting, leading the model to verbatim memorize the few existing complaints. Therefore, we implemented a drastic undersampling strategy. We established a strict threshold, limiting the admitted samples to a maximum of 2,000 records per polarity level. Retained records were randomly selected to maintain geographic diversity, reducing the operative corpus to a perfectly balanced subset of 10,000 reviews.

3.3. Conversational Format and Structural Prompt Engineering

Modern Llama 3 models are instruction-tuned. To ensure the SFT modified the latent parameters appropriately, the balanced samples were programmatically transformed into a restrictive pseudo-XML conversational format. This design maps the target metadata directly into the system prompt, forcing the LLM to condition its generation before predicting the first token:

Listing 1: Prompt Template for Synthetic Text Generation

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
Eres un turista compartiendo una reseña original en TripAdvisor
   sobre un {tipo} en un Pueblo Mágico mexicano.
El nivel de tu satisfacción es exactamente {polaridad} estrellas
    de 5.
Usa vocabulario mexicano coloquial pero natural.<|eot_id|><|
   start_header_id|>user<|end_header_id|>
Escribe primero el "Título: " y en la siguiente línea "Reseña: "
   <|eot_id|><|start_header_id|>assistant<|end_header_id|>
Título:
```

4. System Architecture and Adaptation Framework

Coupling the vast prior knowledge of an LLM to a niche domain like rural Mexican tourism requires an algorithmic adaptation architecture capable of altering the model's probability trajectory without catastrophic forgetting, all under extreme infrastructure constraints.

4.1. The Foundational Model: Meta Llama 3 8B

The computational core selected for this generative task is the Meta Llama 3 8-billion parameter model [5]. Three architectural innovations define its relevance for REST-MEX:

1. **High-Density Tokenizer:** A massively expanded vocabulary of 128,000 tokens [5]. This prevents specific Spanish words, indigenous localisms, or Nahuatl gastronomic terms from fragmenting into meaningless sub-tokens, increasing inference efficiency.
2. **Native Context Length:** Support for an 8,192-token context window ensures that extensive, discursive tourist reviews can be processed without truncation.
3. **Grouped-Query Attention (GQA):** GQA interpolates between Multi-Head Attention and Multi-Query Attention, retaining high inferential precision while drastically reducing the KV Cache memory consumption [5].

4.2. Hardware Constraints and Efficient Adaptation (QLoRA)

The primary operational bottleneck was infrastructural: executing all training and inference locally on an NVIDIA RTX 3060 limited to 12GB of VRAM. Full-parameter fine-tuning of an 8B model requires an absolute minimum of 84GB of VRAM.

To circumvent this limitation, the methodology relied on Quantized Low-Rank Adaptation (QLoRA) [11, 12]. QLoRA introduces a synergy of three critical memory reduction mechanisms:

1. **4-bit NormalFloat (NF4) Quantization:** Assuming pre-trained weights exhibit a normal distribution, NF4 maps these weights non-linearly to 16 discrete values, guaranteeing theoretically minimal information degradation when compressing foundational weights from 16 bits to 4 bits [11].
2. **Double Quantization:** Quantization constants are subjected to a second quantization step (from 32-bit to 8-bit), liberating additional VRAM [11].
3. **Paged Optimizers:** Dynamically transfers optimizer states between GPU VRAM and CPU RAM during gradient checkpoint spikes, preventing Out-of-Memory (OOM) failures [11].

4.3. Training Orchestration: Triton Kernels and Unsloth

The translation of QLoRA's theoretical framework into fast computational execution was orchestrated via the Unsloth framework. Unsloth aggressively rewrites low-level computations using custom Triton kernels. By utilizing fused kernels (for RoPE, RMSNorm, and SwiGLU), modifying Cross-Entropy math to drastically reduce logit memory, and applying uncontaminated sequence packing, Unsloth eliminates up to 75% of unnecessary memory traffic.

The SFT hyperparameters were meticulously calibrated. We applied a LoRA rank (r) of 16, targeting all linear modules (including GQA projections and MLP layers). The model was trained over 1250 batch steps using an 8-bit AdamW optimizer and a cosine learning rate scheduler. The progression confirmed the success of the joint architecture; by the end of the 1250 steps, a stable, consistent, and convergent decline in the scalar loss function was recorded, reflecting the correct absorption of the tourist jargon.

Figure 2: Progression of Training Loss over 1250 steps demonstrating algorithmic convergence.

5. Quantitative, Qualitative Analysis and Originality Verification

Following the SFT, the network was shifted to generation mode. To preserve human idiosyncrasy and avoid repetitive echoes, we implemented a parametrically controlled stochastic sampling decoding methodology.

Qualitative heuristic analysis exposes a fascinating mimicry achieved by the quantized vectors. The infrastructure overcame the inherent standardized formality of base LLMs, dynamically deploying undeniable assimilation of local Mexican narratives. Hostile ratings articulated incisive critiques tied to spatiotemporal variables, while superlative reviews captured the poetics of ancestral towns.

To extinguish any suspicion of data leakage or extractive memorization (plagiarism), every synthetic sequence generated was mathematically cross-referenced using the Jaccard Similarity Index against the 208,000 records of the mother repository. Adhering to methodological protocol, any string with a similarity index above a strict threshold of 0.7 was invalidated. The practically non-existent volume of

Table 2
Text Generation Hyperparameters

Parameter	Value
Samples per Class	200
Temperature	1.05
Top-p	0.95
Repetition Penalty	1.15
Max New Tokens	130,000

Table 3
Verbatim Examples of Synthetic Reviews Generated by the Fine-Tuned Model

Polarity	Generated Title	Generated Review Extract
1 Star	¡Totalmente olvidado!	Me llevé mi familia a cenar a este restaurant en Valladolid, Cenotes Zaci, porque queríamos comer platos típicos... El problema empezó desde el primer bocado en general todo estaba frió... Preguntales a mí para ser sinceramente voy hacer recomendaciones si he hecho una gran confianza al elevar mis asistentes con esta decepción de forma tan particular: si vendría a cenar allá? ¡Jamás!
5 Stars	¡Sorprendente experiencia en Valle de Bravo!	¡Me encantó mi reciente visita a este charmin pueblo mágico! La armonía entre naturaleza, arte y arquitectura hizo que me sintiera verdaderamente conectado con el lugar. Valle de Bravo es como si se tratara de un escenario perfecto para relajarse... Lo peñón es uno de los mejores lugares donde he estado jamás.

vetoed strings conclusively testifies that Llama 3 mastered the sociolinguistic constructs rather than becoming a robotic replicator.

6. Critical Error Analysis and Systemic Boundaries

The primary limiting operational barrier emerged from the hardware characteristics. Maneuvering an 8-billion parameter model within the 12GB limit of the RTX 3060 dramatically compressed the maximum sequence length and concurrent batch size, inevitably inflating the global time cost of training.

The second major limitation arose from the prophylactic undersampling required to cure mode collapse. By capping the samples at 2,000 blocks per polarity, we were forced into the painful digital exile of over 134,000 highly positive pre-existing reviews. This amputation silently blinded the parametric vectors to the enriched learning available in the long-tail semantic curve of those ignored files.

7. Carbon Emissions and Sustainable Technological Impact

Massive Transformer training structures historically imply massive kilowatt consumption via centralized corporate farms of H100s or A100s GPUs. Thanks to the technical achievements of QLoRA (NF4) and Unsloth optimization kernels, the entire iterative convergence cycle concluded successfully using basic home hardware (NVIDIA RTX 3060). This systemic decoupling demonstrates the democratization and global ecological profitability of deploying NLP applications tailored for Ibero-American linguistics.

8. Conclusions and Future Work

Addressing the structural challenges linked to the regional dialects of the Spanish language, particularly given the imbalanced volumetric spectrum, demands a concerted blend of theory and tactics. Responding

to the REST-MEX 2026 challenge, this work empirically validated an operatively revolutionary route to rectify statistical biases. By leveraging Meta Llama 3 under QLoRA and Unsloth packing, we fluidly generated organic, geographically-conditioned synthetic texts without mode collapse.

Future trajectories should imperatively interconnect toward advanced alignment paradigms such as Direct Preference Optimization (DPO) to holistically minimize grammatical hallucinations. Concurrently, embedding-based metrics should be utilized to mathematically quantify the lexical richness deployed within the synthetic corpus.

Declaration on Generative AI

During the preparation of this work, the authors used an AI assistant (Grmini 3.1 Pro) to support the formatting of the LaTeX manuscript in compliance with CEUR-WS guidelines and was used for the revision of translations into English, as well as for grammatical and spelling correction. Further, the authors utilized the Unsloth framework and a Llama model to perform the Supervised Fine-Tuning and subsequent generation of the synthetic text for the REST-MEX 2026 challenge. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and takes 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] 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.

[3] J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance, Journal of Big Data 6 (2019) 1–54.

[4] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 968–988.

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[6] J. D. Jimenez, J. E. Serrano, J. C. Martinez-Santos, E. Puertas, Verbanexai at mentalriskes 2025: Early detection of gambling disorders using transformer architectures and machine learning models, in: CEUR Workshop Proceedings, volume 4098, 2025.

[7] A. Morillo, D. Peña, J. C. Martinez-Santos, E. Puertas, Verbanexai lab at semeval-2024 task 1: A multilayer artificial intelligence model for semantic relationship detection, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 1334–1340.

[8] L. Mendoza, J. Suarez, E. Puertas, J. Martinez, J. Serrano, Cotecmar-utb at erisk 2025: Semantic-centroid symptom ranking and early depression detection using adaptive decision rule, in: CEUR Workshop Proceedings, volume 4038, 2025, pp. 1562–1582.

[9] D. Almanza-Gonzalez, J. E. Serrano, J. C. Martinez-Santos, E. Puertas, Prediction of human preferences and explanation generation with llm: An approach based on rag, few-shot learning, and auto-cot, in: CEUR Workshop Proceedings, volume 4038, 2025, pp. 1360–1369.

[10] A. Morillo, C. Agamez, E. Puertas, J. C. Martinez-Santos, J. Serrano, Pubmed api and llm-driven hybrid retrieval system for biomedical question answering, in: 2025 IEEE Colombian Caribbean Conference (C3), 2025, pp. 1–6.

[11] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QloRA: Efficient finetuning of quantized llms, Advances in Neural Information Processing Systems 36 (2024).

[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, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.