LSM-INAOE at IberLEF 2026: Bidirectional Translation Between Mexican Sign Language Glosses and Spanish Using Llama 3.2

Jesús Joshua Muñoz-Pacheco[1,2,3,*,†], Alejandro Antonio Torres-García[1,2,3,†] and Jesús Ricardo López-Gutiérrez[2,†]

[1]*Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro No. 1, San Andrés Cholula, Puebla, 72840, Mexico*

[2]*SECIHTI-INAOE, Luis Enrique Erro No. 1, San Andrés Cholula, Puebla, 72840, Mexico*

[3]*Laboratorio de Procesamiento de Bioseñales y Computación Médica, INAOE, Luis Enrique Erro No. 1, San Andrés Cholula, Puebla, 72840, Mexico*

Abstract

From a social perspective, the Deaf community in Mexico faces significant communication barriers when interacting with the hearing majority, who often lack the ability to interpret Mexican Sign Language (LSM). Conversely, LSM users frequently encounter difficulties understanding standard written Spanish due to its fundamentally distinct morphosyntactic structure. Advancements in bidirectional translation systems are crucial to bridging this communication gap and fostering true social inclusion. This article describes the LSM-INAOE team's participation in the shared MSLG-SPA task at IberLEF 2026, which focuses on the challenge of bidirectional translation between LSM glosses and Spanish. Our approach leverages a lightweight and optimized language model, specifically Llama 3.2 with approximately three billion parameters. Using this approach, we addressed two tasks: the translation of LSM glosses into natural Spanish (MSLG2SPA) and the reverse process of generating structured glosses from Spanish sentences (SPA2MSLG). Through strategic instructional design, it is possible to obtain favorable metrics for the interpretation and generation of glosses in LSM, even using small language models with a number of parameters that may be suitable for local systems with limited processing capacity. In this regard, our bidirectional system demonstrated highly consistent performance in both directions. In the official evaluations, our system obtained a BLEU score of 32.1097 and a chrF score of 60.5358 for the MSLG2SPA task, and a BLEU score of 13.8826 with a chrF score of 55.3879 for the SPA2MSLG task, placing LSM-INAOE in seventh place in the World Ranking out of 18 participating teams.

Keywords

Mexican Sign Language, Glosses translation, Large Language Models, Bidirectional Translation, Llama 3.2, IberLEF 2026

1. Introduction

Globally, the World Health Organization estimates that approximately 430 million people face some degree of disabling hearing loss [1]. This figure represents a population that encounters not only clinical challenges but also a structural communication barrier that has historically limited their access to fundamental rights. In Mexico, it is estimated that around 2.4 million people present severe hearing difficulties, a significant portion of whom utilize Mexican Sign Language (Lengua de Señas Mexicana, LSM) as their primary means of interaction [2].

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

jjoshua.mpacheco@inaoep.mx (J. J. Muñoz-Pacheco); alejandro.torres@ccc.inaoep.mx (A. A. Torres-García); jrlopez@inaoe.mx (J. R. López-Gutiérrez)

https://sites.google.com/site/alejandroantoniotorresgarcia (A. A. Torres-García); https://www-elec.inaoep.mx/directorio-de-la-coordinacion/investigadores/jesus-ricardo-lopez-gutierrez (J. R. López-Gutiérrez)

0009-0001-6970-0836 (J. J. Muñoz-Pacheco); 0000-0001-5091-0764 (A. A. Torres-García); 0000-0001-5326-5342 (J. R. López-Gutiérrez)

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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-09-29

CEUR-WS.org/Vol-4273/MSLG-SPA2026_paper10.pdf

Mexican Sign Language, like other visual-gestural languages, utilizes "glosas" (glosses) as foundational units. These glosses capture semantic concepts without incorporating the morphological structures or word order rules of spoken Spanish. Consequently, the correspondence between these two languages presents a complex challenge, since generating glosses involves not only generating the sign concept but also its method of representation, which can be a combination of symbols, the manual spelling of a specific word, among others. Furthermore, the interpretability of the glosses entails the morphological challenge of considering the spatial grammar of LSM, which can lead to misinterpretations for models that do not take it into account.

This paper details the system developed by the LSM-INAOE team for the MSLG-SPA shared task [3], evaluated as part of the Iberian Languages Evaluation Forum (IberLEF 2026) [4]. The challenge primarily consists of two independent but complementary translation tracks:

- **MSLG2SPA:** This subtask requires the translation of LSM gloss sequences into grammatically correct and fluid written Spanish.
- **SPA2MSLG:** This subtask consists of generating structured LSM glosses based on standard Spanish text inputs.

Our team approached these tasks by deploying Llama 3.2, an open-weight Large Language Model (LLM) [5]. This specific model architecture was chosen because, while larger language models exist and often achieve marginal gains in performance metrics, they rely heavily on high-cost cloud computing infrastructures that demand persistent internet connectivity. In contrast, our approach aims to demonstrate that robust bidirectional translation can be achieved efficiently on resource-constrained setups, paving the way for localized, low-latency deployment scenarios that do not depend on remote server dependencies.

To contextualize the hardware requirements for a localized implementation, it is essential to describe the technical profile of the target execution environments. Thanks to the parameter quantization strategy applied to the 3-billion-parameter model, the total memory consumption is drastically reduced, requiring approximately 2 to 3 gigabytes of VRAM (video memory) for efficient inference. This compact size makes the system highly viable for execution on embedded edge computing devices, such as the Jetson TX2 or Jetson AGX Xavier embedded systems. We believe these platforms are exceptionally well-suited for our system, as they integrate low-power GPUs compatible with CUDA, which has become a standard in the literature for running this type of model. This design allows the hardware to handle the necessary tensor operations and load the quantized model directly into the GPU's accessible memory, avoiding the prohibitive cost, size, and power consumption of conventional desktop workstations. Consequently, implementation in these integrated architectures would facilitate the creation of portable, real-time translation modules capable of operating autonomously in different environments without relying on external servers.

2. Related Work

Bidirectional machine translation between Spanish and Mexican Sign Language (LSM) glosses is formally defined within Natural Language Processing (NLP) as a text-to-text translation challenge. This task faces severe morphosyntactic challenges due to profound structural asymmetries between both languages [6]. While Mexican Spanish predominantly employs a Subject-Verb-Object (SVO) sentence structure along with a rich inflectional morphology, LSM glosses frequently adopt flexible order variants or Subject-Object-Verb (SOV) configurations. Furthermore, gloss sequences significantly compress sequence length by omitting articles, prepositions, and traditional grammatical markers for time and gender [7]. Over recent years, computational systems modeling this textual mapping have evolved from rigid rule-based paradigms to state-of-the-art deep generative models.

2.1. Rule-Based Machine Translation (RBMT) and Grammatical Knowledge

In the early absence of large parallel corpora, initial computational approaches for translating Spanish text to LSM glosses relied heavily on Rule-Based Machine Translation (RBMT) frameworks [8]. The architecture of these methods begins with a text preprocessing stage encompassing character cleaning, spelling validation, and Part-of-Speech (POS) tagging via standard natural language analyzers such as spaCy or FreeLing [8, 9].

Once parsed, the translation core evaluates the input sentence against predefined hierarchical syntactic trees categorized into simple, intermediate, and complex patterns [8]. If a structural match is detected, direct grammatical transfer rules are triggered: redundant phonetic and functional elements (such as articles and verb inflections) are suppressed, and verbs are lemmatized to their infinitive forms. Concurrently, support gloss injection rules are applied; for instance, pluralization is achieved by appending the gloss "MUCHOS" adjacent to the noun, and gender is explicitly denoted by concatenating "MUJER" over base lemmas [8, 9]. Other systems incorporated lexical reference lemmatizers, such as TreeTagger, as baseline approaches; however, standalone lexical root extraction proved insufficient due to the lack of a mathematical mechanism capable of managing the complex syntactic reordering of LSM [10].

2.2. Transformer Architectures and Transfer Learning

The consolidation of the Transformer architecture established a definitive milestone in Sign Language Processing (SLP) by utilizing multi-head self-attention mechanisms to capture long-range syntactic dependencies without recurrence [11]. Initial attempts to train Transformer models from scratch using LSM glosses failed due to performance collapse during the validation phase. This barrier was overcome following the compilation of the first structured parallel corpus consisting of 3,000 paired Spanish sentences and LSM glosses, enabling the application of Transfer Learning methodologies [12].

Recent studies have adapted powerful models pre-trained on the Spanish language, such as *Helsinki-NLP* and *BARTO* (a bidirectional auto-regressive model featuring 139 million parameters), tailoring them to visual-gestural syntax [12]. During fine-tuning, the model assimilates pronoun elision and morphological inflections, learning to restructure sentence token sequences into native gloss order. Optimized via Native Automated Mixed Precision (AMP) and linear learning rate decay, fine-tuned BARTO variants have achieved state-of-the-art metrics for LSM translation, yielding minimal convergence losses (0.0118) and stable BLEU scores ranging between 82.26 and 83.68 [12].

2.3. Cross-Lingual Data Augmentation and Refinement via LLMs

In the most recent frontier of research, modeling techniques have targeted the dimensional constraints of available data through Cross-Lingual Data Augmentation strategies. The underlying hypothesis dictates that, despite regional lexical divergences, distinct sign languages share universal patterns within their textual gloss representations, such as high syllabic compression and uninflected lemma representations [10].

To evaluate this, cross-training regimes were implemented by feeding a baseline BARTO model mixed dataset pairs: the native Spanish-LSM corpus (3,000 sentences) combined with a filtered subset of English-ASL (American Sign Language) pairs drawn from the ASLG-PC12 corpus [10]. By oversampling local native data, the syntactic knowledge of the high-resource source language (ASL) operated as a morphosyntactic regularizer. This strategy vastly improved the translation engine's performance, scaling BLEU-1 metrics from 62 to 85 and reducing the Translation Edit Rate (TER) from 44 down to 20 [10]. Parallel to these developments, indirect generation architectures have begun exploring multi-stage decoding frameworks that utilize larger text-only Large Language Models (LLMs), such as LLaMA3, to generate and refine gloss sequences without relying exclusively on small task-specific models.

3. Methodology

This section describes the system design utilized for the IberLEF shared task. The proposed pipeline solves the morphosyntactic asymmetries inherent between both linguistic modalities through instruction fine-tuning and low-rank parameter optimization, omitting extensive mathematical formalism in favor of implementation specifics.

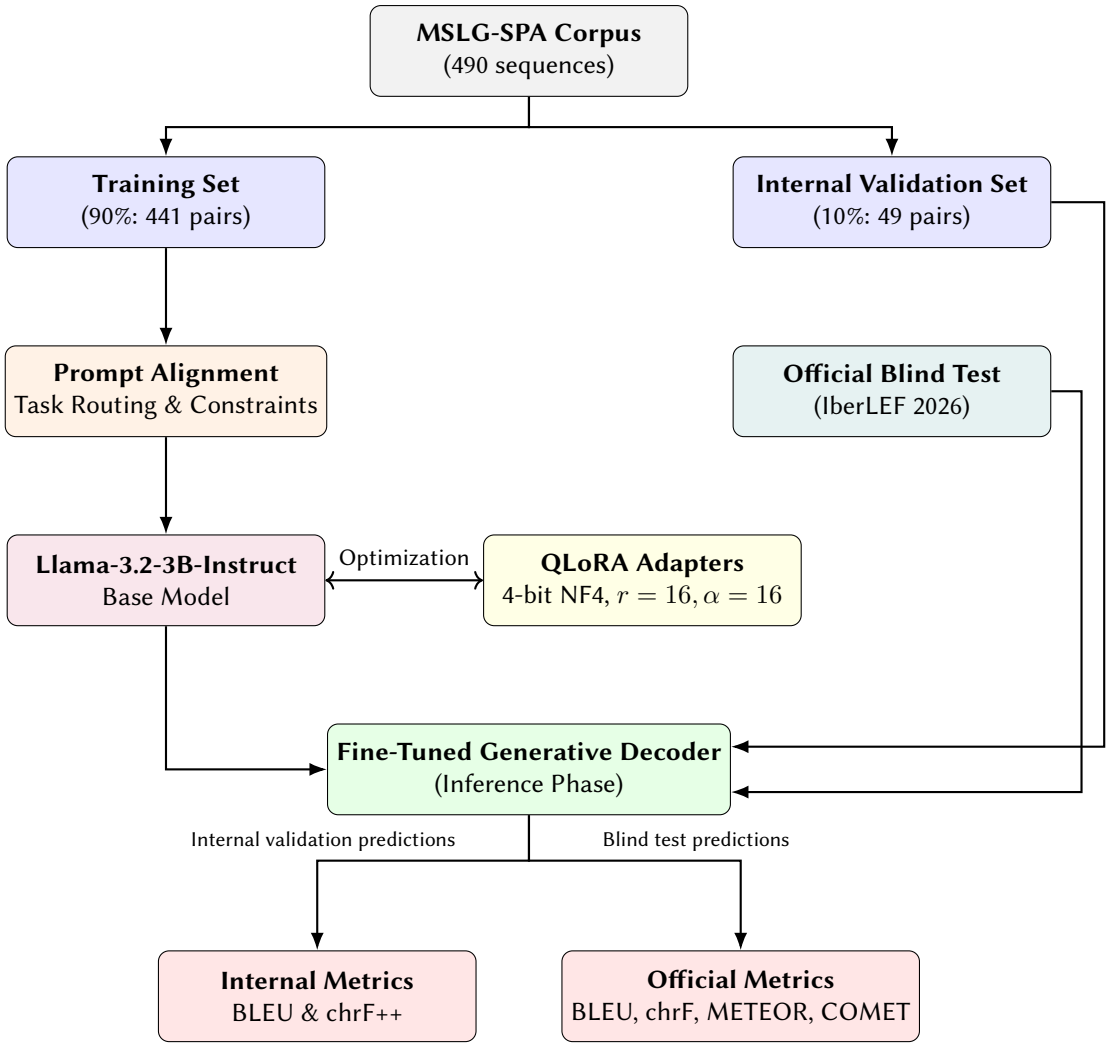

Figure 1: System architecture and optimization pipeline: demonstrating data splitting, QLoRA fine-tuning, and the separated inference flows for both internal validation and official blind testing.

3.1. Dataset and Splitting Methodology

The training and internal evaluation of the system were based on the bilingual parallel corpus provided by the MSLG-SPA 2026 task organizing committee. The training database consists of a total of 490 valid instances, comprising natural Spanish sentences aligned with their respective Mexican Sign Language (LSM) gloss representations.

To ensure a rigorous evaluation of the model's performance during the development phase and to mitigate the risk of overfitting, a strict hold-out partitioning scheme with an exact 90/10 ratio was implemented. The separation was executed pseudo-randomly by fixing an algorithmic control seed (`random_state = 42`). As a result, the corpus was divided into:

- **Training Set:** Composed of 441 sequence pairs (90% of the corpus). This data block was used exclusively for fine-tuning the model's adapters, operating independently for each translation

direction.

- **Internal Validation Set:** Composed of the remaining 49 pairs (10% of the corpus). This subset was strictly isolated from the optimization process, operating as a local blind test set for empirical control metrics.

3.2. Model Architecture and Fine-Tuning Strategy

The core enabling the interpretation and generation of glosses and coherent Spanish text is the autoregressive `Llama-3.2-3B-Instruct` model. To optimize computational efficiency and enable robust training within restricted Video RAM (VRAM) environments, we applied a 4-bit *NormalFloat4* (NF4) parameter quantization coupled with Low-Rank Adaptation (QLoRA).

During the adaptive optimization process, the base weights of the quantized network (W_0) were frozen. Instead of updating all parameters, updatable low-rank decomposition matrices (A and B) were injected into the architecture. The weight modification is parameterized as $\Delta W = \frac{\alpha}{r}(B \cdot A)$, where r denotes the intrinsic rank of the decomposition and α is a scaling constant. For this implementation, we empirically set $r = 16$ and $\alpha = 16$. These adapters were coupled to all essential linear projection layers of the attention mechanism and the Feed-Forward Network:

$$\mathcal{M}_{\text{target}} = \{q_{\text{proj}}, k_{\text{proj}}, v_{\text{proj}}, o_{\text{proj}}, \text{gate}_{\text{proj}}, \text{up}_{\text{proj}}, \text{down}_{\text{proj}}\}$$

System training operated under a supervised causal approach. Since the choice of the objective function can vary depending on the generative architecture, it is pertinent to specify that our model minimizes the Cross-Entropy Loss ($\mathcal{L}_{\text{CE}}$). To prevent penalizing the instructions, the loss was calculated exclusively over the positions corresponding to the actual output sequence tokens:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{V} y_{i,j} \log \hat{y}_{i,j}$$

where N represents the length of valid tokens in the response, V is the total dimension of the tokenizer's vocabulary, $y_{i,j}$ is a binary indicator of the actual token's presence, and $\hat{y}_{i,j}$ denotes the probability assigned by the network's Softmax distribution.

The convergence process was optimized using the 8-bit AdamW algorithm. The hyperparameter tuning of the implemented pipeline is synthesized in Table 1.

Table 1
Configuration and Optimization Parameters of the Causal Pipeline.

Hyperparameter	Assigned Value
Base Learning Rate (η)	2×10^{-4}
Learning Rate Scheduler	Linear with Warmup
Warmup Steps	15
Max Training Steps	300
Local Batch Size	2
Gradient Accumulation Steps	4
Global Effective Batch Size	8
Computational Optimizer	AdamW 8-bit
Hardware Accelerated Precision	bfloat16 / fp16

As detailed in Table 1, the selection of these configuration parameters was established to achieve an optimal balance between empirical convergence stability and the strict memory limitations of the hardware. Specifically, a global effective batch size of 8 was chosen to fully accommodate the limited VRAM capabilities of edge devices. A learning rate of 2×10^{-4} with 15 warm-up steps was strategically defined to mitigate early gradient spikes. Finally, a training step limit of 300 was experimentally determined to ensure domain adaptation while avoiding overfitting.

3.3. Prompt Alignment and Linguistic Constraint Injection

In our methodology, the interaction with the language model is structured as a clear, step-by-step communication process. Rather than simply feeding raw text into the network, we use carefully designed prompts to define the model's behavior, establish its role, and isolate the two distinct translation tasks. This structural alignment relies on Llama 3's native instruction tags (such as <|start_header_id|> and <|eot_id|>) to clearly separate the system instructions, the input data, and the expected output.

3.3.1. Subtask A: Glosses to Spanish (MSLG to SPA)

The purpose of this prompt is to guide the model in expanding condensed concepts into fluid language. The process follows these steps:

1. **Role Assignment:** The system is initially instructed to act as an expert translator to set the contextual baseline.
2. **Input Processing:** The system receives a sequence of LSM glosses. Recognizing the gloss as a precise technical linguistic unit rather than a simple word equivalent, the model must interpret its spatial and semantic weight.
3. **Output Generation:** The model translates these abstract units into a grammatically coherent and natural written Spanish sentence.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
Eres un traductor experto de Lengua de Señas Mexicana (LSM). Traduce la glosa
proporcionada a un español natural y fluido.<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Glosa: {MSLG}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Traducción:
```

3.3.2. Subtask B: Spanish to Glosses (SPA to MSLG)

For the reverse task, the prompt serves as a strict regulatory mechanism. The translation pipeline is defined as follows:

1. **Constraint Injection:** The system is assigned the expert role, but with high-priority restrictions. It is explicitly commanded to drop standard Spanish orthographic features (such as tildes) to match the formal notation of sign language linguistics.
2. **Input Processing:** The system receives a natural, conversational Spanish sentence.
3. **Output Generation:** The model must strip the conversational grammar and output a highly structured, exact sequence of glosses.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
Eres un traductor experto de español a Lengua de Señas Mexicana (LSM). Traduce
el texto en español proporcionado a su representación exacta en glosa,
respetando estrictamente las convenciones de anotación (guiones, signos de
suma, prefijos dm- y #). No utilices tildes ni acentos ortográficos en las
glosas generadas.<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Español: {SPA}<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Glosa:
```

This structure ensures the exact preservation of the following fundamental linguistic operators of LSM:

- **Hyphen (-):** Coheres a complex semantic concept expressed through a single spatial sign (e.g., YA-VEO).

- **Plus Sign (+):** Defines the contraction or construction of a compound term (e.g., MAMÁ+PAPÁ to conceptually denote parents).
- **Number Sign (#):** Identifies a lexical borrowing derived from highly rapid and lexicalized fingerspelling (e.g., #TV).
- **Fingerspelling prefix (dm-):** Explicitly indicates the descriptive letter-by-letter manual spelling of proper nouns or homonyms (e.g., dm-JUAN).

3.4. Continuous Evaluation Statistical Metrics

The internal validation of the system and its subsequent algorithmic contrast are supported by standard machine translation metrics:

- **BLEU (Bilingual Evaluation Understudy):** Quantitatively evaluates the modified n-gram precision of the test sequence against human validation references, incorporating a brevity penalty to counteract artificially short sequences.
- **chrF++:** Because glosses lack traditional inflectional grammatical structures and are significantly affected by minor typographical variations, the chrF++ metric was additionally implemented. This metric fuses character and word n-grams using the F-score statistic. A parametric weight of $\beta = 2$ was assigned to favor recall, making it highly efficient for capturing specific morphemes and complex structural prefixes like dm-.

4. Experiments and Results

The performance of the proposed bidirectional system was evaluated using the official blind test set provided by the MSLG-SPA 2026 organizers. The evaluation framework employed standard automated metrics for machine translation tasks, primarily BLEU, chrF, METEOR, and COMET.

4.1. Hardware Environment and Computational Cost

To assess the computational feasibility of our framework, the instruction fine-tuning process was run on a standard cloud infrastructure using a single NVIDIA Tesla T4 GPU with 14.75 GB of GDDR6 VRAM, provisioned via Google Colab. Thanks to the efficiency of 4-bit NormalFloat4 (NF4) quantization combined with QLoRA, the network exhibited highly optimized memory consumption, requiring a maximum allocation of only 7.2 GB of VRAM during the training phase. The convergence threshold was reached quickly, with the optimization process requiring approximately 5 minutes and 36 seconds per translation direction for 300 steps. This computational profile confirms that the model's memory requirements and training time are significantly reduced compared to larger, non-quantized architectures.

4.2. Official Results

Table 2 presents the quantitative results of the first run submitted by the LSM-INAOE team for both tracks.

Table 2
Official Preliminary Performance of the LSM-INAOE System.

Task / Direction	BLEU	chrF	METEOR	COMET
MSLG2SPA (Glosses → Spanish)	32.1097	60.5358	0.5290	0.8280
SPA2MSLG (Spanish → Glosses)	13.8826	55.3879	0.3796	–

In the competitive landscape, our system achieved the 7th position in the Global Ranking among 18 participating teams, with an aggregated global score of 0.2934. When breaking down the performance

by subtask, the model secured the 8th position among 20 submissions in the MSLG2SPA category, and the same 8th position among 18 submissions in the highly complex SPA2MSLG category.

4.3. Discussion

4.3.1. Bidirectional Symmetry Analysis

A fundamental conclusion from these results is the symmetric robustness of the adapted Llama-3.2-3B architecture. Historically, the effectiveness and performance of machine translation systems vary drastically depending on the task direction: Spanish-to-glosses versus glosses-to-Spanish conversion. This bidirectional variation stems from profound information density asymmetries between both languages.

Specifically in Mexican Sign Language (LSM), this historical disparity is even more evident in the metric evaluation literature. The vast majority of quantitative and optimization efforts have concentrated on the Spanish $\rightarrow$ Glosses (T2G) direction. In this track, fine-tuned Transformer models enhanced with cross-lingual data augmentation have documented outstanding metrics, with BLEU scores scaling up to 85 and Translation Edit Rates (TER) reduced to 20 [10].

Conversely, algorithmic effectiveness in the LSM Glosses to Spanish direction has historically lacked robust quantitative reporting through small dedicated modelsl. To resolve this expansive direction without suffering from hallucinations or grammatical errors, recent research has begun to bypass compact bidirectional architectures in favor of multi-stage generative decoders. Our strategy of adapting a resource-constrained language model optimized for the tasks can maintain competitive stability regardless of the translation direction, avoiding severe overfitting to a single linguistic modality.

4.3.2. Validation Variance and Generalization Challenges

It is essential to contextualize the official blind test results with our internal validation metrics. During the development phase, internal tests utilizing a 10% validation split yielded significantly higher metrics (77.65 BLEU and 87.11 chrF++) for the MSLG2SPA task. The pronounced discrepancy between these internal scores and the official test results (32.11 BLEU and 60.54 chrF) reveals inherent challenges in applying rigid n-gram evaluation metrics to generative Large Language Models within low-resource domains.

We attribute this performance variance primarily to three structural factors:

1. **In-Domain vs. Out-of-Domain Distribution:** The internal validation set (49 instances) was generated via a randomized split from the same base corpus as the training data. Consequently, it shared the exact same annotator bias, stylistic phrasing, and restricted vocabulary. Conversely, the official blind test set introduced completely unseen semantic concepts and Out-of-Vocabulary (OOV) glosses. In a constrained dataset of fewer than 500 training pairs, the model lacks sufficient lexical diversity to generalize perfectly on zero-shot OOV structures.
2. **Generative Paraphrasing vs. Exact N-gram Matching:** Unlike traditional rule-based or specialized Seq2Seq translation models that rigidly output memorized patterns, LLMs like `Llama-3.2` are generative. They produce natural, fluent paraphrases that preserve semantic meaning but diverge in exact word choice (e.g., generating *auto* instead of *carro*). Metrics such as BLEU severely penalize these legitimate variations because they rely on exact word-for-word overlaps against a single human reference.
3. **Strict Orthographic Penalties in Gloss Generation:** Specifically for the SPA $\rightarrow$ MSLG task, evaluation metrics heavily penalize minor typographical deviations. If the LLM generates a conceptually correct sign sequence but slightly alters a compound delimiter (e.g., using a space instead of a plus sign +) or hallucinates the capitalization of a fingerspelling prefix (`dm-`), the precision scores drop precipitously, even if the translation remains fully comprehensible to a human evaluator.

These results indicate that while the model successfully captures the core morphosyntactic mapping between Spanish and LSM, future evaluations of LLM-based sign language translators would benefit

significantly from semantic-aware metrics (such as BERTScore) that better reflect natural language generative capabilities.

4.4. Qualitative Analysis and Error Inspection

To provide a deeper understanding of the model's generative behavior and its impact on the evaluation metrics, Table 3 presents a qualitative sample of the MSLG2SPA translations extracted directly from our internal validation phase.

Table 3
Qualitative Translation Examples (MSLG → SPA).

Input Gloss (MSLG)	Human Reference (SPA)	Model Prediction (SPA)
JABÓN BURBUJAS	El jabón hace burbujas.	El jabón hace burbujas.
APAGAR LUZ POR-FAVOR	Por favor, apaga la luz.	Apaga la luz, por favor.
TÚ NO PREOCUPAR DEJARLO-ASÍ	No te preocupes, déjalo así.	No te preocupes, déjalo así.

As observed in Table 3, the model correctly infers grammatical connectors and conjugates verbs appropriately (e.g., converting BURBUJAS to *hace burbujas*). However, the second example (APAGAR LUZ POR-FAVOR) perfectly illustrates the evaluation challenge discussed previously: while the model's prediction (*"Apaga la luz, por favor."*) is semantically identical and grammatically correct compared to the reference (*"Por favor, apaga la luz."*), the syntactic reordering inevitably degrades the exact n-gram matching score penalized by BLEU.

4.4.1. Error Analysis for SPA2MSLG and Special Markers

The translation from standard Spanish to LSM glosses (SPA2MSLG) introduced a specific set of failure modes tied directly to the rigid orthographic conventions of the shared task. Error inspection reveals that the autoregressive nature of `Llama-3.2` struggles with deterministic, non-natural linguistic operators.

The most frequent errors involved the fingerspelling prefix (`dm-`) and contraction markers (+). For instance, when prompted to translate proper nouns, the model frequently hallucinated casing formats, outputting `DM-NOMBRE` instead of the required `dm-NOMBRE`, or incorrectly inserting spaces around operators (e.g., generating `MAMÁ + PAPÁ` instead of the strict `MAMÁ+PAPÁ` token). Furthermore, lexical borrowings denoted by the hash symbol (#) were occasionally omitted entirely in favor of a descriptive gloss sequence. Since the chrF++ metric is highly sensitive to character-level sequence alignment, these minor tokenization artifacts drastically reduced the official scores for the SPA2MSLG track, despite the underlying semantic structure remaining largely accurate.

4.5. Ablation Study

To quantify the absolute contribution of the fine-tuning process against the foundational capabilities of the underlying language model, we conducted an ablation study on the MSLG → SPA translation direction using our internal 10% validation split.

The experiment isolated the generative capabilities into three distinct configurations:

1. **Base Llama:** The base `Llama-3.2-3B-Instruct` model deployed in a zero-shot configuration, receiving only the raw LSM gloss sequence without any contextual framing.
2. **Prompt-Only:** The base model enhanced solely with our designed expert system prompt, providing task context but lacking weight adaptation.
3. **QLoRA-Adapted (Ours):** The complete pipeline, integrating both the expert system prompt and the task-specific low-rank adapted matrices trained on the corpus.

Table 4
Ablation Study Results on the Internal Validation Set (MSLG → SPA).

Configuration	BLEU	chrF++
Base Llama	0.29	11.94
Prompt-Only	5.38	26.37
QLoRA-Adapted (Full System)	**77.65**	**87.11**

As demonstrated in Table 4, deploying the base model without adaptation results in severe metric degradation, as the generic instruct-tuned weights fail to parse the heavily compressed syntax of LSM glosses. While the injection of the expert prompt (Prompt-Only) provides a marginal structural improvement, it is insufficient to overcome the morphosyntactic asymmetry between the languages. The massive performance spike observed in the QLoRA-adapted version confirms that low-rank fine-tuning is indispensable for bridging the structural divide, allowing the model to learn the specific generative mapping required for sign language translation.

5. Conclusion and Future Work

In this work, we presented the participation of the LSM-INAOE team in the MSLG-SPA shared task at IberLEF 2026. Our approach evaluated the viability of a lightweight, open-weight Large Language Model (Llama-3.2-3B-Instruct) optimized through 4-bit NF4 quantization and Low-Rank Adaptation (QLoRA) to handle the complex, bidirectional translation mapping between Mexican Sign Language (LSM) glosses and standard written Spanish.

The official preliminary evaluation results demonstrated that our system maintains a highly stable and balanced performance across both translation task, securing the 7th position in the Global Ranking out of 18 active international teams. While many contemporary architectures exhibit asymmetric degradation when changing the translation direction, our model framework consistently achieved eighth place in both individual tasks (MSLG2SPA and SPA2MSLG). This symmetric consistency confirms that compact, quantized LLMs can successfully generalize language alignments without requiring expensive cloud infrastructure or persistent remote connectivity.

For future work, we plan to address the performance variance observed between our internal validation metrics and the official blind test results through two primary avenues:

1. **Dataset Expansion:** We aim to scale the local training corpus to incorporate a wider diversity of semantic domains and regional vocabulary, effectively minimizing Out-of-Vocabulary (OOV) errors during zero-shot generalization.
2. **Advanced Prompt Engineering:** We will refine our prompts within the system's guidelines to ensure stricter compliance with special orthographic tokens (such as hyphens, compound delimiters, and dactylic prefixes like `dm-`), thereby reducing character-level evaluation penalties.

While our localized quantization framework demonstrates that a 3-billion-parameter model can operate efficiently on standard cloud hardware, a limitation of this study is that real-time inference tests were not performed on ultra-low-power physical edge platforms. Nevertheless, the theoretical feasibility of this implementation is strongly supported by the optimized system memory consumption.

Traditionally, Large Language Models (LLMs) require a prohibitive amount of edge computing resources. While commercial architectures like GPT-3.5 or GPT-4 [13] offer excellent translation capabilities, their computational requirements (scaling to hundreds of gigabytes or terabytes of distributed cloud memory) make them impractical for isolated medical, educational, or community environments. Even larger open-source LLMs like Llama 3 70B [5] or Mistral Large 123B [14], even when quantized to 4 bits, require approximately 40–45 GB of net VRAM for the model alone, plus an additional 10–15 GB to manage long context windows via Key-Value (KV) cache. Furthermore, the NF4 4-bit quantization applied to the model we deployed compresses the base weights of the 3B architecture to approximately

2 GB [5]. It is crucial to note that the maximum VRAM allocation of 7.2 GB observed in our experiments corresponds strictly to the fine-tuning phase, which inherently requires a large amount of additional memory for optimizer states and gradient graphs. In contrast, pure autoregressive inference requires significantly less capacity. Consequently, edge AI devices like the NVIDIA Jetson series, which possess between 8 and 32 GB of unified memory, retain sufficient architectural capacity to accommodate both the quantized model and the KV cache necessary for short, sentence-level translations.

Therefore, our future work will focus on the physical deployment and benchmarking of the QLoRA-adapted `Llama-3.2-3B-Instruct` [5] on these edge platforms. This next phase will use optimized C++ inference engines to experimentally evaluate metrics such as Time-To-First-Token (TTFT), overall translation latency (tokens per second), memory efficiency under strict RAM constraints, and heat dissipation. The ultimate goal is to validate a potential real-time, fully portable, bidirectional translation assistant for the Mexican Sign Language community.

Acknowledgments

The authors express their gratitude to the organizers of the MSLG-SPA 2026 task and the IberLEF committee for the coordination, datasets, and automated evaluation platforms. We also acknowledge the computational infrastructure and academic support provided by the National Institute of Astrophysics, Optics and Electronics (INAOE), as well as the Secretariat of Science, Humanities, Technology and Innovation (SECIHTI) and the financial support it provided during the development of this research.

Declaration on Generative AI

During the preparation of this work, the authors utilized Gemini in order to perform English grammar checks, contextual phrase smoothing, and structural alignment of the LaTeX manuscript. Following the use of this service, the authors critically reviewed, verified, and edited the text as required, taking full responsibility for the final publication content.

References

[1] World Health Organization, Deafness and hearing loss, 2024. URL: https://www.who.int/es/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] National Autonomous University of Mexico, Communication barriers in healthcare for the deaf community in mexico, PMC - PubMed Central (2025). URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12787980/.

[3] A. Y. Rodríguez-González, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, M. G. Sánchez-Cervantes, Á. Díaz-Pacheco, R. Aranda, S. Fajardo-Flores, C. H. Martínez-Rodríguez, Overview of mslg-spa at iberlef 2026: Bidirectional translation between mexican sign language glosses and spanish, Procesamiento del Lenguaje Natural 77 (2026).

[4] 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.

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024). doi:10.48550/arXiv.2407.21783.

[6] K. Yin, A. Read, A. Bohr, A. Cervantes, H. Touvron, Including sign languages in natural language processing, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 7347–7360.

[7] M. Cruz Aldrete, Gramática de la lengua de señas mexicana, Tesis de doctorado, El Colegio de México, 2008.

[8] B. Martinez-Seis, O. Pichardo-Lagunas, et al., Automatic translation of sentences to mexican sign language: Rule-based machine translation and animation synthesis in avatar, Computación y Sistemas 29 (2025).
[9] O. Pichardo-Lagunas, L. Partida-Terrón, et al., Sistema de traducción directa de español a lsm con reglas marcadas, Research in Computing Science 115 (2016) 29–41.
[10] V. Lara-Ortiz, S. Padó, Low-resource sign language glossing profits from data augmentation, in: Proceedings of the Workshop on Sign Language Processing (WSLP), ACL Anthology, 2025, pp. 14–19.
[11] N. C. Camgoz, O. Koller, S. Hadfield, R. Bowden, Sign language transformers: Joint end-to-end sign language recognition and translation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10023–10033.
[12] V. Lara-Ortiz, R. Q. Fuentes-Aguilar, I. Chairez, Spanish to mexican sign language (msl) glosses corpus for nlp tasks, Scientific Data 12 (2025).
[13] OpenAI, Gpt-4 technical report, 2023. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.
[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023). URL: https://arxiv.org/abs/2310.06825.

A. Online Resources

The complete source code for the QLoRA fine-tuning process and the ablation study scripts utilized in this research are publicly available in our repository at: https://github.com/KorakiMx/IberLEF-LSM-INAOE