LoRA Fine-Tuning, Frontier-LLM Prompting, and Encoder-Decoder Baselines for Low-Resource Bidirectional Sign Language Gloss Translation

Carlos Minutti-Martinez[1,*], Alejandra Torres-Nájera[1], Boris Escalante-Ramirez[2] and Jimena Olveres[2]

[1]INFOTEC, Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Aguascalientes, Mexico
[2]Departamento de Procesamiento de Señales, División de Ingeniería Eléctrica, Universidad Nacional Autónoma de México, Mexico City, Mexico

Abstract

Bidirectional translation between sign-language glosses and spoken language is an inherently low-resource problem: parallel corpora rarely exceed a few thousand pairs, and many language pairs lack any curated resource at all. We compare three families of approaches that reuse external knowledge in different ways: a compact encoder-decoder baseline, prompting of commercial frontier LLMs (Gemini 3 Pro, Claude Opus 4.5), and parameter-efficient LoRA fine-tuning of an 8-bit quantized 9B open-weight LLM, trained jointly on both directions under a two-phase curriculum that first warms up on a broader external corpus and then specializes on the curated challenge data, optionally distilling from the frontier-LLM predictions. We also introduce a reference-free, interpretable translation-quality scorer used to rank candidate models when official references are withheld. We instantiate the comparison on the MSLG-SPA 2026 shared task, which provides only 490 parallel (MSLG, SPA) sentence pairs. Submitted as team AxoloTux, our system ranked first overall, first on SPA→MSLG, and second on MSLG→SPA. Beyond the official ranking, three findings generalize across approaches: prompted frontier LLMs are not interchangeable on under-resourced pairs, with Claude performing strongly on both directions while Gemini dropped nine ranks on SPA→MSLG; under the COMET metric, all six of our submissions land in the top fifth of the field and a non-distilled local fine-tune outperforms prompted Gemini despite being orders of magnitude smaller; and the linguistically harder MSLG→SPA direction, which requires recovering articles, copulas, tense, and gender that are absent from the gloss, benefits the most from frontier-LLM prior knowledge.

Keywords

Mexican Sign Language, Sign Language Glosses, Neural Machine Translation, Low-Resource Translation, Transfer Learning, LoRA, Large Language Models

1. Introduction

Hearing loss is one of the most common sensory impairments worldwide, affecting more than 1.5 billion people, of whom at least 430 million have moderate or higher levels of loss and require rehabilitation or other forms of care [1]. Critically, nearly 80% of those affected live in low- and middle-income countries, where human resources and services for ear and hearing care remain largely inaccessible [1]. In Mexico, over 4 million people face communication barriers due to hearing loss, of whom approximately 1.3 million have severe or profound hearing loss [2]. Mexican Sign Language (MSL) is the primary communication medium of this community and has been recognized as a national language since 2005. Despite this legal status, access to education, public services, and digital communication remains structurally limited for Deaf signers, because the surrounding institutional infrastructure operates almost exclusively in written Spanish. Developing automatic recognition and translation technologies

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
carlos.minutti@infotec.mx (C. Minutti-Martinez); alejandratornaj@gmail.com (A. Torres-Nájera); boris@lapi.unam.mx (B. Escalante-Ramirez); jolveres@lapi.unam.mx (J. Olveres)
https://cminuttim.github.io (C. Minutti-Martinez); https://lapi.fi-p.unam.mx/ (B. Escalante-Ramirez); https://lapi.fi-p.unam.mx/ (J. Olveres)
0000-0002-4002-0773 (C. Minutti-Martinez); 0009-0000-1174-3633 (A. Torres-Nájera); 0000-0003-4936-8714 (B. Escalante-Ramirez); 0000-0002-1514-4520 (J. Olveres)

for MSL thus represents a concrete path toward communicative inclusion for a community that faces both the global burden of unaddressed hearing loss and compounding structural barriers at the local level.

1.1. MSL and its glosses

MSL is a natural language with a distinct grammar and lexicon, independent of Spanish. It is neither a manual encoding of spoken Spanish nor derived from it [3]. Similar to other sign languages, MSL employs a subject–object–verb (SOV) word order, omits copular verbs, lacks articles and gender morphology, and makes extensive use of spatial classifiers [3, 4]. The linguistic distance from other sign systems is also substantial: MSL shares only 23% lexical similarity with American Sign Language (ASL) and has a mutual intelligibility rate of just 14%, underscoring its status as a fully independent linguistic system [3].

Due to the absence of a standardized written form for sign languages, computational research relies on an intermediate representation known as *glosses*: transcriptions in which each sign is rendered as one or more words from the corresponding spoken language, often supplemented by linguistic markers [5]. Glossification transforms a natural signed utterance into a gloss sequence that typically contains fewer tokens than the original and conveys the core propositional content, although it may not fully represent all phonological and prosodic features [4]. For instance, the Spanish sentence *Max fue al parque el jueves* ("Max went to the park on Thursday") is transcribed as the MSLG sequence JUEVES PARQUE MAX IR ("Thursday park Max go"). Despite their limitations, glosses currently provide the only representation of sign languages for which parallel corpora of sufficient size exist to support the training of automatic translation systems [5, 4]. Throughout this paper we use the abbreviation **MSLG** (Mexican Sign Language Glosses) for this written representation, and **SPA** for written Spanish.

1.2. MSL as a low-resource language

In natural language processing, a language is classified as low-resource not only by the number of speakers but primarily by the availability of parallel data for a given language pair [6]. Even languages with extensive monolingual corpora are considered low-resource for machine translation if parallel data are limited. MSL exemplifies this definition: it lacks a standardized written form, has limited electronic resources, and, until recently, lacked a validated parallel corpus of *(MSLG, SPA)* pairs suitable for training neural machine translation models [4, 6].

The quantity of available parallel data is the primary factor correlated with research progress in machine translation for any language [6]. To address this scarcity, methods such as transfer learning, data augmentation, and multilingual modeling have consistently improved performance for low-resource translation pairs [6, 5, 4]. These strategies inform the most competitive MSL translation system designs to date, and are the conceptual basis for the families of approaches that we compare in the rest of the paper.

1.3. State of the art in MSL–Spanish automatic translation

Research on automatic translation between Spanish and MSL can be categorized into rule-based and data-driven approaches. Rule-based systems were initially developed for MSL because insufficient parallel data prevented direct neural training [7]. These systems explicitly encode grammatical rules for both languages, such as removing articles, reordering constituents to match SOV structure, and applying morphological transformations for number and gender. While they can achieve acceptable BLEU scores on small test sets, their scalability is limited, and their coverage relies heavily on manually codified expert knowledge [7].

Data-driven systems, particularly those based on neural machine translation, demonstrate greater promise for scalability. In the context of sign languages, techniques such as data augmentation, transfer learning, and multilingual translation have produced improvements of up to 6 BLEU points over strong baselines [5]. For the Spanish–MSL pair, recent initiatives have concentrated on developing the essential

linguistic resources required to enable neural approaches [4]. Persistent challenges include regional variation within MSL, difficulties in recruiting expert annotators, and the computational complexity of modeling non-manual features such as facial expressions and body posture [7, 5].

1.4. The reference corpus

The most directly relevant published resource for the present work is the Spanish-to-MSLG corpus released by Lara-Ortiz et al. [4, 8], which contains approximately 3,000 parallel *(MSLG, SPA)* sentence pairs covering everyday domains and produced under the supervision of expert annotators. To our knowledge, this is the largest publicly available parallel corpus for the language pair to date. In the same work, the authors report sequence-to-sequence translation results in the MSLG→SPA direction using a relatively compact encoder–decoder model, which establishes the natural data-driven baseline for the language pair. Several of the design decisions taken in our pipeline (the gloss conventions encoded in the system prompt, the negative-example generators used to train our translation-quality scorer, and the warm-up vs. specialization split in our two-phase curriculum) are derived directly from the linguistic conventions documented in that corpus.

1.5. The MSLG-SPA 2026 shared task

The MSLG-SPA 2026 shared task [9], organized as part of IberLEF 2026 [10], represents the first public benchmark specifically designed for bidirectional translation between MSLG and Spanish. The task comprises two complementary subtasks. Subtask A (MSLG→SPA) requires systems to generate fluent and semantically accurate Spanish sentences from MSLG sequences. Subtask B (SPA→MSLG) involves translating Spanish sentences into MSLG sequences, capturing both semantic content and the grammatical features characteristic of MSL, such as SOV order, absence of articles, and lack of copular verbs and gender morphology [3, 4]. Both subtasks are evaluated using standard automatic translation metrics, establishing a common protocol for comparing data-driven and linguistically informed approaches. The task recognizes the unique challenges that sign languages present compared to spoken-language translation, including the visual-gestural modality, rich morphology, and the absence of standardized written forms. For this reason, glosses are adopted as the intermediate symbolic representation that enables computational processing and automatic evaluation [9, 5]. The central difficulty for participants is data scarcity: only 490 parallel sentence pairs are released for training, which is roughly six times smaller than the reference corpus of Section 1.4 and effectively rules out training a neural translation system from scratch on the in-domain data alone.

1.6. Contribution and paper organization

We participated in the MSLG-SPA 2026 shared task as team **AxoloTux**, and approached the bidirectional translation problem by comparing three families of methods that differ in the amount and type of prior knowledge they reuse: (i) a compact encoder–decoder model trained from scratch as a strong baseline, following Lara-Ortiz et al. [4]; (ii) prompting commercial frontier large language models (Gemini 3 Pro and Claude Opus 4.5) with task instructions and a small set of in-context examples; and (iii) parameter-efficient fine-tuning of open-weight decoder-only LLMs using low-rank adaptation (LoRA) [11], training both translation directions jointly under a curriculum that first exposes the model to a noisier broader corpus and then specializes it on the curated challenge data. Our final submissions are drawn from the second and third families. The contributions of the paper are as follows:

- A practical recipe for low-resource bidirectional gloss-to-text translation that combines (i) a two-phase curriculum that warms the model up on the broader external corpus before specializing it on the curated challenge data, and (ii) distillation from frontier closed-source LLMs to expand the in-domain supervision.

- A reference-free translation-quality scorer based on interpretable surface, morphosyntactic, and MSLG-specific cross-modal features, used both to rank candidate models during development (when test-set references were withheld) and to ensemble submission variants at inference time.
- An empirical, per-direction comparison of the three families of approaches on the same task, which exposes asymmetric strengths of prompted frontier models, fine-tuned open-weight LLMs, and traditional encoder–decoder baselines, and which we discuss with a view to other low-resource gloss-to-text translation problems.
- A public release of the system code, prompts, and training configuration at https://github.com/cminuttim/MSLG-SPA_2026, to support reproducibility and reuse on other low-resource sign language gloss translation tasks.

Our final system obtained first place in the SPA→MSLG track, second place in the MSLG→SPA track, and ranked first in the global ranking of the shared task. The remainder of the paper is organized as follows. Section 2 describes the candidate models, the translation-quality scorer used for model selection, and the training corpora and configurations for the final submissions. Section 3 reports the official evaluation results. Section 4 analyzes these results and the design choices behind them, and Section 5 concludes.

2. Methodology

Our work proceeded in two consecutive stages. **Stage 1 (Section 2.1)** performs *best-model identification*: candidates from each of the three families introduced in Section 1 are trained or prompted and ranked using an internally trained translation-quality classifier, since the ground-truth references for the official test set were withheld throughout the challenge. **Stage 2 (Section 2.2)** performs *variant generation*: the model selected in Stage 1 is fine-tuned under several training-data and prompting configurations to produce the final set of submissions.

2.1. Stage 1: Best-Model Identification

2.1.1. Candidate models

We considered three families of approaches. The first is a compact encoder-decoder model, following the baseline of Lara-Ortiz et al. [4]: a Spanish-pretrained BART model (BARTO) [12, 13], which is bidirectional and autoregressive and was originally pretrained on a SPA-to-SPA denoising objective. Because the encoder-decoder formulation cannot easily condition on a direction tag through a system prompt, BARTO was trained only for MSLG→SPA, which is the direction reported in the reference paper.

The second family comprises open-weight decoder-only large language models, all fine-tuned with LoRA [11] on 4-bit QLoRA-style quantization [14]: LLaMA 3.2 3B [15], SalamandraTA-7b-instruct [16, 17] (an instruction-tuned translation model from the Barcelona Supercomputing Center, proficient in 40 languages and trained for sentence-, paragraph- and document-level translation), Qwen 3.5 4B [18], and Qwen 3.5 9B [18]. For Qwen we used the *abliterated* variant of Qwen 3.5 9B released by `huihui-ai` [19], which removes refusal behaviors that would otherwise interfere with translating sentences containing medical or other terms (e.g., SORDO, CIEGO, EMBARAZADA, ALCOHOL, AUTISMO) that general-purpose alignment occasionally edits or softens. Unlike the BARTO baseline, the LLMs are decoder-only and so the source/target direction is conveyed through the prompt, allowing both MSLG→SPA and SPA→MSLG to be learned jointly.

The third family is direct prompting of frontier closed-source LLMs: Gemini 3 Pro [20] and Claude Opus 4.5 [21]. These models received the task instructions in the system prompt together with a small set of in-context examples, and the in-domain training pairs were used only to inform the in-context demonstrations rather than to update any parameters.

The full list of candidate models with the citations used in this paper is summarized in Table 1.

Table 1
Candidate models evaluated in Stage 1.

Family	Model	Reference
Encoder-decoder	BARTO (Spanish BART)	[12, 13, 4]
Fine-tuned LLM	LLaMA 3.2 3B (4-bit QLoRA)	[15]
Fine-tuned LLM	SalamandraTA 7B-instruct (4-bit QLoRA)	[16, 17]
Fine-tuned LLM	Qwen 3.5 4B (4-bit QLoRA)	[18]
Fine-tuned LLM	Qwen 3.5 9B abliterated (4-bit QLoRA)	[18, 19]
Fine-tuned LLM	Qwen 3.5 9B abliterated (8-bit QLoRA)	[18, 19]
Prompted LLM	Gemini 3 Pro	[20]
Prompted LLM	Claude Opus 4.5	[21]

2.1.2. Translation-quality scoring without references

The official test set consists of 245 sentences per direction whose reference translations were not released during the challenge, so neither standard reference-based metrics (BLEU, chrF, COMET) nor a held-out evaluation loss can be used to compare candidate models. We therefore trained a binary classifier that predicts whether a $(\text{MSLG}, \text{SPA})$ pair is a correct translation and used its calibrated probability $p_{\text{correct}} \in [0, 1]$ as a proxy quality score.

Negative-example generation. Following standard practice in synthetic data augmentation, we constructed 490 *negative* translations by perturbing the 490 training pairs. For each pair we selected one of the two sides at random and applied one to three of the following operations: random token *deletion*, frequency-weighted token *insertion* or *substitution*, *synonym replacement* from manually curated MSLG and Spanish dictionaries, *temporal swap* on time-marker tokens and the corresponding Spanish tense (e.g. swapping AYER/MAÑANA and compré/compraré), *compound swap* or *split* on MSLG '+'-compounds (e.g. MAMÁ+PAPÁ → PAPÁ+MAMÁ), and *gender* or *number flips* on Spanish nouns and articles. A complementary set of *hard* negatives was produced by TF-IDF cosine pairing of distinct sentences whose Spanish sides share lexical material but are not paraphrases. Combined with the original 490 positive pairs, this yields a balanced binary classification corpus.

Features. We extracted 54 features organized in three groups: (i) *surface and structural* features (21) including MSLG and Spanish lengths and their ratio, counts of '+'-compounds, hyphenated multi-token signs, and fingerspelled tokens (dm- and #), Boolean and agreement indicators for negation, temporal markers and interrogatives, fingerspelled-name overlap with the Spanish sentence, and two prefix-based lexical similarities (Jaccard overlap and asymmetric MSLG-prefix coverage); (ii) *Spanish morphosyntactic* features (27) computed with spaCy [22] using the `es_core_news_md` model, comprising counts and ratios for the main POS classes, gender and number distributions, coarse tense counts, reliability-filtered binary flags for present/past indicative, participle, conditional, gerund and subjunctive forms, infinitive counts, named-entity counts including a PER-entity flag, and presence flags for proper nouns, feminine human nouns and intensity adverbs; and (iii) *cross-modal alignment* rules (6) that fire when an MSLG construction has the expected Spanish counterpart, encoded as soft agreement scores in $\{0, 0.5, 1\}$ to penalize one-sided presence: temporal-marker ↔ verb-tense agreement, dm-name ↔ PER-entity / textual mention, MSLG-vs-Spanish content-word ratio, +MUJER compound ↔ feminine human noun, MUCHO/TANTO ↔ Spanish intensifier (*muy*, *mucho*, *bastante*, *tan*), and dm- ↔ PROPN.

The full feature set was reduced to the top 20 features by mutual information with the label using `SelectKBest` [23] before model fitting.

Classifier selection. On the 974-pair balanced corpus (487 positive, 487 negative) we trained logistic regression, an RBF support vector machine, a random forest, a gradient-boosting classifier and a histogram gradient-boosting classifier, using stratified 5-fold cross-validation with feature standardization

Table 2
Stratified 5-fold cross-validated performance of the translation-quality classifiers on the 974-pair balanced corpus.

Classifier	ROC-AUC	PR-AUC
Logistic Regression	0.778 ± 0.037	0.782 ± 0.038
SVM (RBF)	0.783 ± 0.026	0.792 ± 0.030
Random Forest	0.794 ± 0.032	0.800 ± 0.039
Gradient Boosting	$\mathbf{0.803 \pm 0.038}$	$\mathbf{0.801 \pm 0.042}$
Hist. Gradient Boosting	0.799 ± 0.034	0.800 ± 0.037

fit inside each fold. Results are summarized in Table 2. The gradient-boosting model achieved the best mean ROC-AUC (0.803 ± 0.038) and was retained as the *translation-quality scorer* for the rest of the paper.

We emphasize that the scorer is a proxy: it can be fooled by translations that satisfy its surface rules without being faithful, and it cannot evaluate dimensions that the official COMET-based evaluation captures, such as cross-lingual semantic adequacy [24]. However, it has two properties that were essential during model selection. First, it can be applied to the 245-sentence test set without reference translations. Second, unlike COMET its features are interpretable and grounded in MSLG gloss conventions (compounds, fingerspelling, temporal markers, intensifiers), so an inspection of feature importance directly suggests which translation phenomena a model handles correctly.

Feature importance. Table 3 lists the contribution of each of the 20 pre-selected features under the gradient-boosting scorer refit on the full balanced corpus. The distribution is heavily concentrated: the top feature alone accounts for roughly half of the total importance, and the top five features together account for about three quarters of it. We describe the most informative features individually below.

- `jaccard_sim` **(rank 1, 49.9%)** measures the Jaccard overlap between the sets of length-4 lowercased prefixes extracted from MSLG signs and Spanish content tokens, after removing stop words and normalizing MSLG markup (compounds, fingerspelling). Prefixes are used instead of full word forms so that morphological variants such as COMPRAR and *compré* are still matched on `comp`. A faithful translation should share most content stems with its source, so this feature dominates as a sanity check on lexical overlap.
- `spa_noun_ratio` **(rank 2, 10.3%)** is the fraction of Spanish tokens whose spaCy POS tag is NOUN or PROPN. MSLG sentences are heavily nominal because articles, copulas and auxiliaries are absent; the Spanish side must therefore have a comparable nominal density. Very high or very low ratios flag truncated, over-explained or off-topic outputs.
- `len_ratio` **(rank 3, 9.5%)** is the ratio of MSLG token count to Spanish token count. Spanish is consistently longer than its MSLG transcription because it adds function words that the gloss omits, so well-formed pairs cluster around a stable ratio; gross deviations indicate truncation, padding or hallucination.
- `len_mslg` **(rank 4, 7.5%)** is the absolute MSLG token count, which the scorer uses jointly with `len_ratio` to calibrate scale-dependent expectations (e.g., short imperatives behave differently from multi-clause utterances).
- `spa_n_verbs` **(rank 5, 4.7%)** is the number of Spanish VERB or AUX tokens. Most pairs have one main predicate; outputs with zero verbs (nominal-only responses) or many verbs (run-on or repeated clauses) are penalized.
- `prefix_overlap` **(rank 6, 3.4%)** is the asymmetric counterpart of `jaccard_sim`, defined as the fraction of MSLG content prefixes that are also present in the Spanish side. Whereas Jaccard is symmetric and rewards mutual overlap, this feature specifically penalizes Spanish outputs that drop MSLG content (a common failure mode for MSLG→SPA).

Table 3
Feature importance of the gradient-boosting translation-quality scorer (fit on the full 974-pair balanced corpus over the 20 pre-selected features). Groups: V1 = surface and structural; spaCy = Spanish morphosyntax; cross = cross-modal alignment rule.

Rank	Group	Feature	Importance
1	V1	`jaccard_sim`	0.499
2	spaCy	`spa_noun_ratio`	0.103
3	V1	`len_ratio`	0.095
4	V1	`len_mslg`	0.075
5	spaCy	`spa_n_verbs`	0.047
6	V1	`prefix_overlap`	0.034
7	cross	`rule_mucho_intensity_agree`	0.025
8	spaCy	`spa_has_fem_human`	0.019
9	V1	`len_spa`	0.017
10	spaCy	`spa_n_nouns`	0.015
11	cross	`rule_mujer_fem_agree`	0.014
12	V1	`has_interrog_mslg`	0.013
13	spaCy	`spa_n_pres`	0.012
14	V1	`temp_match`	0.009
15	spaCy	`spa_n_entities`	0.007
16	V1	`neg_match`	0.006
17	spaCy	`spa_has_subj`	0.006
18	V1	`has_neg_spa`	0.003
19	spaCy	`spa_n_fut`	0.002
20	spaCy	`spa_has_cnd`	0.000

- `rule_mucho_intensity_agree` **(rank 7, 2.5%)** is the highest-ranked cross-modal rule. It fires when MSLG contains a member of the MUCHO/TANTO intensifier family and checks whether the Spanish side contains a matching intensifier (*muy*, *mucho*, *bastante*, *tan*), returning a soft agreement score in $\{0, 0.5, 1\}$.
- `spa_has_fem_human` **(rank 8, 1.9%)** and `rule_mujer_fem_agree` **(rank 11, 1.4%)** jointly encode the MSLG convention that feminine human referents are marked by a +MUJER compound on a base masculine sign: HERMANO+MUJER ↔ *hermana*, HIJO+MUJER ↔ *hija*. The first feature merely detects the Spanish side; the rule checks the bilateral agreement.

The remaining features (ranks 9–20) carry small but non-zero weight and cover Spanish length, named-entity counts, additional tense and mood flags (present indicative, subjunctive, future, conditional, participle), and the symmetry indicators for interrogatives, temporal markers and negation. The overall structure of the table is therefore that the scorer rewards candidate models for producing Spanish of the right size and lexical shape first, and for honoring MSLG-specific lexical contracts second. We note that this also explains why the scorer is a useful but imperfect proxy: a fluent paraphrase that respects length and stem overlap may receive a high score even when it silently swaps a word for a near-synonym, while a literal but ungrammatical word-by-word translation may also score well. We return to this caveat in Section 4.

2.1.3. Comparison of candidate models

Each candidate model was used to translate the 245 sentences of the MSLG→SPA test set, which we adopted as the harder of the two directions because gloss-to-Spanish requires recovering material (articles, copulas, tense, gender, periphrastic future) that is absent from the gloss. Each (source MSLG, predicted SPA) pair was scored with the gradient-boosting scorer of Section 2.1.2. Table 4 reports the fraction of pairs predicted correct at the default 0.5 threshold and the mean probability $\bar{p}_{\text{correct}}$.

Table 4
Translation-quality scoring on the 245-sentence MSLG→SPA test set. References are unavailable, so the classifier of Section 2.1.2 is used to predict p_{correct} on each model's own output.

Model	% predicted correct	$\bar{p}_{\text{correct}}$ (std)
BARTO	56.7%	0.519 (0.229)
LLaMA 3.2 3B (4-bit)	70.6%	0.605 (0.223)
SalamandraTA 7B (4-bit)	72.2%	0.629 (0.235)
Qwen 3.5 4B (4-bit)	73.9%	0.623 (0.220)
Qwen 3.5 9B (4-bit)	72.2%	0.631 (0.220)
Qwen 3.5 9B (8-bit)	**74.3%**	**0.646** (0.218)

Two observations drove the selection of the final base model. First, the LLM-based systems all outperform the encoder-decoder baseline by 14–18 percentage points of predicted-correct fraction, which is consistent with our hypothesis that the limited supervision available in this challenge favours models that can transfer broad prior linguistic knowledge. Second, within the Qwen 9B family, qualitative inspection of hard examples revealed that the LoRA-fine-tuned 4-bit variant was outperformed by the same base model served by `ollama` with carefully written prompts and no fine-tuning at all. Further tests traced the gap to a quantization mismatch: the 4-bit GGUF format used by `ollama` preserves more capability than the 4-bit BitsAndBytes [25] NF4 quantization used during training. Re-running the fine-tune with LLM.int8() 8-bit quantization recovered and slightly surpassed the prompted baseline (last row of Table 4). We therefore adopted the abliterated Qwen 3.5 9B model under 8-bit quantization as the base model for Stage 2.

2.2. Stage 2: Variant Generation

2.2.1. Training corpora

Three corpora were combined across our Stage 2 submissions:

- **Base corpus** (3,000 pairs): the Spanish-to-MSLG glosses • corpus released by Lara-Ortiz et al. [8]. It is noisier and broader in lexical coverage than the challenge corpus and is used for a single warm-up epoch when included.
- **Challenge corpus** (490 pairs): the official MSLG-SPA 2026 training data, treated as the curated specialization corpus and used for three epochs when included.
- **Extended corpus** (Challenge + 950 augmented pairs): the challenge corpus augmented with 950 additional sentence pairs generated by Claude Opus 4.5 (475 pairs) and Gemini 3 Pro (475 pairs), prompted to produce new *(MSLG, SPA)* examples consistent with the gloss conventions described in Section 2.2.2. This is a form of distillation: the closed-source LLMs are used to expand the curated data with plausible additional examples, which the open-weight LLM then consumes during fine-tuning.

2.2.2. Prompt format and system instruction

Both translation directions are unified under a single JSON input/output protocol so that one fine-tuned model can serve both sub-tasks. The user message is a JSON object with three fields: `"fuente"` (the source language, `"MSLG"` or `"SPA"`), `"texto"` (the input text), and `"objetivo"` (the target language). The assistant must emit a single JSON object with both `"MSLG"` and `"SPA"` fields, where the field corresponding to the source side is echoed verbatim and the field corresponding to the target side is the translation. During training the loss is masked over the prompt tokens so it is computed only over the JSON completion. For Qwen 3.5 the prompt uses the model's native chat tokens (`<|im_start|>` / `<|im_end|>`), and the thinking mode is disabled by injecting an empty

<think>\n\n</think>\n\n block before the JSON completion so that the model emits the translation directly without an intermediate reasoning trace.

The system prompt itself encodes a compact reference grammar of MSLG. It declares that MSLG is a language with its own grammar (not a variant of Spanish) and that the translator must be literal, in particular not softening, omitting or reformulating terms that general-purpose alignment might edit (medical, descriptive of disability, or colloquial). It then defines the MSLG written conventions used in the corpus: hyphenated multi-word signs (YA-VEO, TARJETA-DE-CRÉDITO), '+'-compound signs (MAMÁ+PAPÁ, HERMANO+MUJER), fingerspelled borrowings (#TV, #SEP) and manually fingerspelled proper nouns (dm-LUIS), reduplication marking plural or intensity, omission of copulas and articles, sentence-final or NO- negation, perfective marking with YA, temporal markers at sentence start and wh-questions at sentence end. The prompt closes with six worked MSLG↔SPA examples that illustrate each rule, followed by the JSON input/output schema and one canonical example of it.

2.2.3. Implementation details

The base model is the abliterated Qwen 3.5 9B variant `huihui-ai/Huihui-Qwen3.5-9B-abliterated` [19] loaded with `transformers` [26] and quantized to 8-bit with the LLM.int8() algorithm [25] of the BitsAndBytes library. LoRA adapters [11] are inserted into all linear projection matrices of the attention and MLP blocks (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with rank $r = 16$, scaling $\alpha = 32$, dropout 0.10 and no bias term. The choice of $r = 16$ trades expressive capacity for regularization given the small specialization corpus, and $\alpha = 2r$ follows common practice.

Training uses a per-device batch size of 1 with gradient accumulation over 16 steps (effective batch size 16), learning rate 1×10^{-4} with a 5% linear warm-up and weight decay 0.01, in mixed precision (bf16 on Ampere-or-later GPUs, fp16 otherwise). The maximum sequence length is 1024 tokens, which covers the system prompt (about 700 tokens) together with the JSON input and the completion. Examples are emitted in both directions for every pair so the two sub-tasks contribute equally to each batch. When the base corpus is used, it constitutes a single warm-up epoch (Phase 1) on which a fresh LoRA adapter is trained; the resulting adapter is then specialized for three further epochs (Phase 2) on either the challenge corpus or the extended corpus. Three intermediate checkpoints per epoch are saved during Phase 2; the checkpoint with the lowest evaluation loss on a 10% random held-out split of the corresponding corpus is retained as the final adapter, although as noted in Section 2.1.3 the evaluation loss is only weakly informative across runs because the validation split changes with the corpus.

At inference time we use beam search with four beams and greedy beam selection (`do_sample=False`), up to 128 new tokens, and the JSON output is parsed back into the MSLG and SPA fields. Beam search was preferred over sampling because the task is near-deterministic at the sentence level and precision dominates creativity.

All fine-tuning and inference for the open-weight Qwen 3.5 9B runs were performed on a single consumer-grade NVIDIA RTX 4090 (24 GB VRAM). The 8-bit quantization keeps the base model at roughly 10 GB, leaving sufficient memory for the LoRA adapter, the optimizer state and a maximum sequence length of 1024 tokens with gradient accumulation (effective batch size 16, per-device batch size 1). One Phase 2 specialization run of 3 epochs on the 490-pair challenge corpus completes in under one hour on this hardware.

2.2.4. Submitted variants

We submitted six systems per direction, identified by the prefixes listed in Table 5. Each prefix is paired with _MSLG2SPA or _SPA2MSLG to indicate the direction. The QDistilMix system is an ensemble that scores every candidate translation produced by all the other submissions using the gradient-boosting scorer of Section 2.1.2 and, for each test sentence, retains the candidate with the highest p_{correct}.

Table 5
Submitted systems. *Q* denotes the abliterated Qwen 3.5 9B base in 8-bit quantization; *C* denotes Claude Opus 4.5; *G* denotes Gemini 3 Pro.

Prefix	Approach	Training data
`AxoloTux_QBase_`	Q + LoRA fine-tune	Challenge (3 epochs)
`AxoloTux_QDistil2_`	Q + LoRA fine-tune	Base (1 epoch) + Extended (3 epochs)
`AxoloTux_QDistilTest_`	Q + LoRA fine-tune	Challenge (3 epochs) + Claude/Gemini predictions on test set
`AxoloTux_QDistilMix_`	Ensemble selected by scorer	n/a (post-hoc over the other submissions)
`AxoloTux_G_`	Gemini 3 Pro prompted	Challenge as in-context examples only
`AxoloTux_C_`	Claude Opus 4.5 prompted	Challenge as in-context examples only

Table 6
Official global ranking of the 19 participating teams. The two right-most columns show each team's best Task Score per direction, on which the Global Score is computed.

Rank	Team	Global	MSLG→SPA	SPA→MSLG
1	AxoloTux	**1.549**	1.340	**1.758**
2	VerbaNexAI	1.240	1.034	1.447
3	UAM_ChineseRoom	1.231	1.189	1.272
4	CICESE-LCDAA	1.175	**1.342**	1.007
5	UCOUCPlenitas	0.648	1.063	0.233
6	RETUYT-INCO	0.387	0.204	0.571
7	ComunicadorInnato	0.263	0.881	−0.355
8	LSM-INAOE	0.262	0.411	0.113
9	FisBioUNAM	0.236	0.372	0.100
10	ReCore	0.153	−0.117	0.422
11	DunderMifflin	−0.094	−0.229	0.041
12	MBM_Solution	−0.287	−0.510	−0.064
13	Bortolotii	−0.312	−0.464	−0.161
14	WangKongqiang	−0.497	−0.661	−0.334
15	SignNexus	−0.569	−0.530	−0.608
16	SeñLP	−0.599	−0.702	−0.496
17	TeamCICESE	−1.271	−1.400	−1.142
18	SeñasDelDesierto	−1.412	−1.332	−1.492
19	Lepopardo	−1.619	−1.539	−1.700

3. Results

The MSLG-SPA 2026 shared task [9] received submissions from 19 teams, each entering one or more runs in the two sub-tasks. The official Task Score for each direction is a per-run standardized average of BLEU, METEOR and chrF (and additionally COMET for MSLG→SPA), and the Global Score per team averages a team's best Task Score on each direction. We received 6 runs per direction (the variants listed in Table 5), competing against 48 submitted runs in each direction.

3.1. Global ranking

The official global ranking is reported in Table 6. AxoloTux obtained a Global Score of 1.549, ranking first by a comfortable margin (the second team, VerbaNexAI, scored 1.240). Among the top four teams, AxoloTux is also the only one to lead in a sub-task: VerbaNexAI and UAM ChineseRoom rank strictly behind AxoloTux on both tracks, and CICESE-LCDAA, which obtained the top Task Score on MSLG→SPA, ranks fourth overall due to a comparatively weaker SPA→MSLG submission.

Table 7
AxoloTux submissions to the MSLG→SPA track, ranked against the 48 submitted runs. Rank is the official Task Rank of each run.

Rank	Entry	Task Score	BLEU	METEOR	chrF	COMET
2	`AxoloTux_C_`	**1.340**	**52.70**	**0.727**	**76.40**	**0.902**
3	`AxoloTux_QDistilTest_`	1.283	51.15	0.714	76.08	0.895
5	`AxoloTux_G_`	1.143	47.10	0.691	73.94	0.885
6	`AxoloTux_QDistilMix_`	1.121	46.76	0.685	73.01	0.888
12	`AxoloTux_QDistil2_`	1.017	45.18	0.650	70.92	0.885
13	`AxoloTux_QBase_`	1.009	45.24	0.646	70.82	0.884

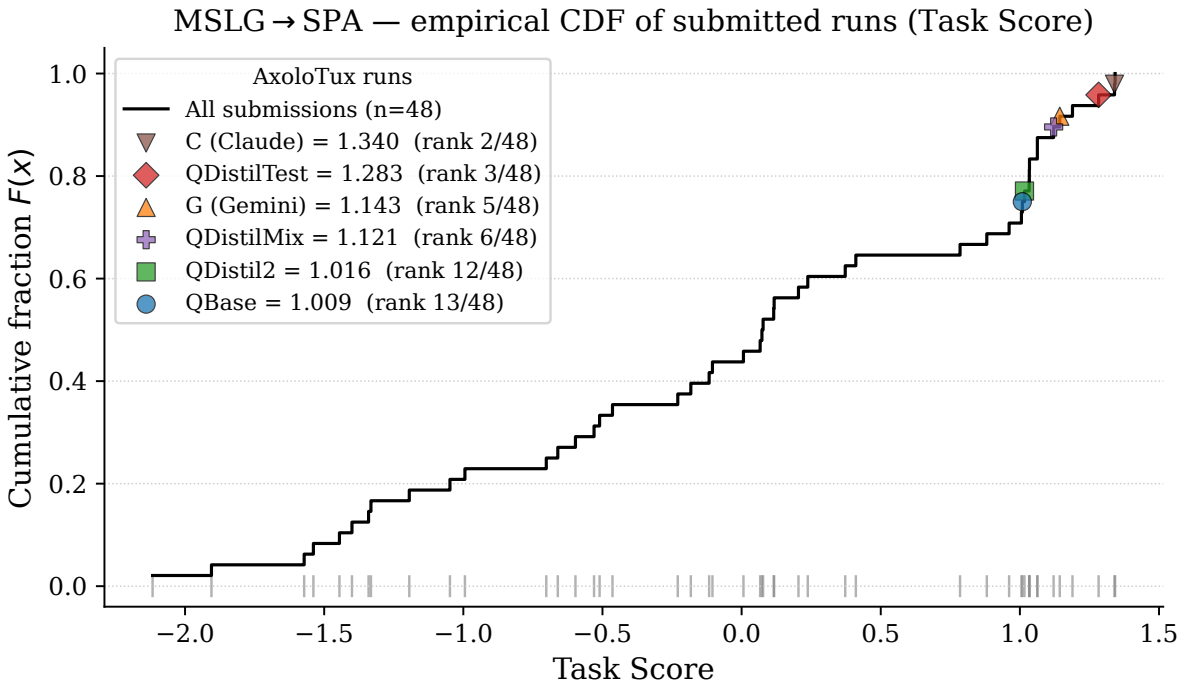

Figure 1: Empirical CDF of the MSLG→SPA Task Score over the 48 submitted runs, with the six AxoloTux runs annotated on the curve.

3.2. MSLG→SPA track

Table 7 summarizes the position of our six runs on the MSLG→SPA track. The best AxoloTux submission on this direction is the Claude Opus 4.5 prompted system (`AxoloTux_C_`), which ranks 2nd out of 48 submitted runs with a Task Score of 1.340. The LoRA-fine-tuned Qwen 9B distilled on Claude/Gemini predictions (`AxoloTux_QDistilTest_`) is the next AxoloTux entry at rank 3 with Task Score 1.283, closely followed by Gemini 3 Pro at rank 5. The two fine-tunes that do not use frontier-LLM distillation (`QDistil2` on Base+Extended and `QBase` on Challenge only) trail at ranks 12 and 13. The post-hoc ensemble `QDistilMix` sits between them at rank 6 because, on this direction, the scorer picks the better of two close candidates rather than producing an improvement of its own.

Figure 1 situates these six runs on the empirical CDF of the Task Score over all 48 runs. The two best AxoloTux entries (`C` and `QDistilTest`) fall in the top right tail of the curve, well above the bulk of the distribution; the bottom two (`QDistil2`, `QBase`) sit slightly above the median.

A different picture emerges from the COMET metric [24], the only neural reference-based metric in the evaluation. Table 8 re-sorts the same six runs by COMET, and Figure 2 shows their position on the empirical CDF of COMET over the field. Under COMET, `AxoloTux_C_` is now the top run overall, and all six AxoloTux entries are in the top fifth of the field: the *worst* of our submissions on COMET

Table 8
AxoloTux MSLG→SPA submissions sorted by COMET. *COMET Rank* and *chrF Rank* are computed over the 48 submitted runs.

Entry	Task Rank	Task Score	COMET	COMET Rank	chrF	chrF Rank
`AxoloTux_C_`	2	1.340	**0.902**	**1**	76.40	2
`AxoloTux_QDistilTest_`	3	1.283	0.895	3	76.08	3
`AxoloTux_QDistilMix_`	6	1.121	0.888	5	73.01	6
`AxoloTux_QDistil2_`	12	1.017	0.885	8	70.92	12
`AxoloTux_G_`	5	1.143	0.885	9	73.94	4
`AxoloTux_QBase_`	13	1.009	0.884	10	70.82	13

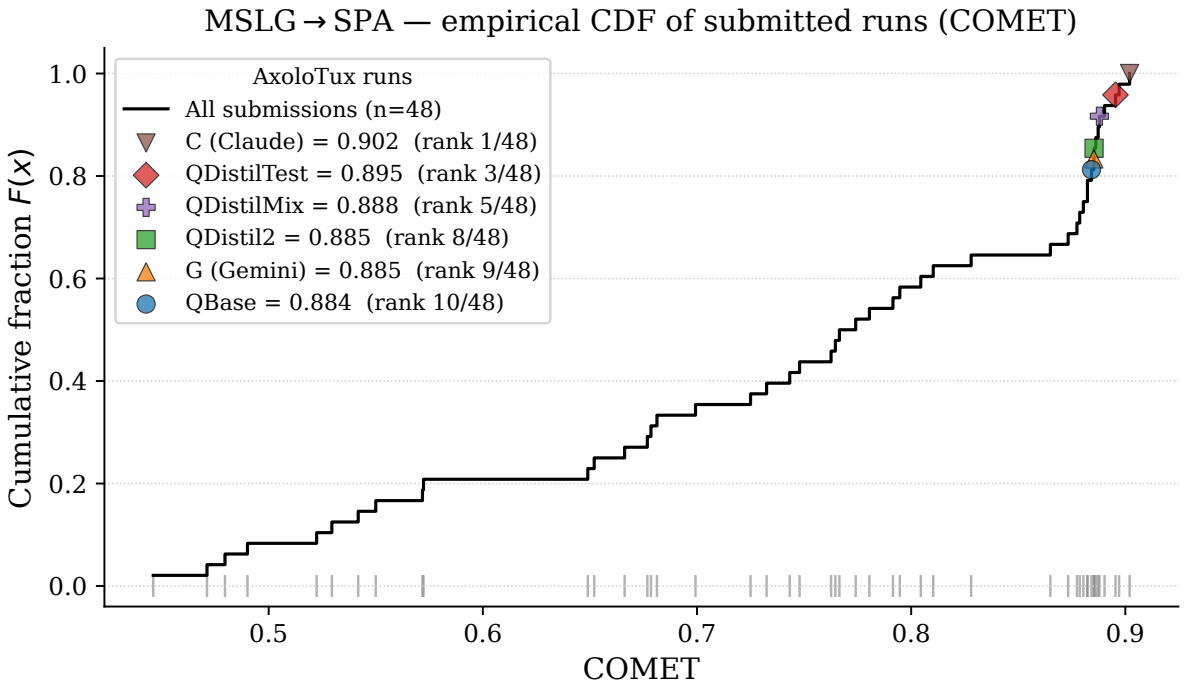

Figure 2: Empirical CDF of MSLG→SPA COMET over the 48 submitted runs. All six AxoloTux runs lie in the top fifth of the COMET distribution, and `AxoloTux_C_` is the top run on this metric.

(`QBase`, rank 10/48) still lands at the 79th percentile of the distribution. By contrast, the Gemini and the `QDistil2` entries lose 7–8 positions when re-sorted by COMET vs. Task Score, suggesting that they translate the semantic content well but trade chrF/BLEU points by producing surface forms that diverge more from the references; `QDistilTest` and `QDistilMix` are near-monotonic across the two metrics.

3.3. SPA→MSLG track

On the SPA→MSLG direction, Claude Opus 4.5 ranks first overall with a Task Score of 1.758, more than 0.2 above the second-ranked run (Table 9 and Figure 3). The LoRA-fine-tuned Qwen 9B distilled on Claude/Gemini predictions (`QDistilTest`) again comes in second. Notable in this direction is the strong showing of `QDistil2`, the variant warmed up on the 3,000-pair Lara-Ortiz corpus and specialized on the extended challenge corpus: it ranks 4th on SPA→MSLG, eight places higher than its rank on MSLG→SPA. Conversely, Gemini 3 Pro drops to rank 14, a loss of nine places compared with its MSLG→SPA standing; its SPA→MSLG Task Score (0.580) is roughly half of its MSLG→SPA score. The ensemble `QDistilMix` is the worst of our runs on this direction at rank 8.

Table 9
AxoloTux submissions to the SPA→MSLG track, ranked against the 48 submitted runs.

Rank	Entry	Task Score	BLEU	METEOR	chrF
1	`AxoloTux_C_`	**1.758**	**26.49**	**0.597**	**68.29**
2	`AxoloTux_QDistilTest_`	1.554	24.42	0.581	66.55
4	`AxoloTux_QDistil2_`	1.338	23.24	0.549	64.39
7	`AxoloTux_QBase_`	1.100	19.28	0.540	64.06
8	`AxoloTux_QDistilMix_`	1.055	18.17	0.541	64.35
14	`AxoloTux_G_`	0.580	12.89	0.494	61.90

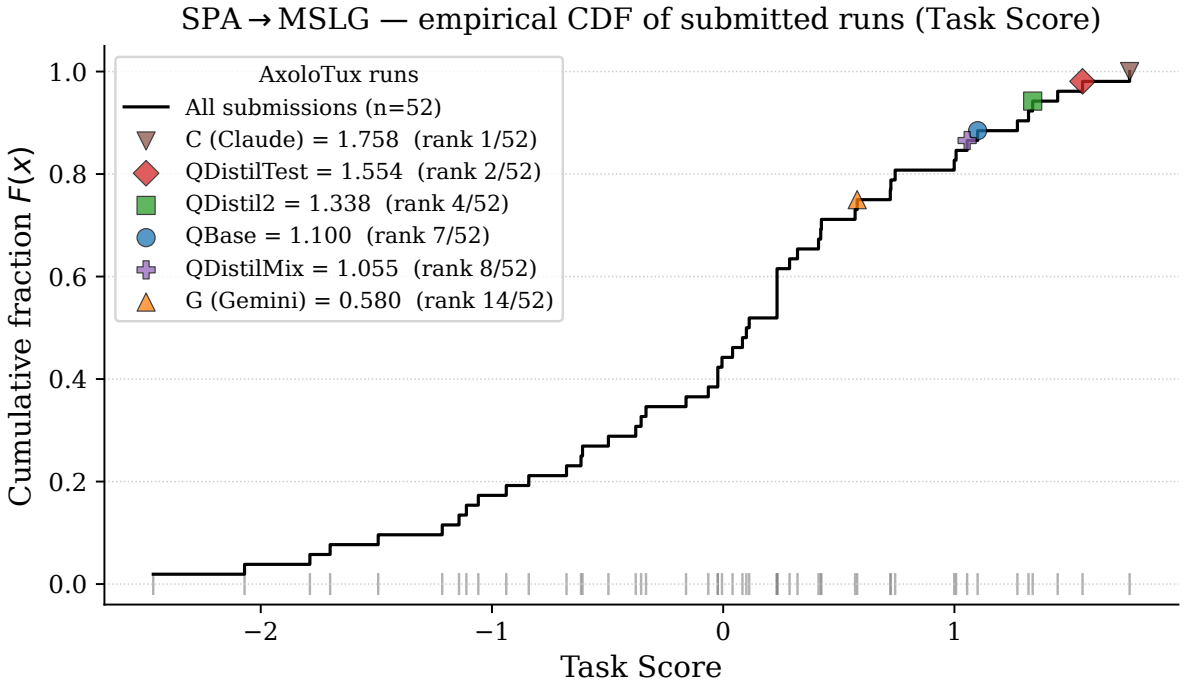

Figure 3: Empirical CDF of the SPA→MSLG Task Score over the 48 submitted runs, with the six AxoloTux runs annotated on the curve.

3.4. Per-approach summary across both tracks

Table 10 consolidates the six AxoloTux approaches by averaging their Task Rank and Task Score across the two directions, providing a single ordering of our submissions. Sorted by the average rank, the prompted Claude Opus 4.5 (C) leads with $\bar{r} = 1.5$ and $\bar{s} = 1.549$, followed by `QDistilTest` ($\bar{r} = 2.5$, $\bar{s} = 1.419$). The remaining four approaches form a second tier: `QDistilMix` ($\bar{r} = 7.0$), `QDistil2` ($\bar{r} = 8.0$), prompted Gemini ($\bar{r} = 9.5$), and the challenge-only fine-tune `QBase` ($\bar{r} = 10.0$). The per-direction ranks reveal that the second tier is not flat: the two prompted frontier LLMs and the two non-distilled LoRA fine-tunes each show large rank differences between the two directions, while `QDistilTest` and `QDistilMix` are close to symmetric. A detailed discussion of these patterns is deferred to Section 4.

4. Discussion

The Results section reports per-track and per-approach numbers; here we use those numbers to highlight what they imply for the methodological choices behind a low-resource gloss-to-text translation system. We organize the discussion around four observations: (i) frontier commercial LLMs are not uniformly strong across directions, (ii) MSLG→SPA is the linguistically harder direction despite producing larger

Table 10
Average ranking of the six AxoloTux approaches across the two tracks. Per-direction ranks are taken from the official ranking of the 48 submitted runs in each direction; *Avg Rank* and *Avg Score* are the arithmetic mean of the two corresponding columns.

Approach	M→S Rank	M→S Score	S→M Rank	S→M Score	Avg Rank	Avg Score
C (Claude)	2	1.340	1	1.758	**1.5**	**1.549**
QDistilTest	3	1.283	2	1.554	2.5	1.419
QDistilMix	6	1.121	8	1.055	7.0	1.088
QDistil2	12	1.017	4	1.338	8.0	1.177
G (Gemini)	5	1.143	14	0.580	9.5	0.861
QBase	13	1.009	7	1.100	10.0	1.055

surface-metric values, (iii) COMET tells a substantially different story than the n-gram-based metrics, and (iv) for an in-domain low-resource pair, a local LoRA-fine-tuned LLM is a competitive, and in some respects preferable, alternative to a frontier API.

4.1. Commercial frontier LLMs are not uniformly strong

A common assumption in low-resource translation work is that prompting a frontier LLM provides a strong, near-zero-cost baseline on any language pair, simply because the model has absorbed broad multilingual signal during pretraining. Our results only partially support this assumption. Claude Opus 4.5 (C) ranked 2nd on MSLG→SPA and 1st on SPA→MSLG, with an average rank of 1.5 (Table 10), confirming the assumption for that particular model. Gemini 3 Pro (G), in contrast, ranked 5th on MSLG→SPA but dropped to 14th on SPA→MSLG, with a Task Score (0.580) that is roughly half of its MSLG→SPA value. The drop reflects systematic failures on the reverse direction: Gemini tends to over-generate glosses, keep Spanish function words in the output, and ignore the documented MSLG conventions (+-compounds, fingerspelled-name marking, sentence-final wh-words) even when they are included in the prompt as in-context examples. We could not identify a prompt-engineering fix that closed the gap with Claude.

The practical implication is that the prompted-LLM family cannot be treated as a single class of system: which specific provider one prompts matters substantially, especially in the spoken-language → gloss direction, where the target side is the under-resourced one.

4.2. MSLG→SPA is the linguistically harder direction

The raw surface metrics (BLEU, METEOR, chrF) are systematically higher on MSLG→SPA than on SPA→MSLG: e.g. chrF 76.4 vs. 68.3 on our best system. This is not evidence that MSLG→SPA is the easier direction; rather, it reflects properties of the target side and of the evaluation. In the MSLG→SPA direction the target is Spanish, with a larger and more redundant vocabulary, longer surface forms, and many high-frequency function words that an n-gram-based metric is likely to match by chance. In SPA→MSLG the target is a short gloss sequence drawn from a restricted, partly expert-controlled vocabulary, where each token carries more information and a single substitution or reordering error is penalized harshly.

From a generation perspective, however, MSLG→SPA is the harder problem: it requires the system to *recover* information that the source side does not encode. Spanish articles, copulas, periphrastic future, gender and number agreement, and overt subjects in pro-drop contexts are all absent from a well-formed MSLG sequence, and must be reconstructed from world knowledge and from sentence-level context. SPA→MSLG, by contrast, is mostly a *deletion-and-reordering* task: the system has to drop articles and copulas, apply SOV reordering, mark feminine human referents with a +MUJER compound, and move wh-words and temporal markers to their canonical positions. Both directions are non-trivial, but the information-theoretic asymmetry is reflected in the much wider score range observed on MSLG→SPA

across the field (Task Score ranging from -1.5 to $+1.4$), and is consistent with our choice in Section 2.1.3 to use MSLG→SPA as the direction in which to compare candidate models.

4.3. COMET as the more informative metric

The official Task Score on MSLG→SPA aggregates BLEU, METEOR, chrF and COMET [24] in a standardized average, so the n-gram-based metrics (which share a single source of evidence, lexical overlap with the reference) collectively dominate. This matters because surface-metric agreement is a noisy proxy of translation quality for languages with morphology-rich inflection and free word order on the source side (here Spanish): two paraphrases with identical meaning can differ substantially in their BLEU or chrF score against a single reference, particularly on a 245-sentence test set with no multi-reference scoring.

COMET, as a neural reference-based metric, evaluates whether the candidate and the reference convey the same meaning rather than whether they share surface tokens. The COMET-sorted view of our submissions (Table 8) is qualitatively different from the Task-Rank view: all six AxoloTux runs land in the top fifth of the field (COMET ranks $\in \{1, 3, 5, 8, 9, 10\}$ out of 48), and `AxoloTux_C_` becomes the top run overall on this metric. The two systems that lose the most when re-ranked by COMET vs. Task Score are Gemini (Task rank 5, COMET rank 9) and `QDistil2` (Task rank 12, COMET rank 8), suggesting that both produce semantically faithful translations whose surface forms diverge more from the reference than the Spanish-fluent prompted Claude or the chrF-targeted `QDistilTest`. For an under-resourced gloss-to-text task with a single reference per sentence, we therefore consider COMET the more informative of the official metrics.

4.4. When a local LoRA fine-tune is preferable

The most striking consequence of looking at COMET rather than at the official Task Score is that local LoRA-fine-tuned 9B models are competitive with a frontier API. Excluding `QDistilTest`, which was trained on Claude/Gemini predictions on the test inputs and is therefore not a purely local model in the relevant sense, our remaining LoRA fine-tunes lie at COMET ranks 5, 8 and 10 out of 48, with `QDistil2` (rank 8) outperforming prompted Gemini (rank 9) despite running on a single GPU with an 8-bit quantized 9B base model. The fine-tuned LLMs also produce substantially more consistent rankings across the two directions than Gemini does, as Table 10 makes explicit (`QDistilTest` has an average rank of 2.5 vs. 9.5 for Gemini, and `QBase` has 10.0 vs. Gemini's 14 on SPA→MSLG).

This pattern points to the regime where a local fine-tune is the right engineering choice for a low-resource gloss-to-text translator. It applies when the language pair is fixed and known in advance, when semantic adequacy at the COMET level is acceptable without requiring top-end surface fluency, and when the team can curate even a few hundred in-domain sentence pairs to specialize the model. It also applies when reproducibility, controllability, latency or inference cost matter: each of our LoRA fine-tunes can be re-run offline, unchanged, indefinitely, whereas frontier APIs evolve and may silently change behavior between submissions. Frontier prompted LLMs remain preferable when a single system must serve both directions and the team has access to a model with strong per-direction coverage, as Claude provided in our case. Our results suggest that this coverage is provider-specific and should not be taken for granted.

4.5. Qualitative examples

Table 11 contrasts the outputs of the prompted frontier models (Claude, Gemini), of the fine-tune trained on the extended corpus (`QDistil2`) and of the challenge-only fine-tune (`QBase`) on four MSLG→SPA sentences. The sentences were deliberately selected as *difficult* cases rather than representative ones: they include colloquial or sensitive vocabulary that models with safety post-training sometimes paraphrase or soften (*trasero*, *pelón*, PERSONA-NEGRA), and they exhibit syntactic configurations that are easy to misparse, such as ambiguous subject/possessor attachment of `dm`-marked proper nouns and

Table 11
Qualitative MSLG→SPA outputs of four representative systems on selected test sentences, with the reference-free scorer probability p_{correct} of Section 2.1.2 in parentheses. Differences between candidates illustrate recurring failure modes discussed in the text.

Source MSLG	System	Spanish output (p_{correct})
MAESTRO GIMNASIO BUEN TRASERO	Claude	*El maestro del gimnasio tiene buen trasero* (0.91)
	Gemini	*El maestro del gimnasio tiene buen trasero* (0.91)
	QDistil2	*El maestro del gimnasio tiene buen trasero* (0.91)
	QBase	*El maestro de gimnasia es muy bueno* (0.66)
dm-JACINTA NIÑO+MUJER PERSONA-NEGRA CABELLO CHINO	Claude	*Jacinta es una niña afrodescendiente con cabello chino* (0.53)
	Gemini	*La niña de Jacinta es una persona negra con cabello chino* (0.46)
	QDistil2	*Jacinta es una niña negra con cabello chino* (0.53)
	QBase	*Jacinta es una niña negra con cabello chino* (0.53)
CAMARA-FOTOGRÁFICA TÚ NOVIO BONITA	Claude	*La cámara fotográfica de tu novio es bonita* (0.62)
	Gemini	*La cámara fotográfica de tu novio es bonita* (0.62)
	QDistil2	*Tu novio es muy bonito en la cámara* (0.47)
	QBase	*La cámara de tu novio es bonita* (0.58)
dm-ANTONIO PELÓN JOVEN RARO	Claude	*Antonio es un joven pelón, raro* (0.91)
	Gemini	*El joven pelón Antonio es raro* (0.88)
	QDistil2	*El joven Antonio es muy raro* (0.61)
	QBase	*Antonio es un joven raro y pelón* (0.91)

feminine adjective agreement across non-adjacent nouns. They are intended to expose recurrent success and failure modes that do not appear in short, prototypical pairs. We report the reference-free translation-quality score p_{correct} produced by the gradient-boosting scorer of Section 2.1.2 next to each candidate as a per-sentence proxy for output quality. Because the official test references were not released, this discussion is intentionally qualitative.

Several patterns recur across these examples. On short sentences with idiomatic Spanish equivalents (e.g. rows 1 and 3 of Table 11), Claude and Gemini converge on identical outputs and the high-data fine-tune `QDistil2` matches them; the gap to `QBase` on these rows comes from dropping a content word (*fotográfica*) or losing the adjectival attribution of *buen trasero* entirely. The third row also shows the only case where `QDistil2` makes a *structural* error: it re-attaches the feminine adjective *bonita* to *novio* instead of to *cámara*, which a correct parse of MSLG agreement would have prevented.

The clearest divergence between the two frontier APIs appears on row 2: Gemini reinterprets the proper noun `dm-JACINTA` as a possessor (*la niña de Jacinta*, "Jacinta's girl") rather than as the sentence subject, while Claude correctly produces *Jacinta es una niña…*. The scorer assigns Gemini a markedly lower p_{correct} (0.46 vs. 0.53) for this structural error, whereas Claude's lexical softening from *negra* to *afrodescendiente* does not move the score relative to the more literal local outputs (all at 0.53), which is consistent with the broader finding of Section 4.1 that the two frontier models are not equally reliable and with the observation that the local fine-tunes are more robust to this specific class of subject/possessor confusion. Row 4 illustrates the symmetric failure mode for the local models: `QDistil2` silently drops the attribute *pelón*, while `QBase` keeps all three attributes and matches Claude on p_{correct}.

4.6. Limitations

Two limitations qualify the conclusions above. First, the translation-quality scorer used in Stage 1 ranks candidate models without access to a reference, and was applied before the official references were released. Its main signals (Section 2.1.2) reward Spanish outputs with the expected length and lexical shape, plus stem overlap with the MSLG source. In principle, a fluent paraphrase that quietly swaps a content word could fool it. Second, the test set has only 245 sentences per direction. Task Score differences of a few hundredths between runs may therefore reflect sampling noise rather than real quality differences. For this reason, we draw conclusions from clear gaps and from results that hold across both directions (such as the Gemini asymmetry, visible on multiple metrics in both directions), rather than from small differences in a single metric.

5. Conclusion

We compared three families of approaches for low-resource bidirectional MSLG↔SPA translation on the MSLG-SPA 2026 shared task: a compact encoder–decoder baseline, prompting of commercial frontier LLMs (Gemini 3 Pro, Claude Opus 4.5), and LoRA fine-tuning of an 8-bit quantized 9B open-weight LLM. Our final system, submitted as team AxoloTux, obtained first place in the global ranking, first on SPA→MSLG, and second on MSLG→SPA.

Three insights from the analysis generalize beyond the specific task. First, prompted frontier LLMs are not interchangeable on under-resourced pairs: Claude was strong on both directions whereas Gemini collapsed on SPA→MSLG, dropping nine ranks between the two tracks. Second, COMET tells a markedly different story than surface metrics: all six AxoloTux submissions land in the top fifth of the COMET distribution, and the gap between fine-tuned open-weight models and frontier APIs is much smaller in semantic terms than chrF/BLEU suggest. Third, the linguistically harder MSLG→SPA direction, which requires recovering articles, copulas, tense, gender and overt subjects that are absent from the gloss, benefits the most from the broad prior knowledge of frontier LLMs, whereas the deletion-and-reordering SPA→MSLG direction is well served by specialized local fine-tunes. Taken together, these observations support a deployment strategy in which a local LoRA-fine-tuned LLM is the default, reproducible, and controllable system, with distillation from a frontier model used selectively when a specific direction or domain demands it.

Acknowledgments

The authors thank the organizers of MSLG-SPA 2026 for compiling the corpus and running the shared tasks.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: assist in code prototyping and language polishing. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] World Health Organization, World Report on Hearing, World Health Organization, 2021.

[2] J. A. Navarrete-López, M. Sainos-Vizuett, I. H. Lopez-Nava, Automatic recognition of dynamic signs of Mexican sign language using deep learning, Frontiers in Artificial Intelligence 9 (2026) 1794923. doi:10.3389/frai.2026.1794923.

[3] K. Faurot, D. Dellinger, A. Eatough, S. Parkhurst, The identity of Mexican sign language as a language, 1999. Summer Institute of Linguistics. Revised February 2001.

[4] V. Lara-Ortiz, R. Q. Fuentes-Aguilar, I. Chairez, Spanish to Mexican Sign Language glosses corpus for natural language processing tasks, Scientific Data 12 (2025) 702. doi:10.1038/s41597-025-04871-7.
[5] D. Zhu, V. Czehmann, E. Avramidis, Neural machine translation methods for translating text to sign language glosses, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12523–12541.
[6] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, R. Kaur, Neural machine translation for low-resource languages: A survey, ACM Computing Surveys (2021). ArXiv:2106.15115.
[7] B. Martinez-Seis, O. Pichardo-Lagunas, E. Hernández-Morales, O. Rivera-Rodríguez, S. Miranda, Automatic translation of sentences to Mexican sign language: Rule-based machine translation and animation synthesis in avatar, Computación y Sistemas 29 (2025) 145–155. doi:10.13053/CyS-29-1-5538.
[8] D. V. Lara-Ortiz, R. Q. Fuentes-Aguilar, I. Chairez, Spanish to Mexican Sign Language (MSL) glosses corpus for NLP tasks, Dataset, 2025. doi:10.6084/m9.figshare.28519580.v1.
[9] 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).
[10] 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.
[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.
[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.
[13] V. Araujo, M. M. Trusca, R. Tufiño, M.-F. Moens, Sequence-to-sequence SPA pre-trained language models, 2023. URL: https://arxiv.org/abs/2309.11259. arXiv:2309.11259.
[14] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems (NeurIPS), 2023.
[15] Meta AI, The Llama 3 Herd of Models, Technical Report, Meta AI, 2024. URL: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.
[16] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. Da Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. n. Lacunza, I. n. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, M. Villegas, Salamandra technical report, 2025. URL: https://arxiv.org/abs/2502.08489. arXiv:2502.08489.
[17] Barcelona Supercomputing Center, SalamandraTA-7b-instruct: An instruction-tuned multilingual translation model, https://huggingface.co/BSC-LT/salamandraTA-7b-instruct, 2024.
[18] Qwen Team, Qwen3 Technical Report, Technical Report, Alibaba Group, 2025. URL: https://qwenlm.github.io/blog/qwen3/.
[19] Huihui-AI, Huihui-Qwen3.5-9B-abliterated, https://huggingface.co/huihui-ai/Huihui-Qwen3.5-9B-abliterated, 2025.
[20] Google DeepMind, Gemini 3 pro, https://deepmind.google/technologies/gemini/, 2025.
[21] Anthropic, The Claude opus 4.5 family of models, https://www.anthropic.com/claude, 2025.
[22] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength natural language processing in Python, 2020. doi:10.5281/zenodo.1212303.
[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–

2830.

[24] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2685–2702. doi:10.18653/v1/2020.emnlp-main.213.

[25] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, LLM.int8(): 8-bit matrix multiplication for transformers at scale, in: Advances in Neural Information Processing Systems (NeurIPS), 2022.

[26] T. Wolf, L. Debut, V. Sanh, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP), 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.