

Hierarchical Prompt Engineering for Synthetic Tourist Review Generation in REST-MEX 2026

Federico Sandoval^{1,*}

¹*Algiedi Solutions, Cholula de Rivadavia, Puebla, México*

Abstract

This paper presents a hierarchical prompt engineering strategy for synthetic tourist review generation in the REST-MEX 2026 Artificial Data Generation Track. Unlike conventional prompt-based approaches that rely on a single instruction, the proposed methodology decomposes the generation process into multiple specialized prompts that independently control the role of the language model, the tourism context, the target sentiment polarity, and several linguistic constraints. The resulting synthetic reviews are incorporated into the official REST-MEX 2026 training corpus to fine-tune the baseline sentiment classifier provided by the organizers. Experimental results demonstrate that structured prompt design can generate coherent and diverse synthetic reviews that improve the official evaluation metrics over the Majority Class baseline, highlighting prompt engineering as a lightweight and reproducible alternative for synthetic data generation in tourism sentiment analysis.

Keywords

Prompt Engineering, Rest-Mex, Synthetic Text Generation, Large Language Models, Sentiment Analysis, Tourism Reviews

1. Introduction

The rapid adoption of Large Language Models (LLMs) has significantly transformed the landscape of Natural Language Processing (NLP). Beyond their remarkable capabilities for text understanding, these models have demonstrated an unprecedented ability to generate coherent, context-aware, and linguistically diverse documents across multiple domains. This generative capacity has opened new research directions in data augmentation, where synthetic texts are used to enrich training corpora and improve the performance of downstream machine learning models.

Tourism represents one of the domains that has benefited most from recent advances in NLP. Millions of online reviews published on platforms such as TripAdvisor, Google Reviews, and Booking.com provide valuable information regarding customer satisfaction, destination image, service quality, and visitor experiences. Consequently, sentiment analysis has become a fundamental component of intelligent tourism systems, supporting recommendation engines, destination management, digital reputation monitoring, and decision-support applications [1, 2, 3, 4, 5, 6, 7, 8].

The REST-MEX evaluation campaign has progressively established itself as the principal benchmark for sentiment analysis in Mexican tourism reviews. Since its first edition, the shared task has evolved from recommendation and polarity classification toward increasingly challenging problems involving multitask learning, domain adaptation, and large-scale tourism datasets [9, 10, 11, 12]. The REST-MEX 2026 edition introduces a fundamentally different challenge by asking participants to improve sentiment classification through synthetic review generation rather than classifier optimization [13, 14].

Recent research has shown that prompt engineering constitutes an effective mechanism for adapting foundation models to domain-specific tasks without modifying their internal parameters. Instead of learning new representations through fine-tuning, carefully designed prompts exploit the knowledge already encoded within LLMs to perform reasoning, classification, summarization, and text generation. Previous studies have demonstrated that prompt formulation can substantially influence model behavior, particularly in subjective tasks such as sentiment analysis and opinion mining [15]. Nevertheless, most

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

✉ fsandoval@algiedi.com.mx (F. Sandoval)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

prompt engineering research in tourism has focused on inference tasks rather than synthetic corpus generation.

In this work, we extend the prompt engineering paradigm to the problem of synthetic review generation. Rather than employing a single instruction, we propose a hierarchical prompting strategy in which multiple complementary prompts independently control different characteristics of the generated opinion, including sentiment polarity, tourism context, writing style, and linguistic constraints. By decomposing the generation process into specialized instructions, the proposed methodology aims to produce reviews that are coherent, diverse, and faithful to the lexical characteristics of authentic Mexican tourism opinions.

The resulting synthetic reviews are incorporated into the official REST-MEX 2026 training corpus and subsequently used to fine-tune the baseline sentiment classifier provided by the organizers. This work investigates whether structured prompt engineering can produce synthetic data that contributes to sentiment classification while maintaining a lightweight, reproducible, and model-agnostic generation framework.

2. The REST-MEX Benchmark

The REST-MEX benchmark has become one of the principal evaluation frameworks for Natural Language Processing research in Mexican tourism. Since its introduction in 2021, the shared task has provided progressively larger and more challenging datasets, fostering the development of sentiment analysis techniques specifically designed for Spanish-language tourism reviews. Beyond serving as a competitive evaluation platform, REST-MEX has contributed standardized corpora, reproducible evaluation protocols, and common baselines that facilitate fair comparisons among different methodologies [9, 10, 11, 12, 13].

The first edition, REST-MEX 2021, focused primarily on recommendation systems and opinion analysis over Mexican tourism reviews. Participants were challenged to predict customer satisfaction using textual information collected from online review platforms, establishing the first publicly available benchmark dedicated to tourism-oriented NLP in Mexican Spanish [9].

The benchmark evolved considerably during REST-MEX 2022. In addition to recommendation and sentiment analysis tasks, the competition incorporated new challenges related to the automatic prediction of the Mexican COVID-19 epidemiological traffic-light system from news articles. This extension demonstrated the flexibility of the benchmark and encouraged the development of more general-purpose Spanish-language text classification systems [10, 16, 17].

REST-MEX 2023 marked another important milestone by concentrating on fine-grained sentiment analysis over tourism reviews while expanding the diversity of participating systems. During this edition, transformer-based architectures, domain adaptation strategies, ensemble learning, and corpus balancing techniques became the dominant approaches, substantially improving the state of the art for sentiment analysis in tourism [11, 18, 19, 20, 21, 22, 23].

The 2025 edition significantly increased the complexity of the benchmark by introducing a large-scale corpus containing more than 250,000 reviews collected from forty Mexican *Pueblos Mágicos*. Besides sentiment polarity, participants were required to classify the type of tourism service and identify the town associated with each review, transforming the competition into a multi-task learning problem with highly imbalanced data distributions [12].

REST-MEX 2026 introduced a substantially different research perspective through the Artificial Data Generation Track. Instead of optimizing classification architectures, participants were required to generate up to 3,000 synthetic tourist reviews that could be incorporated into the original training corpus. The effectiveness of each submission was evaluated indirectly by fine-tuning a common baseline sentiment classifier using the generated reviews and measuring its performance on a hidden test set. This formulation shifted the focus of the competition from model design toward data-centric Artificial Intelligence, emphasizing the quality, diversity, and usefulness of synthetic training instances rather than classifier architecture [13, 14].

The official REST-MEX 2026 training corpus contains 208,051 manually annotated tourist reviews

covering forty Mexican *Pueblos Mágicos*. Each opinion includes a title, the review text, a five-level sentiment polarity label, the corresponding town, its geographical region, and the type of tourism service. The corpus presents a severe class imbalance, with highly positive reviews representing the majority of the dataset, making synthetic data generation particularly relevant for improving the representation of minority sentiment categories.

Through its continuous evolution, REST-MEX has progressively reflected the major trends in Natural Language Processing research. The benchmark has transitioned from traditional supervised sentiment classification toward multitask learning, domain adaptation, transformer-based architectures, and, more recently, data-centric AI through synthetic text generation. Consequently, it provides an ideal experimental framework for evaluating novel strategies aimed at improving sentiment analysis systems for Spanish-language tourism reviews.

3. Proposal

The central hypothesis of this work is that the quality of synthetic reviews depends not only on the language model employed, but also on the structure of the instructions provided to the model. Rather than relying on a single generic prompt, we propose a hierarchical prompt engineering strategy in which multiple complementary instructions independently control different characteristics of the generated reviews. This decomposition allows the language model to simultaneously satisfy semantic, stylistic, and task-specific constraints while maintaining coherent and diverse outputs.

Let $D = (x_i, y_i)_{i=1}^N$ denote the original REST-MEX 2026 training corpus, where x_i represents a tourist review and $y_i \in 1, 2, 3, 4, 5$ its corresponding sentiment polarity. The objective is to generate a synthetic corpus $D' = (x' * i, y_i) * i = 1^M$, with $M \leq 3000$, that complements the original dataset while preserving the target sentiment labels.

Instead of directly querying the language model with a single instruction, each synthetic review is generated by composing a sequence of specialized prompts. Formally, the complete prompt is represented as

$$P = P_s \oplus P_c \oplus P_l \oplus P_r,$$

where P_s defines the system behavior, P_c provides contextual information, P_l specifies the desired sentiment polarity, and P_r introduces generation constraints. The operator $\oplus$ denotes the concatenation of prompt components before being submitted to the language model.

3.1. System Prompt

The first component establishes the role adopted by the language model. Instead of acting as a generic text generator, the model is instructed to behave as an experienced tourist writing authentic online reviews in Mexican Spanish. The prompt explicitly emphasizes natural language, realistic descriptions, and the writing style commonly observed in platforms such as TripAdvisor or Google Reviews.

3.2. Context Prompt

The contextual prompt specifies the type of establishment being reviewed, such as a hotel, restaurant, or tourist attraction. This information encourages the model to generate domain-specific vocabulary and descriptions appropriate for each tourism service.

Let $c \in \text{Hotel, Restaurant, Attractive}$ denote the selected context. The generated review is therefore conditioned on both the sentiment label and the tourism category.

3.3. Sentiment Prompt

The sentiment prompt explicitly controls the polarity of the generated opinion. Given a target sentiment $y \in 1, 2, 3, 4, 5$, the language model receives instructions describing the expected emotional intensity

associated with that polarity level.

This conditioning process can be represented as

$$x' = G(P, y, c),$$

where $G(\cdot)$ denotes the language model responsible for generating the synthetic review.

By explicitly specifying the desired sentiment before generation, the model is encouraged to produce reviews consistent with the target polarity instead of relying solely on implicit contextual cues.

3.4. Generation Constraints

The final prompt introduces a set of linguistic and semantic constraints intended to improve the quality of the generated corpus. The language model is instructed to:

- write exclusively in Mexican Spanish;
- generate realistic tourist experiences;
- avoid copying or paraphrasing reviews from the original dataset;
- produce reviews of moderate length;
- maintain consistency between the expressed sentiment and the assigned polarity;
- avoid repetitive expressions across different generated reviews.

These constraints reduce lexical redundancy while promoting greater diversity among the synthetic instances.

3.5. Synthetic Corpus Construction

The previous prompt sequence is executed repeatedly until the desired number of synthetic reviews is obtained. Every generated review inherits the sentiment polarity specified during prompt construction, eliminating the need for additional annotation.

Finally, the synthetic corpus D' is merged with the original training dataset,

$$D^* = D \cup D',$$

and the resulting corpus is used to fine-tune the official baseline sentiment classifier provided by the REST-MEX 2026 organizers.

Unlike conventional prompt engineering approaches that rely on a single instruction, the proposed methodology decomposes the generation process into multiple specialized prompts. This modular design simplifies prompt refinement, increases control over the generated content, and facilitates the production of balanced and coherent synthetic reviews tailored to the requirements of the REST-MEX benchmark.

4. Results and Discussion

The proposed hierarchical prompt engineering strategy was evaluated following the official protocol of the REST-MEX 2026 Artificial Data Generation Track. As in the other participating systems, the effectiveness of the generated corpus was assessed indirectly by fine-tuning the baseline sentiment classifier released by the organizers and evaluating its performance on the hidden test set. The official evaluation relies primarily on the *Ranking Performance* (RP) metric, which emphasizes minority sentiment classes, together with Macro F1, overall Accuracy, and the F1-score obtained for each individual polarity level.

The proposed methodology obtained an RP score of 0.0138 and a Macro F1 score of 0.0359. The overall evaluation score reached 11.9601, while the per-class F1 scores were 0.0005, 0.0441, 0.0005, 0.0008, and 0.1335 for polarity classes 1 through 5, respectively. Although the improvements remain moderate,

these results demonstrate that carefully designed prompts are capable of generating synthetic reviews that contribute useful information during the fine-tuning stage.

A detailed inspection of the per-class results reveals that the best performance was achieved for polarity class 5, obtaining an F1-score of 0.1335. This behavior is expected, since highly positive opinions dominate the original REST-MEX corpus, allowing the language model to better capture the linguistic characteristics associated with this sentiment category. Polarity class 2 also achieved a measurable improvement, with an F1-score of 0.0441, whereas classes 1, 3, and 4 remained considerably more difficult to identify. These observations indicate that prompt engineering alone cannot completely compensate for the severe imbalance present in the training data.

One of the principal advantages of the proposed methodology lies in the explicit decomposition of the generation process. Rather than relying on a single generic instruction, the prompt is divided into independent components responsible for defining the role of the language model, the tourism context, the target sentiment, and a collection of linguistic constraints. This modular organization provides greater control over the generated content and facilitates systematic prompt refinement without modifying the underlying language model.

Another important observation concerns the consistency of the generated reviews. During manual inspection, the hierarchical prompting strategy produced opinions that exhibited coherent sentiment, realistic tourism vocabulary, and natural writing styles characteristic of online review platforms. The explicit constraints included in the prompts also reduced repetitive expressions and encouraged greater lexical variability across generated instances, producing a synthetic corpus that closely resembles authentic tourist opinions.

Despite these advantages, the quantitative improvements remain relatively limited. This suggests that the quality of synthetic data depends not only on prompt design but also on the intrinsic knowledge encoded within the language model and the diversity of the generated corpus. While hierarchical prompting improves generation control, it cannot entirely eliminate semantic redundancy or guarantee balanced coverage of minority sentiment classes.

Overall, the experimental results indicate that prompt engineering constitutes an effective and computationally inexpensive strategy for synthetic review generation. The proposed hierarchical formulation provides a structured framework for controlling multiple aspects of the generation process and demonstrates that carefully engineered prompts can produce synthetic data suitable for data-centric sentiment analysis. These findings also suggest that prompt engineering can serve as a complementary component within more sophisticated generation pipelines that combine retrieval, lexical augmentation, or iterative refinement techniques.

5. Conclusions

This paper presented a hierarchical prompt engineering strategy for synthetic tourist review generation in the REST-MEX 2026 Artificial Data Generation Track. By decomposing the generation process into multiple specialized prompts, the proposed methodology provides explicit control over sentiment polarity, tourism context, writing style, and linguistic constraints while requiring no model fine-tuning.

The experimental results demonstrate that structured prompt design can generate synthetic reviews that positively contribute to the training process of sentiment classifiers. Although the overall improvements remain modest, the proposed framework offers a simple, reproducible, and model-agnostic alternative for synthetic data generation based entirely on prompt engineering.

Future work will investigate adaptive prompt optimization, automatic prompt refinement, and hybrid generation strategies that combine hierarchical prompting with retrieval-augmented generation and iterative self-evaluation to produce more diverse and balanced synthetic corpora.

Declaration on Generative AI

Generative AI tools were used solely to improve the language, grammar, and readability of the manuscript. The authors reviewed and edited all AI-assisted content and take full responsibility for the final version of the manuscript.

References

- [1] M. Á. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-Gonzalez, D. Fajardo-Delgado, M. G. Sánchez, H. Pérez-Espinosa, J. Martínez-Miranda, R. Guerrero-Rodríguez, L. Bustio-Martínez, Á. Díaz-Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, *Journal of king Saud university-computer and information sciences* 34 (2022) 10125–10144.
- [2] A. Diaz-Pacheco, M. Á. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, R. Aranda, Artificial intelligence methods to support the research of destination image in tourism. a systematic review, *Journal of Experimental & Theoretical Artificial Intelligence* 36 (2024) 1415–1445.
- [3] R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico, *Current issues in tourism* 26 (2023) 289–304.
- [4] R. Guerrero-Rodríguez, M. A. Álvarez-Carmona, R. Aranda, Á. Díaz-Pacheco, Ai-driven analysis of hotel digital reputation: insights from a mexican destination, *International Journal of Tourism Cities* 12 (2026) 19–36.
- [5] E. P. Ramirez-Villaseñor, H. Pérez-Espinosa, M. A. Álvarez-Carmona, R. Aranda, Design, development, and evaluation of a chatbot for hospitality services assistance in spanish, *Acta universitaria* 33 (2023).
- [6] S. Arce-Cardenas, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, J. P. Ramírez-Silva, A tourist recommendation system: a study case in mexico, in: *Mexican international conference on artificial intelligence*, Springer, 2021, pp. 184–195.
- [7] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, R. Aranda, A comprehensive deep learning approach for topic discovering and sentiment analysis of textual information in tourism, *Journal of King Saud University-Computer and Information Sciences* 35 (2023) 101746.
- [8] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, R. Aranda, Quantifying differences between ugc and dmo's image content on instagram using deep learning, *Information Technology & Tourism* 26 (2024) 293–329.
- [9] 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. Y. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism (2021).
- [10] 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.
- [11] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, V. Muñoz-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.
- [12] 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, *Procesamiento del Lenguaje Natural* 75 (2025) 499–511.
- [13] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, 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).
- [14] 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.
 - [15] J. Stack-Sánchez, M. Álvarez-Carmona, A. P. L. Monroy, Nlp_cimat at semeval-2025 task 3: Just ask gpt or look inside. a prompt and neural networks approach to hallucination detection, in: *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, 2025, pp. 1683–1689.
 - [16] M. Á. Álvarez-Carmona, R. Aranda, Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias (2022).
 - [17] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, *Journal of Information Science* 50 (2024) 568–589.
 - [18] V. G. Morales-Murillo, H. Gómez-Adorno, D. Pinto, I. A. Cortés-Miranda, P. Delice, Lke-iimas team at rest-mex 2023: Sentiment analysis on mexican tourism reviews using transformer-based domain adaptation (2023).
 - [19] J. D. Jurado-Buch, S. Minayo-Díaz, J. Tello, K. Chaucanes, L. Salazar, M. Oquendo-Coral, M. Á. Álvarez-Carmona, A single model based on beto to classify spanish tourist opinions through the random instances selection, 2023.
 - [20] O. G. Toledano-López, M. Á. Álvarez-Carmona, J. Madera, A. Simón-Cuevas, Y. A. López-Rodríguez, H. González Díez, Polarity prediction in tourism cuban reviews using transformer with estimation of distribution algorithms, in: *International Workshop on Artificial Intelligence and Pattern Recognition*, Springer, 2023, pp. 335–346.
 - [21] A. B. García-Gutiérrez, P. E. López-Ávila, P. A. Gallegos-Ávila, R. Aranda, M. Á. Álvarez-Carmona, Balancing of tourist opinions for sentiment analysis task., in: *IberLEF@ SEPLN*, 2023.
 - [22] M. Á. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-González, A. P. López-Monroy, A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one, *Computación y Sistemas* 26 (2022) 977–987.
 - [23] A. Díaz-Pacheco, A. B. García-Gutiérrez, A. Y. Rodríguez-González, R. Aranda, M. Á. Álvarez-Carmona, Balancing sentiment analysis datasets through representative-word-guided synthetic review generation: A case study on mexican spanish tourism reviews, *Applied Sciences* 16 (2026) 7398.