UCO-UC-CICESE-Plenitas Team at MSLG-SPA 2026: Bidirectional Translation between Mexican Sign Language Glosses and Spanish Using Large Language Models

Yoan Martínez-López[1,2,*], Yanaima Jauriga[3], Lismaida Agüero Rodríguez[3], Julio Madera[3], Ansel Rodríguez-González[4], Kristell Aguilar Alarcón[1,2], Carlos de Castro Lozano[1,2] and José Miguel Ramírez Uceda[1,2]

[1]*EATCO. 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]*Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad Académica Tepic, México*

Abstract

This paper describes our participation in MSLG-SPA 2026, the first IberLEF shared task on bidirectional translation between Mexican Sign Language glosses (MSLG) and Spanish. The task comprises two low-resource sequence-to-sequence translation subtasks: MSLG2SPA and SPA2MSLG, using a parallel corpus of 489 training pairs. We propose a modular translation pipeline with two interchangeable backends: a HuggingFace sequence-to-sequence architecture with QLoRA-based fine-tuning, and an Ollama-based large language model backend. Both approaches employ a structured bilingual prompting strategy that explicitly preserves the annotation conventions present in the official corpus. Our final submission used the Ollama backend with low-temperature decoding and achieved competitive results in both subtasks. The system ranked 4th out of 20 teams in MSLG2SPA (Task Score 1.0628), 7th out of 19 teams in SPA2MSLG (Task Score 0.2329), and 5th overall in the global ranking (Task Score 0.6478). These results show that prompt-based LLM approaches can provide effective performance for low-resource sign language gloss translation.

Keywords

Mexican Sign Language, gloss translation, LLMs, MSLG-SPA

1. Introduction

Automatic translation between sign languages and spoken languages remains a challenging and relatively underexplored problem in Natural Language Processing (NLP) [1, 2, 3]. Despite recent advances in neural machine translation (NMT) and large language models (LLMs), computational resources for sign languages are still scarce, particularly for languages with limited annotated corpora and reduced institutional support. This limitation is especially evident in Mexican Sign Language (LSM, Lengua de Señas Mexicana), for which publicly available datasets and computational benchmarks remain limited [4].

The MSLG-SPA 2026 shared task, organized within IberLEF 2026 [5], introduces the first benchmark for bidirectional translation between Mexican Sign Language glosses (MSLG) and Spanish. The task comprises two subtasks: MSLG2SPA, which translates gloss sequences into Spanish text, and SPA2MSLG, which generates gloss sequences from Spanish sentences. The dataset is intentionally small, containing

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
ymlopez2022@gmail.com (Y. Martínez-López); yanaima.jauriga@reduc.edu.cu (Y. Jauriga); lismaida.aguero@reduc.edu.cu (L. A. Rodríguez); 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); 0009-0000-6891-0068 (Y. Jauriga); 0000-0002-6476-9498 (L. A. Rodríguez); 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)

only 489 training pairs and 244 test instances per direction, thus representing a realistic low-resource translation scenario.

The task presents several challenges beyond data scarcity. MSL glosses differ substantially from Spanish in syntax and grammatical structure, often omitting inflectional and functional information that must be reconstructed during translation. In addition, the official corpus preserves simplified annotation conventions such as compound signs, contractions, lexical borrowings and dactylology markers, which introduce additional ambiguity for automatic systems.

In this paper, we present a modular translation pipeline designed for low-resource bidirectional translation between MSL glosses and Spanish. Our approach supports two interchangeable backends: a HuggingFace sequence-to-sequence architecture with QLoRA-based fine-tuning, and an Ollama-based LLM backend using structured bilingual prompting. The proposed system achieved competitive results in the shared task, ranking 4th out of 20 teams in MSLG2SPA, 7th out of 19 teams in SPA2MSLG, and 5th overall in the global ranking.

2. Methodology

The MSLG-SPA 2026 shared task is formulated as two sequence-to-sequence translation problems over the same parallel corpus: MSL gloss → Spanish (MSLG2SPA) and Spanish → MSL gloss (SPA2MSLG) [4]. For each direction, systems must generate a target sequence for every instance in the official test set and submit the predictions following the organizers' format specifications. Evaluation is performed using BLEU, METEOR and chrF for both subtasks, complemented by COMET for MSLG2SPA, and aggregated through z-score-normalized metric values.

2.1. Large Language Models (LLMs)

Large Language Models (LLMs), predominantly based on the Transformer architecture, have become the dominant paradigm in Natural Language Processing due to their strong performance across a wide range of tasks [6, 7, 8]. Their ability to leverage large-scale pretrained linguistic knowledge makes them particularly attractive for low-resource translation scenarios such as MSLG-SPA 2026, where only 489 parallel training pairs are available.

Our pipeline evaluates several instruction-tuned open-weight LLMs as translation backends for bidirectional translation between Mexican Sign Language glosses and Spanish. The evaluated models include Qwen2.5 [9], Mistral-7B-Instruct [10], Gemma-2 and Gemma-3 [11]. These models were selected due to their multilingual capabilities and compatibility with the Ollama serving framework.

The Qwen family provides strong multilingual coverage, including Spanish, while Mistral-7B-Instruct serves as a competitive medium-scale multilingual baseline. Gemma-2 and Gemma-3 were evaluated to compare different generations of instruction-tuned models under the same prompting strategy. Most configurations were executed locally through an Ollama daemon, while Gemma-3 was additionally evaluated through the Ollama Cloud API.

2.2. Data Loading and Annotation Handling

The official corpus is distributed as tab-separated files containing parallel MSL gloss and Spanish sentences [4]. Our pipeline preserves the original text verbatim throughout all processing stages, including the annotation conventions retained in the simplified gloss sequences after preprocessing by the organizers:

- Hyphen (-): compound signs, e.g. YA-VEO, CON-PERMISO.
- Plus sign (+): contractions or compounds, e.g. MAMÁ+PAPÁ denoting padres ('parents').
- Hash sign (#): lexical borrowing through fingerspelling, e.g. #OK.
- Prefix dm-: proper-noun dactylology, e.g. dm-LUIS ('Luis' spelled out).

Rather than removing these markers, our system propagates them unchanged throughout the entire pipeline, from data loading and prompt construction to model inference and submission generation. Their meaning is explicitly communicated to the model through the structured bilingual prompting strategy.

2.3. System Architecture

The system is organized as a modular sequential pipeline composed of four decoupled stages:

- **Data ingestion**: reads and normalizes the official training and test files.
- **LLM provider**: encapsulates the translation backend behind a common abstract interface (`BaseLLMProvider`) exposing a unified `generate(text)` method. Two implementations are provided: `HuggingFaceLLM` and `OllamaLLM`.
- **Prediction generation**: iterates over the test partition, selects the appropriate source field according to the configured translation direction, invokes the provider backend and collects prediction pairs.
- **Submission writing**: serializes the generated predictions into UTF-8 formatted submission files compliant with the official evaluation protocol.

All experimental settings —including backend selection, model identifiers, decoding parameters and training hyperparameters— are centralized through command-line arguments, making the pipeline fully reproducible from its invocation configuration.

The proposed architecture is designed around three principles: modular separation between translation backends and the rest of the pipeline, explicit preservation of MSL annotation conventions through structured prompting, and strict compatibility with the official MSLG-SPA 2026 submission format. The implementation is developed in Python 3 using PyTorch, HuggingFace Transformers, PEFT (LoRA/QLoRA), bitsandbytes and the Ollama REST API.

Figure 1 summarizes the end-to-end workflow of the proposed translation pipeline. The process begins with data ingestion and normalization of the official corpus while preserving the annotation conventions retained in the gloss sequences. A direction-specific bilingual prompt is then constructed and forwarded to the selected backend through the unified provider abstraction.

The architecture supports both HuggingFace-based sequence-to-sequence models with QLoRA adaptation and Ollama-based LLM inference, allowing local and remote deployment configurations under the same interface. Finally, generated predictions are collected and serialized into submission files compatible with the official MSLG-SPA 2026 evaluation format.

2.4. HuggingFace Backend with QLoRA Fine-Tuning

The first backend wraps sequence-to-sequence models available through the HuggingFace ecosystem. Translation models and tokenizers are loaded using `AutoTokenizer` and `AutoModelForSeq2SeqLM`, with automatic device mapping (`device_map="auto"`) and optional 4-bit quantization through `BitsAndBytesConfig` using NF4 quantization with `bfloat16` compute precision. This configuration enables efficient adaptation of multilingual models on consumer-grade GPUs.

For NLLB-family models, source and target language codes are configured explicitly according to the translation direction, and the corresponding `forced_bos_token_id` is applied during generation. For non-NLLB architectures, decoding relies on the standard EOS/PAD mechanism, allowing compatibility with other encoder-decoder models such as mT5, MarianMT and mBART.

When fine-tuning is enabled, the model is prepared through `prepare_model_for_kbit_training` and adapted using Low-Rank Adaptation (LoRA) with rank $r = 16$, scaling factor $\alpha = 32$ and dropout 0.05. LoRA adapters are applied to the attention projection matrices (`q_proj`, `k_proj`, `v_proj` and `out_proj`) using the `SEQ_2_SEQ_LM` task configuration.

The training corpus is wrapped in a custom `TranslationDataset` that tokenizes source and target sequences with truncation to 128 subword tokens. Training is performed through the HuggingFace

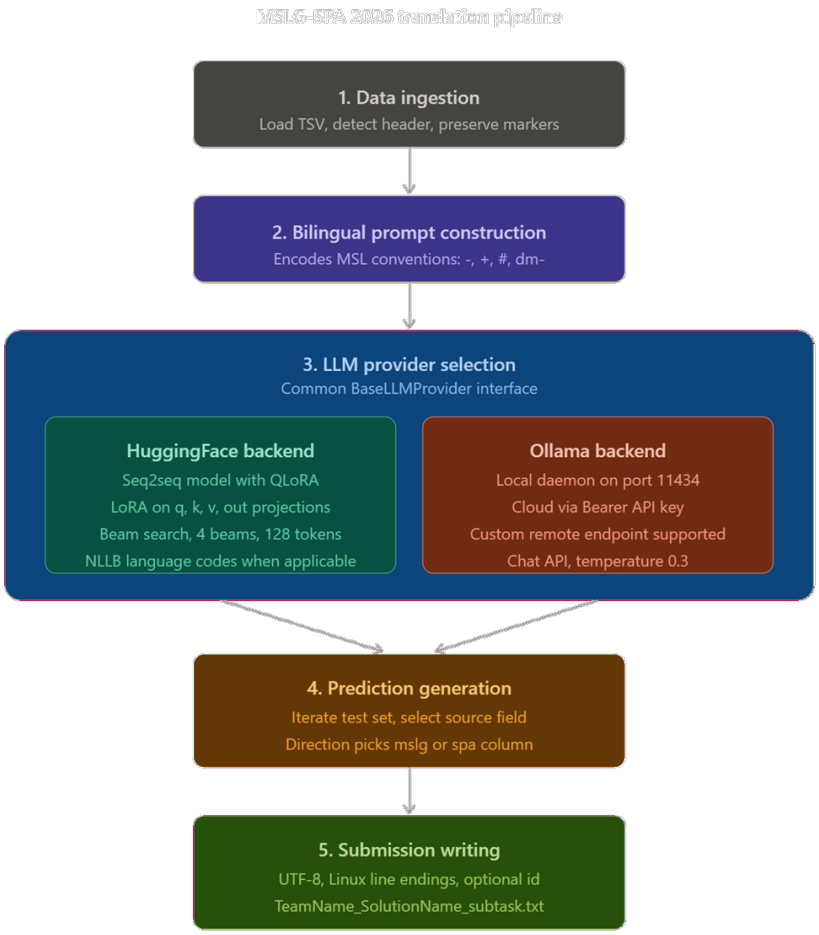

Figure 1: Overview of the proposed MSLG-SPA 2026 translation pipeline.

`Trainer` API using 5 epochs, batch size 4, learning rate 2×10^{-4} and FP16 mixed precision when CUDA is available. The best checkpoint is stored as a PEFT adapter for later inference reuse.

During inference, the backend applies beam-search decoding (`num_beams=4`, `early_stopping=True`, `max_new_tokens=128`) under `torch.no_grad()` and decodes predictions using `skip_special_tokens=True`. The same decoding configuration is used in both translation directions.

2.5. Bilingual Structured Prompting

For both backends —particularly the prompt-based Ollama backend— translation is framed through a structured bilingual prompt written in Spanish. The prompt is direction-specific and follows a four-block structure:

- A high-level task instruction specifying the source and target languages.
- A declaration of the annotation conventions preserved in the MSL gloss sequences (`-`, `+`, `#` and `dm-`), including canonical examples.
- The source sequence provided verbatim with its corresponding language label.
- A trailing instruction explicitly requesting the translation output.

The prompting strategy is asymmetric across translation directions. In MSLG2SPA, the model is instructed to interpret the annotation markers in order to recover semantic meaning without reproducing them in the Spanish output. In SPA2MSLG, the model is instructed to generate these markers when

appropriate. This design aligns the prompting strategy with the annotation conventions of the official corpus and reduces the propagation of glossing artifacts into Spanish translations.

2.6. Ollama Backend (Local and Cloud)

The second backend wraps the native Ollama REST API and supports three deployment modes through a unified interface:

- **Local mode**: inference against a locally running Ollama daemon without authentication.
- **Cloud mode**: inference through the Ollama Cloud API using Bearer authentication via an API key.
- **Custom remote mode**: inference against arbitrary Ollama-compatible endpoints specified by the user.

The backend automatically normalizes endpoint configuration and performs connectivity validation during initialization. Translation requests are executed as chat completions using the configured model identifier and the structured bilingual prompt described in the previous subsection.

Decoding is performed with low-temperature generation (`temperature=0.3` and `num_predict=256` by default), reflecting the deterministic nature of the task, where conservative lexical generation is generally favored over creative paraphrasing under BLEU-, chrF-, METEOR- and COMET-based evaluation. Responses are parsed automatically, with explicit handling of HTTP errors, timeouts and authentication failures to avoid interruption of the full inference pipeline.

2.7. Evaluation Metrics

System outputs are evaluated using the metrics defined by the MSLG-SPA 2026 organizers. For MSLG2SPA, evaluation includes BLEU, METEOR, chrF and COMET, since the target side consists of well-formed Spanish text that supports fluency-aware evaluation through pretrained language models. For SPA2MSLG, evaluation is restricted to BLEU, METEOR and chrF because MSL gloss sequences do not constitute a natural language suitable for COMET-based assessment.

In both subtasks, metric values are standardized through z-score normalization across participating systems and aggregated into a per-subtask Task Score used for ranking. An additional global ranking averages the scores obtained in both translation directions.

Internally, we use the same evaluation metrics on a held-out 10% split of the training corpus to compare model backends and prompting configurations prior to selecting the final submitted runs.

3. Results

MSLG2SPA Results

Table 1 reports the internal comparison of the three Ollama-based configurations evaluated in our held-out validation partition for the MSLG2SPA subtask. The systems are ranked by the per-subtask Task Score, computed as the arithmetic mean of the z-score-normalized BLEU, METEOR, chrF and COMET values across all internally evaluated configurations, following the official aggregation protocol of MSLG-SPA 2026.

Table 1
Internal comparison

Method	Task Score	BLEU	METEOR	chrF	COMET
mistral	1,0628	47,6182	0,6515	71,8559	0,8823
qwen	1,0628	47,6182	0,6515	71,8559	0,8823
gemma4-cloud	1,0345	46,9379	0,6462	71,5308	0,8787

All three evaluated configurations achieved positive Task Scores, indicating that the proposed structured prompting strategy and annotation-aware pipeline are effective under the low-resource conditions of the shared task. The systems reached BLEU scores in the 46–48 range, METEOR values around 0.65 and chrF scores above 71, which represent competitive performance for gloss-to-Spanish translation given the limited training corpus.

The mistral and qwen configurations obtained identical scores across all evaluated metrics, suggesting highly similar translation behavior on the validation partition. In contrast, the gemma4-cloud configuration achieved slightly lower scores across BLEU, METEOR, chrF and COMET. Although the differences are relatively small, the degradation is consistent across all evaluated metrics.

Although a HuggingFace backend with QLoRA fine-tuning was implemented, the Ollama-based configurations consistently achieved better validation scores and were therefore selected for the official submissions.

Table 2 presents the official MSLG2SPA leaderboard released by the organizers, including 20 participating teams ranked according to the z-score-normalized aggregation of BLEU, METEOR, chrF and COMET. Our submission achieved 4th place with a Task Score of 1.0628, placing UCO-UC-CICESE-Plenitas among the top-performing systems in the gloss-to-Spanish translation direction.

Table 2
Official Evaluation Results MSLG2SPA

Rank	Team Name	Task Score
1	CICESE-LCDAA	1,3421
2	AxoloTux	1,3400
3	UAM-ChineseRoom	1,1890
4	**UCO-UC-CICESE-Plenitas**	1,0628
5	VerbaNexAI	1,0339
6	Ludwig	1,0055
7	ComunicadorInnato	0,8811
8	LSM-INAOE	0,4105
9	FisBioUNAM	0,3724
10	RETUYT-INCO	0,2043
11	ReCore	-0,1167
12	DunderMifflin	-0,2294
13	Bortolotii	-0,4640
14	MBM-Solution	-0,5103
15	SignNexus	-0,5303
16	WangKongqiang	-0,6605
17	SeÑLP	-0,7019
18	SeñasDelDesierto	-1,3320
19	TeamCICESE	-1,4002
20	Lepopardo	-1,5388

SPA2MSLG Results

Table 3 reports the internal comparison of the four Ollama-based configurations evaluated on our held-out validation partition for the SPA2MSLG subtask. In this direction, the Task Score is computed as the arithmetic mean of the z-score-normalized BLEU, METEOR and chrF values across all evaluated configurations. Following the official MSLG-SPA 2026 protocol, COMET is not applied because MSL gloss sequences do not constitute a natural language suitable for fluency-based evaluation.

The contrast with the MSLG2SPA direction is substantial. While gloss-to-Spanish translation achieved BLEU scores above 46 and Task Scores above 1.00, SPA2MSLG reached BLEU values between 9 and 11 and Task Scores between -0.02 and 0.23. This performance reduction reflects the intrinsic asymmetry of the task, where generating MSL gloss sequences from Spanish text is considerably more difficult than recovering Spanish sentences from gloss representations.

Despite the lower BLEU and METEOR values, all evaluated configurations maintained chrF scores in

Table 3
Internal comparison

Method	Task Score	BLEU	METEOR	chrF
mistral	0,2329	10,8194	0,4518	57,7966
qwen	0,2329	10,8194	0,4518	57,7966
gemma4-cloud	-0,0233	9,2519	0,4066	56,2141
gemma3	-0,0233	9,2519	0,4066	56,2141

the 56–58 range. The large gap between chrF (57.80 / 56.21) and BLEU (10.82 / 9.25) is considerably wider than the equivalent gap observed in MSLG2SPA, suggesting that the models frequently recover relevant gloss vocabulary while failing to reproduce the exact gloss ordering required by the references. This behavior is consistent with the structural differences between Spanish syntax and the topic-prominent organization characteristic of MSL gloss sequences. The METEOR values above 0.40 further indicate that the generated gloss sequences often preserve relevant semantic content despite deviations in surface ordering.

Table 4 presents the official SPA2MSLG leaderboard released by the organizers. Our submission achieved 7th place out of 19 participating teams with a Task Score of 0.2329. Although the results in this direction were substantially lower than those obtained in MSLG2SPA, the system remained competitive under the structural complexity and low-resource constraints of Spanish-to-gloss translation.

Table 4
Official Evaluation Results SPA2MSLG

Rank	Team Name	Task Score
1	AxoloTux	1,7577
2	VerbaNexAI	1,4466
3	UAM-ChineseRoom	1,2724
4	CICESE-LCDAA	1,0072
5	RETUYT-INCO	0,5705
6	ReCore	0,4218
7	**UCO-UC-CICESE-Plenitas**	0,2329
8	LSM-INAOE	0,1125
9	FisBioUNAM	0,1003
10	DunderMifflin	0,0405
11	MBM-Solution	-0,0644
12	Bortolotii	-0,1607
13	WangKongqiang	-0,3338
14	ComunicadorInnato	-0,3549
15	SeÑLP	-0,4964
16	SignNexus	-0,6083
17	TeamCICESE	-1,1424
18	SeñnasDelDesierto	-1,4920
19	Lepopardo	-1,6999

Our submission achieved 7th place out of 19 participating teams with a Task Score of 0.2329. Although the system remained below the top-performing approaches, it maintained competitive performance despite relying exclusively on prompt-based inference without task-specific fine-tuning or ensemble methods. The relatively small score differences among several mid-table systems further suggest that moderate methodological improvements could substantially affect the final ranking in this translation direction.

Global Ranking

The combined interpretation of the two official leaderboards provides the clearest summary of our overall participation in MSLG-SPA 2026. Our team achieved:

- Rank 4 of 20 in MSLG2SPA with Task Score 1.0628.
- Rank 7 of 19 in SPA2MSLG with Task Score 0.2329.

Table 5 presents the official global ranking released by the organizers, which aggregates the per-subtask Task Scores of all teams participating in both translation directions. The global Task Score is computed as the arithmetic mean of the MSLG2SPA and SPA2MSLG Task Scores, rewarding systems that maintain balanced performance across both tasks rather than specialization in a single direction.

Table 5
Official Evaluation Results Global Ranking

Rank	Team Name	Task Score	MSLG2SPA Task Score	SPA2MSLG Task Score
1	AxoloTux	1,5489	1,3400	1,7577
2	VerbaNexAI	1,2403	1,0339	1,4466
3	UAM-ChineseRoom	1,2307	1,1890	1,2724
4	CICESE-LCDAA	1,1746	1,3421	1,0072
5	**UCO-UC-CICESE-Plenitas**	0,6478	1,0628	0,2329
6	RETUYT-INCO	0,3874	0,2043	0,5705
7	ComunicadorInnato	0,2631	0,8811	-0,3549
8	LSM-INAOE	0,2615	0,4105	0,1125
9	FisBioUNAM	0,2364	0,3724	0,1003
10	ReCore	0,1525	-0,1167	0,4218
11	DunderMifflin	-0,0944	-0,2294	0,0405
12	MBM-Solution	-0,2873	-0,5103	-0,0644
13	Bortolotii	-0,3124	-0,4640	-0,1607
14	WangKongqiang	-0,4972	-0,6605	-0,3338
15	SignNexus	-0,5693	-0,5303	-0,6083
16	SeÑLP	-0,5992	-0,7019	-0,4964
17	TeamCICESE	-1,2713	-1,4002	-1,1424
18	SeñasDelDesierto	-1,4120	-1,3320	-1,4920
19	Lepopardo	-1,6194	-1,5388	-1,6999

Among the 19 teams included in the global evaluation, UCO-UC-CICESE-Plenitas achieved 5th place overall with a combined Task Score of 0.6478. The submission maintained a clear margin over most mid-table systems while remaining below the four highest-ranked teams, indicating that the proposed pipeline achieves robust cross-direction performance under the low-resource conditions of the shared task.

The distribution of scores across the leaderboard also highlights the discriminative nature of the benchmark. Nearly half of the participating teams obtained negative overall Task Scores, suggesting that successful participation in both translation directions requires more than simply generating valid outputs for each task. In this context, the balanced performance achieved by UCO-UC-CICESE-Plenitas across both subtasks constitutes one of the main strengths of the proposed system.

The gap to the first positions remains substantial: 0.5268 points to rank 4 (CICESE-LCDAA, 1.1746) and 0.9011 points to rank 1 (AxoloTux, 1.5489). However, the separation from the lower-middle portion of the leaderboard is also notable. UCO-UC-CICESE-Plenitas maintains a 0.2604-point margin over rank 6 (RETUYT-INCO, 0.3874), which is considerably larger than the average inter-rank gap observed in that region of the leaderboard (≈ 0.06). This positioning suggests that the proposed system achieves a stable performance level above most mid-table competitors while remaining below the top-performing submissions.

The 5th position obtained by UCO-UC-CICESE-Plenitas in the global ranking, with a combined Task Score of 0.6478, provides a concise summary of the overall performance of our participation in MSLG-SPA 2026. The submission outperformed 14 of the 19 teams participating in both directions and successfully translated the strong MSLG2SPA performance into a competitive overall ranking through balanced results across both subtasks.

4. Conclusions

In this paper, we presented a modular and reproducible pipeline for bidirectional translation between Mexican Sign Language glosses and Spanish in the context of the MSLG-SPA 2026 shared task at IberLEF. The proposed methodology combines interchangeable translation backends with structured bilingual prompting and explicit preservation of MSL annotation conventions, allowing the same pipeline to support both HuggingFace-based fine-tuning and Ollama-based prompt-driven inference.

The experimental results demonstrate that prompt-based LLM approaches can achieve competitive performance under the low-resource conditions of the shared task. UCO-UC-CICESE-Plenitas achieved 4th place out of 20 teams in MSLG2SPA, 7th place out of 19 teams in SPA2MSLG and 5th place overall in the official global ranking. The results also reveal a clear asymmetry between translation directions: gloss-to-Spanish translation proved substantially easier than Spanish-to-gloss generation, particularly due to the structural and ordering constraints characteristic of MSL gloss sequences.

The observed gap between BLEU and chrF in SPA2MSLG suggests that current prompt-based systems can often recover the appropriate gloss vocabulary while still struggling with the exact structural organization of gloss sequences. These findings indicate that annotation-aware prompting is effective for low-resource sign language translation, but also that gloss generation remains a substantially harder problem than gloss interpretation.

Future work will explore larger multilingual backbones, task-specific fine-tuning strategies and hybrid approaches combining prompting with supervised adaptation. Additional directions include the incorporation of linguistic constraints for gloss ordering and the evaluation of larger parallel corpora for Mexican Sign Language translation.

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] S. C. Fanni, M. Febi, G. Aghakhanyan, E. Neri, Natural language processing, in: Introduction to Artificial Intelligence, Springer International Publishing, Cham, 2023, pp. 87–99.

[2] A. K. Joshi, Natural language processing, Science 253 (1991) 1242–1249.

[3] P. Shetty, R. Udhayakumar, A. Patil, M. Manwal, P. S. Vadar, Application of natural language processing (NLP) in machine learning, in: 2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE), IEEE, 2023, pp. 949–957.

[4] 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).

[5] 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.

[6] 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).

[7] A. Zubiaga, Natural language processing in the era of large language models, Frontiers in Artificial Intelligence 6 (2024) 1350306.

[8] 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.

[9] 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.

[10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, W. El Sayed, Mistral 7B, arXiv preprint arXiv:2310.06825 (2023). URL: https://arxiv.org/abs/2310.06825. doi:10.48550/arXiv.2310.06825.

[11] Gemma Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, others, S. Iqbal, Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025). doi:10.48550/arXiv.2503.19786.

A. Online Resources

The results of the MSLG-SPA 2026 are available via:

- MSLG-SPA,
- MSLG-SPA 2026 Results,
- MSLG-SPA 2026 Code