RETUYT-InCo at MSLG-SPA 2026: Improving NLLB Fine-tuning with Synthetic Data for Mexican Sign Language Glosses

Luis Chiruzzo[1,*]

[1]Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay

Abstract

We present the participation of RETUYT-InCo to the MSLG-SPA 2026 shared task on machine translation for the Mexican Sign Language - Spanish translation pair. Our submitted systems are created by fine-tuning the NLLB model to the new language pair MSLG-Spanish, using different combinations of real training and synthetic training data generated by two LLMs. We describe the way the data was created, the different experiments, and the results obtained over our own development split as well as the official test set of the shared task.

Keywords

Mexican Sign Language, sign language processing, machine translation, synthetic data

1. Introduction

This paper presents the participation of the team RETUYT-InCo to the MSLG-SPA 2026 [1] shared task, part of the IberLEF 2026 [2] workshop. Sign languages are the main way of communication for the deaf and hard-of-hearing communities. They are complete natural languages that are produced mainly in the visual-spatial modality instead of the oral-auditory modality of spoken languages [3, 4]. Each country has their own sign language (on occasions more than one per country), and even countries that share the same main oral languages (such as Mexico and Spain) could have different non mutually intelligible sign languages.

The Mexican Sign Language (MSLG, or LSM by its acronym in Spanish "Lengua de Señas Mexicana") is the main language used by the deaf and hard-of-hearing community in Mexico, with an estimated number of between 49,000 and 195,000 signers. As well as most of the sign languages in the world, MSLG has not yet been fully explored from the NLP perspective. Sign languages in general suffer from lack of digital resources that could be used to train models and develop NLP tools. All sign languages in the world are considered low-resource languages [5, 6], and MSLG is no exception, with this shared task being the first one to provide a high quality dataset of MSLG for experimentation.

In this work, we will complement the small but high quality data, provided by the organizers of the shared task, with a larger but considerably lower quality synthetic dataset, and try to leverage both corpora in order to improve the translation results.

The rest of the paper is organized as follows: Section 2 presents both the real and synthetic datasets we used. Section 3 describes the experiments we did with real and different combinations of synthetic data. Section 4 shows the results we obtained over the dev and test data, and finally section 5 provides some conclusions and future work.

2. Data

This section describes the dataset we worked with, including our partition of the original MSLG-SPA dataset of the shared task and our synthetic sets.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

luischir@fing.edu.uy (L. Chiruzzo)

0000-0002-1697-4614 (L. Chiruzzo)

	Pairs	MSLG glosses	Spanish words
Training	440	2052	2654
Dev	50	226	304
Subtotal (MSLG Training Set)	490	2278	2958
Synthetic Gemini	6717	25186	33245
Synthetic GPT	3760	12291	15898
Subtotal Synthetic	10477	37477	49143
MSLG Test Set*	245	1156	1585

Table 1
Statistics of the corpus we used for our experiments. The training and dev sets are our subdivision of the original MSLG Training Set. The synthetic corpus was used for complementing the data in our experiments. *The MSLG Test Set is shown for completeness, but is not in fact a parallel corpus but two sets of sentences for each language.

2.1. Corpus

The MSLG-SPA corpus provided by the organizers of the shared task contains 490 pairs of <MSLG gloss sequence, Spanish sentence>. We split this original set into a training set and a development set, that we used to train and compare all our experiments. We took 50 random pairs from the original set to create the dev set, and used the rest for training. In all our experiments, we made all the MSLG glosses lowercase, and then postprocessed the results after generation. The first rows of Table 1 show the composition of our training and dev sets.

Later on the organizers of the shared task released a test set for submissions, comprising 245 new MSLG gloss sequences and 245 Spanish sentences. These are shown in the last row of the table, but please note that in this case the number of glosses and words do not correspond to the same sequences. These are shown in the table only for completeness.

2.2. Synthetic data

In our experiments we used two different datasets to complement the training data. One of them was created using Google Gemini [7] (model `gemini-3.1-pro-preview`). We randomized our training set and split it in chunks of 20 sample pairs. Then we used the Gemini API to obtain more examples with the following prompt:

```
Here are 20 examples of translations from Spanish to LSM:
Perdí el suéter blanco. -> suéter blanco yo perder
Mi primo vive en Orizaba. -> mi primo orizaba vivir
...
Now write 200 more examples in the format: Spanish -> LSM
```

We traversed the whole training set twice obtaining new examples from Gemini, getting a total of 8800 new pairs, but after removing duplicates that number went down to 6717. The high number of duplicates might be explained in part because we are sampling with a temperature value of 0, top p of 0.95, and top k of 20.

After generating the samples with Gemini, we used the OpenAI ChatGPT [8] online interface in a similar way to generate more examples. As we were using the online interface, the underlying GPT version could be switching between 5.5 and 4.5. We also could only generate 100 samples at a time, instead of 200, so we created in total 3760 GPT pairs after deduplication.

The middle rows in Table 1 show the statistics of our synthetic sets. One thing that can be seen is that the word to gloss ratio is perfectly respected by both models, being around 1.3 Spanish words per MSLG gloss.

Table 2 shows some examples of synthetic pairs of text generated by both models. We can see from the examples that both models are able to pick up the basic conventions for representing glosses in MSLG

Model	MSLG	Spanish
Gemini	clase hora dos terminar	La clase termina a las dos.
	centro histórico tú conocer	¿Conoces el centro histórico?
	enfermera inyectar paciente	La enfermera inyecta al paciente.
	enfermero+mujer poner inyección	La enfermera pone inyecciones.
	dm-pedro su papá doctor	El papá de Pedro es doctor.
	desde niña ella bailar dm-ballet	Ella baila ballet desde niña.
	yo dm-whatsapp mensaje enviar tú	Te envié un mensaje por WhatsApp.
	dm-maría su tío+mujer enfermo	La tía de María está enferma.
GPT	mi padre mucho trabajar	Mi padre trabaja mucho.
	yo no poder dormir	No puedo dormir.
	nosotros vivir ciudad	Nosotros vivimos en la ciudad.
	flor muy bonita	La flor es muy bonita.
	dm-juan panadería pan comprar	Juan compró pan en la panadería.
	dm-maría estudiar historia	María estudia historia.
	niño+ plural jugar calle	Los niños juegan en la calle.
	ellos comprar regalo+	Ellos compran regalos.

Table 2
Examples of synthetic MSLG-Spanish pairs generated by Gemini and GPT.

as simple Spanish lemmas without morphology. Both models also seem to have learned to represent proper names as fingerspelling with the "dm-" prefix, but there are discrepancies on the frequency of use of these constructions: Gemini tends to produce sentences with some proper name about 2.7% of the time, varying between names of people, geographic places, and on occasions other things that could be errors like the example "dm-ballet" in the table. On the other hand, GPT uses it very conservatively only around 1% of the time, and almost exclusively on the proper nouns "Juan", "María", and "Pedro".

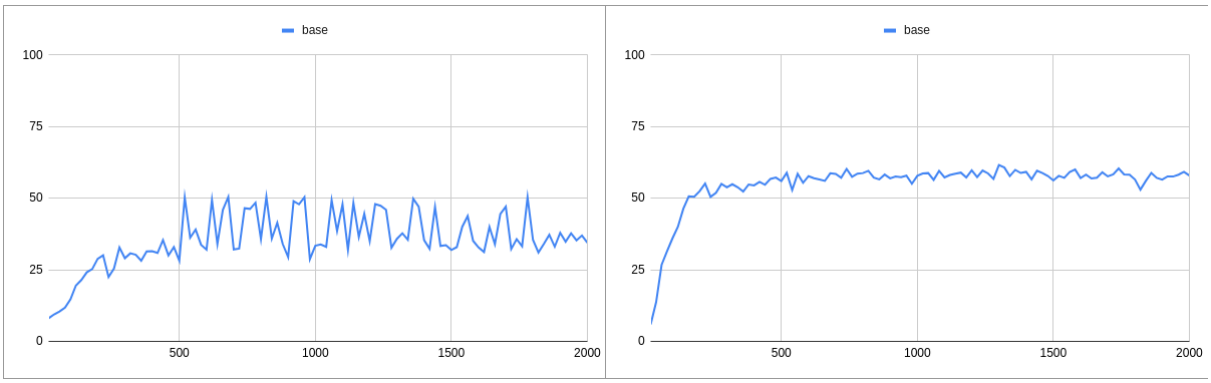
Furthermore, Gemini is able to produce some examples of "+" constructions, used for inflectional morphology in MSLG (e.g. "tío+mujer" to refer to "tía"). Even in some cases it could create new elements not found in the training corpus, like "enfermero+mujer" to mean "enfermera", as shown in the table. But it does so inconsistently because sometimes it also uses "enfermera" as a gloss, which must be a mistake. GPT seems not to understand the proper use of "+" leaving a blank space after the token, and also in many cases using it at the end of sentences, like the example "ellos comprar regalo+", which should be invalid.

3. Experiments

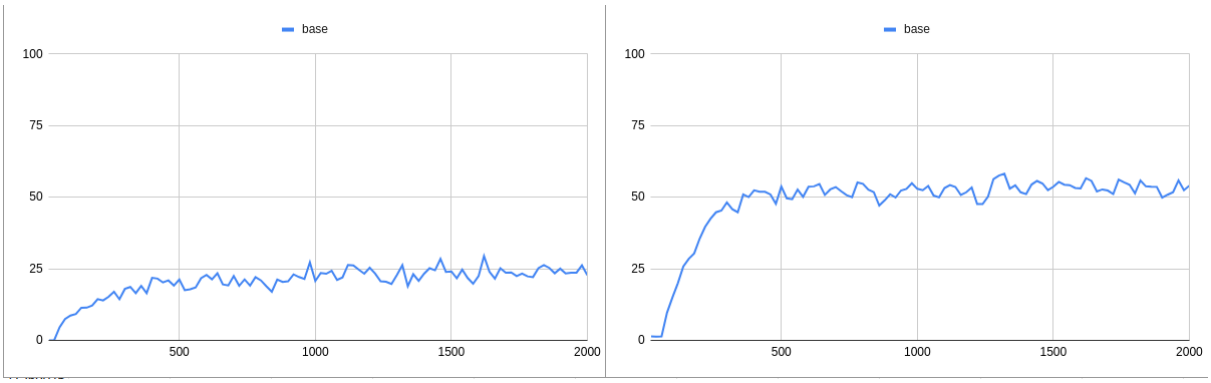
All our experiments are different ways of fine-tuning the NLLB model [9] adding the new language pair of MSLG-Spanish. NLLB [9] is a pretrained encoder-decoder transformer [10] model that has already been trained on a large dataset containing all language pairs for 200 languages, many of them low-resource, and can be fine-tuned to new language pairs as in our case. We first did a baseline experiment of fine-tuning the model with only the gold training data, and then used combinations of the Gemini and GPT synthetic data in two ways for pretraining. Our experimental setting and training strategy have already been tried in the Uruguayan Sign Language (LSU) - Spanish language pair, obtaining good results [11], so we wanted to see if this strategy also worked in this new language pair. All our experiments were run using a NVIDIA Tesla T4 (16 GB of VRAM) GPU in the Google Colab environment.

3.1. Base experiment

Our first baseline experiment is just adding the new language MSLG into the mix of languages of NLLB. The fine-tuning strategy is a continued training of the original NLLB model, but this time using only



(a) BLEU (left) and CHRF (right) on the MSLG->Spanish direction.



(b) BLEU (left) and CHRF (right) on the Spanish->MSLG direction.

Figure 1: Performance of the baseline experiment, fine-tuning the NLLB corpus with the gold training set, over the dev set.

data of MSLG and Spanish. We used a batch size of 8, maximum sentence length of 128 tokens, and learning rate of 1e-4. The training was done for 2000 steps, taking a snapshot every 20 steps to evaluate the performance over the dev set. At every step, a batch was randomly sampled from the training set, and a direction (either MSLG->Spanish or Spanish->MSLG) was randomly selected with 50% chance. We took both BLEU and CHRF metrics for all steps, but in the end we only considered the one with the best BLEU as the candidate to use as one of our submissions. In order to avoid issues with larger order n-grams that brought the BLEU values to 0, and to make the comparisons easier, we calculated BLEU with the Chen and Cherry smoothing method 7 [12].

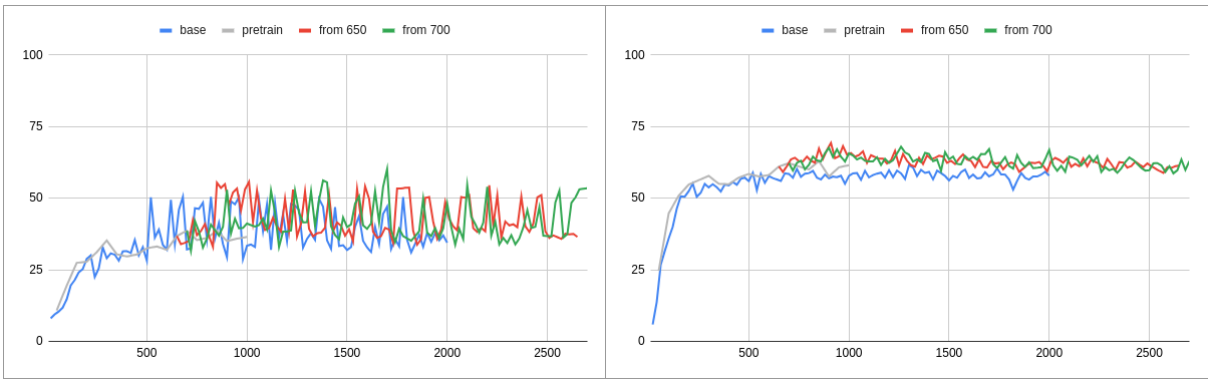
Fig. 1 shows the BLEU and CHRF performance over the dev set during the 2000 training steps. The peak performance according to the BLEU metric was achieved at step 820 for MSLG->Spanish, and at step 1620 for Spanish->MSLG.

3.2. Staged data experiments

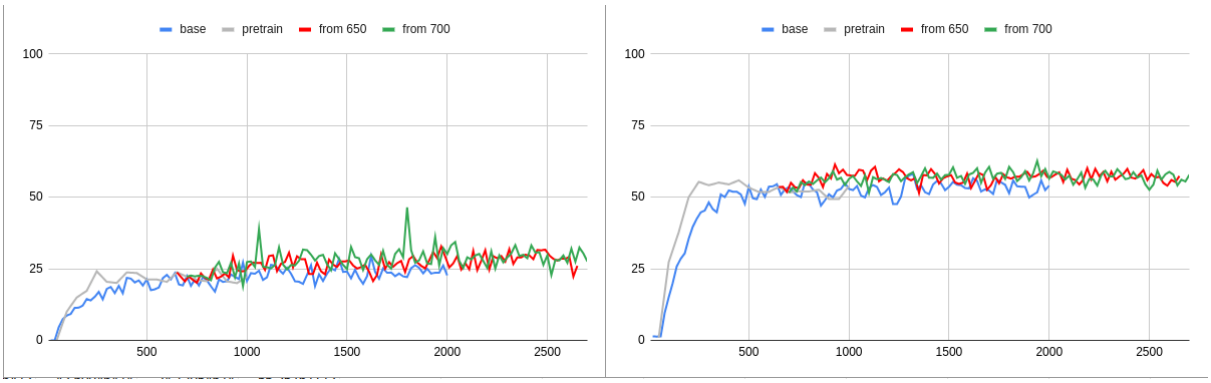
The second set of experiments was inspired in the strategy followed by [13]. In that work the training is done in two stages: first a pretraining phase in which they use only synthetic data (which is more abundant), then they choose a few points that look like local maxima during the pretraining phase, and starting from those points they do a fine-tuning phase using only gold training data (which is a considerably smaller dataset).

We followed the same strategy in two experiments. First we did the pretraining phase only with the Gemini synthetic data, and in a second round of experiments with the mix of Gemini and GPT data.

In the first experiment, the pretraining phase ran for 1000 steps using only synthetic data created with Gemini, taking snapshots every 50 steps. We then chose the following pretraining points, which were local maxima during pretraining in either direction: 150, 250, 300, 450, 500, 650, 700, and 850. We



(a) BLEU (left) and CHRF (right) on the MSLG->Spanish direction.



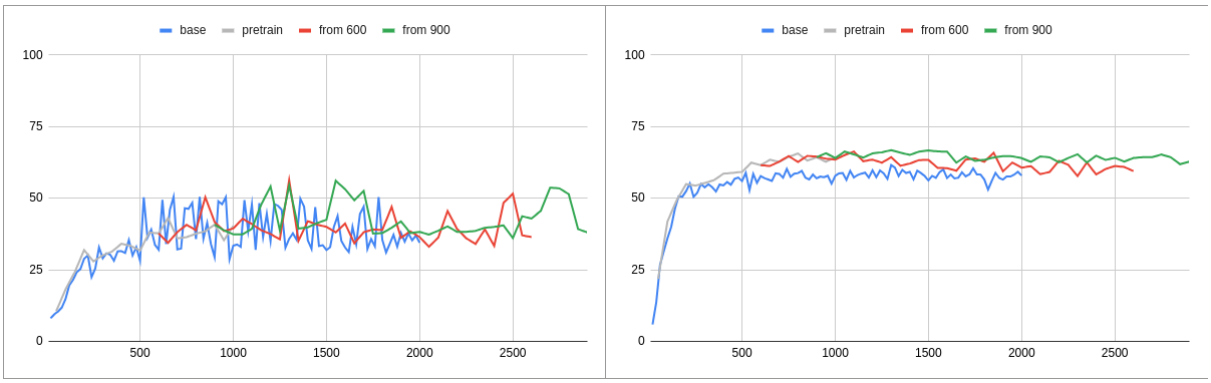
(b) BLEU (left) and CHRF (right) on the Spanish->MSLG direction.

Figure 2: Performance of the staged fine-tuning experiment using Gemini synthetic data over the dev set. The grey line is the pretraining phase using only synthetic data, and then fine-tuning with gold training data from steps 650 (red) and 700 (green) are shown.

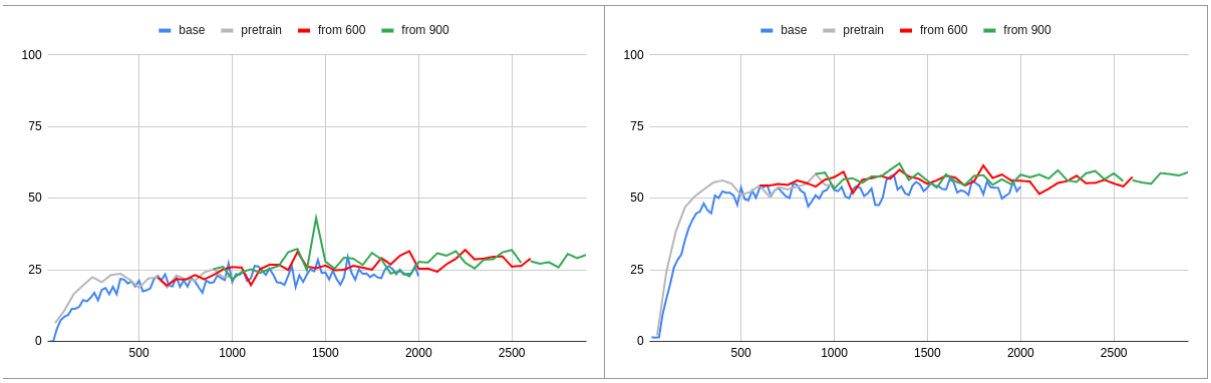
then did eight fine-tuning rounds, starting from each of those points. Each fine-tuning ran for 2000 more steps, taking a snapshot every 20 steps.

Fig 2 shows the performance over the dev set for this experiment. The blue line is the base experiment shown for comparison, and the grey line is the result of the pretraining phase. In the charts we only show two of the fine-tuning rounds (from 650 in red, and from 700 in green), so as not to clutter the chart too much. What can be seen clearly in the charts is that the fine-tuning rounds achieve higher CHRF scores than just using the gold training data for most of the training steps, and also the peaks of the BLEU score are a little higher. Using the maximum BLEU scores, we chose for test submissions the model from pretraining step 700 fine-tuned at step 1000 for the MSLG->Spanish direction, and the model from pretraining 700 fine-tuned at step 1100 for the Spanish->MSLG direction.

In the second experiment, we repeated the same strategy but using the union of Gemini and GPT synthetic data. We first pretrained for 1000 steps, and selected the pretraining steps 200, 400, 550, 600, 650, 700, 800, and 900. Then we did the fine-tuning rounds from each pretraining point for 2000 more steps. Fig. 3 shows the results of this second staged experiment. Again we only show two fine-tuning rounds (600 in red and 900 in green) to make the chart more clear. The same results in general as in the previous experiment can be seen, but the maximum BLEU and CHRF obtained in this experiment were a few points below the one using Gemini data only. Using the maximum BLEU scores, we chose for test submissions the model from pretraining step 600 fine-tuned at step 700 for the MSLG->Spanish direction, and the model from pretraining 900 fine-tuned at step 550 for the Spanish->MSLG direction.



(a) BLEU (left) and CHRF (right) on the MSLG->Spanish direction.



(b) BLEU (left) and CHRF (right) on the Spanish->MSLG direction.

Figure 3: Performance of the staged fine-tuning experiment using Gemini and GPT synthetic data over the dev set. The grey line is the pretraining phase using only synthetic data, and then fine-tuning with gold training data from steps 600 (red) and 900 (green) are shown.

3.3. Combined data experiments

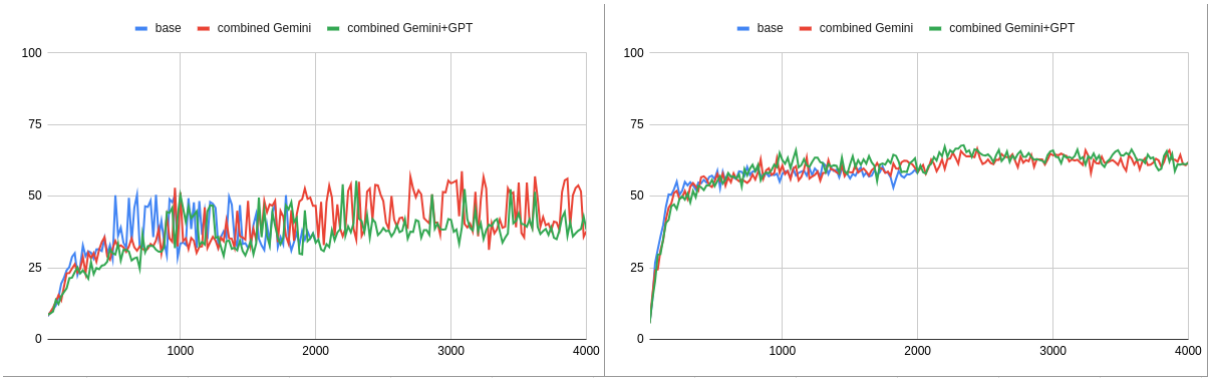
We did two final rounds of experiments but using the synthetic data in a different way. Instead of doing first a pretraining phase with synthetic data only, we ran the pretraining with the union of the synthetic and gold training data. This pretraining phase was executed for 2000 steps, and then was followed by 2000 more steps using only gold training data, totaling 4000 steps. We did this experiment using only Gemini synthetic data, or using Gemini and GPT synthetic data.

Fig. 4 shows the performance across the 4000 steps for the two experiments (red and green) and compared to the base experiment (blue). It is interesting that the point in which the pretraining phase with syntetic data stops and the fine-tuning with gold data starts can be clearly seen in the charts, and the peformance (both BLEU and CHRF) suddenly starts improving, even if it had been almost the same as in the blue line up to that point. However, when comparing the top performance over the dev set, these experiments did not improve over the staged experiment pretraining with Gemini data only. Based on the maximum BLEU scores, we chose for test submissions the model at step 3080 pretrained with Gemini data in the MSLG-Spanish direction, and the model at step 3500 in the other direction. Likewise, we chose the model at step 2300 pretrained with Gemini and GPT data for the MSLG->Spanish direction, and the model at step 2720 in the other direction.

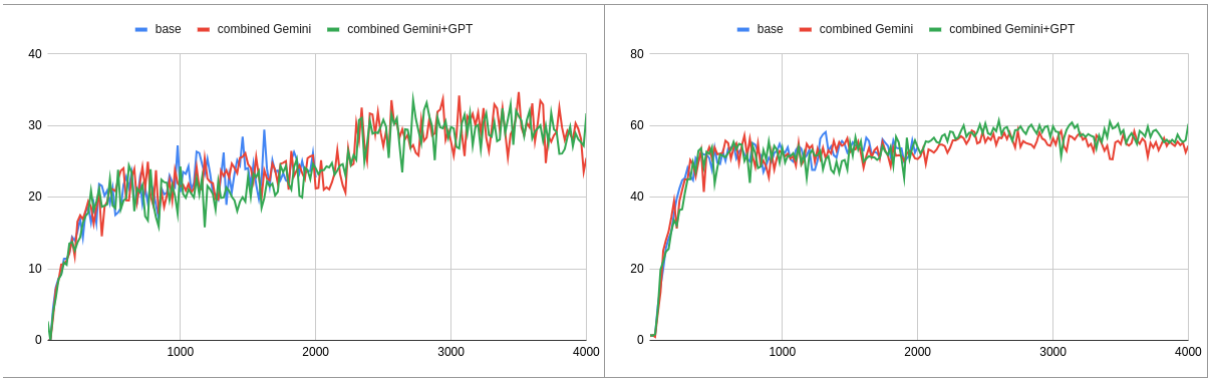
3.4. Postprocessing

The text produced by our models was slightly postprocessed to fix some formatting issues. The postprocessing steps are:

- Balance question (‘¿’ and ‘?’) and exclamation (‘¡’ and ‘!’) marks.



(a) BLEU (left) and CHRF (right) on the MSLG->Spanish direction.



(b) BLEU (left) and CHRF (right) on the Spanish->MSLG direction.

Figure 4: Performance of the fine-tuning experiment using combined synthetic and gold data over the dev set. The red line shows the performance using Gemini synthetic data only, while the green line uses Gemini and GPT synthetic data.

- Use all uppercase for MSLG sequences, except for the "dm-" prefixes (where present).
- Remove "dm-" prefixes in Spanish sentences (if any).

4. Results

Table 3 shows a summary of performance of the best model from each of our experiments on the dev set. As mentioned before, we used the BLEU metric to choose the best model in each case, and submitted those results to the shared task.

The right part of the table shows the results obtained over the test set of the shared task. The first thing that is noticeable from the table is that there is a huge performance gap in terms of the BLEU metric between our dev set and the test set. For the MSLG->Spanish direction the test BLEU performance is about half of the one we calculated in dev, and for the other direction is about a third. The CHRF metric also obtained worse results, but in this case only losing about 8 points in MSLG->Spanish and 4 points in Spanish->MSLG. These big differences in performance could mean that the test set could be significantly different in style or composition from the training set, and in particular from the small subset we used as dev. Another possible explanation is that our choice of using BLEU as the main metric to trust, and the fact that we did many rounds of experiments, could have made that our best models are too overfit to our dev data. Since our dev set is so small and completely from the same distribution of the training data, it could happen that our models could easily reproduce its content but they could not generalize to new data. Also, our use of Chen and Cherry smoothing method 7 could be another source of differences between the calculated BLEU for dev and test.

The other interesting result is that the comparative performance of our different models was also not the same on dev and test. Over the dev set, the staged experiments behaved much better than the

	Dev			Test					
MSLG->Spanish	Step	BLEU	CHRF	BLEU	Rank	CHRF	Rank	Task Score	Rank
base	820	50.48	59.55	23.08	28/48	52.69	28/48	-0.1048	28/48
staged Gemini	700+1000	**60.13**	**67.13**	25.38	23/48	56.41	24/48	0.0666	26/48
staged Gemini+GPT	600+700	56.29	64.30	24.79	25/48	56.00	25/48	0.0730	25/48
combined Gemini	3080	58.68	64.35	27.07	22/48	55.99	26/48	0.1169	22/48
combined Gemini+GPT	2300	55.41	66.58	**28.51**	20/48	**58.70**	21/48	**0.2043**	21/48
Spanish->MSLG	Step	BLEU	CHRF	BLEU	Rank	CHRF	Rank	Task Score	Rank
base	1620	29.44	56.64	13.12	23/52	52.16	33/52	0.0833	28/52
staged Gemini	700+1100	**46.43**	**60.60**	15.19	18/52	**56.88**	20/52	**0.5705**	15/52
staged Gemini+GPT	900+550	43.10	58.73	**17.77**	12/52	56.71	21/52	0.3214	19/52
combined Gemini	3500	34.68	54.82	16.67	15/52	53.64	30/52	0.4123	18/52
combined Gemini+GPT	2720	33.55	58.73	16.07	17/52	55.85	26/52	0.4247	16/52

Table 3
Results of our experiments over the development and test sets.

combined ones. This is true in test for the Spanish->MSLG direction, but for the other direction the combined experiments were better. However, we can see that the experiments using synthetic data (either staged of combined) were always better than the baseline with only gold data, both in dev and test, which confirms our intuition that it's possible to improve the translation results by using synthetic data created in this way.

5. Conclusions

We presented the participation of the team RETUYT-InCo to the MSLG-SPA 2026 shared task. Our team tried different ways of improving the fine-tuning of the NLLB model with synthetic data, obtaining mixed-to-positive results. The best results in the MSLG->Spanish direction were obtained by pretraining with a combination of synthetic data generated by Gemini and GPT, and the gold training data, followed by a fine-tuning phase using only gold data. In the Spanish->MSLG direction, the best model was found by first pretraining with only synthetic Gemini data for a few epochs, and then fine-tuning with the gold training data. However, in all cases our results over the dev set were much higher than the ones obtained over the test set, which might indicate that the over-reliance on our dev data might not have translated well into the different data distribution of the test set.

In the future we would like to analyze more the results on the test set to understand how the differences in distribution or domain might be affecting the results of our models. Furthermore, we would like to analyze more the rules and behavior of MSLG from a theoretical point of view in order to define better ways of translation and generation of synthetic data. We would also like to use more modern and larger models that might yield even better results, like the newer generation Omnilingual MT models [14], or directly fine-tuning the generative auto-regressive models like Gemini or Llama.

Acknowledgments

This research is with support from Google.org and the Google Cloud Research Credits program for the Gemini Academic Program.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools during the writing of this document, but the experiments described here are done with the aforementioned AI tools: Google Gemini, Meta NLLB, OpenAI GPT.

References

[1] 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).

[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[3] K. Yin, A. Moryossef, J. Hochgesang, Y. Goldberg, M. Alikhani, Including signed languages in natural language processing, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7347–7360. URL: https://aclanthology.org/2021.acl-long.570/. doi:10.18653/v1/2021.acl-long.570.

[4] A. E. Baker, Sign languages as natural languages, in: The linguistics of sign languages: An introduction, John Benjamins Publishing Company, 2016, pp. 1–24.

[5] A. Moryossef, K. Yin, G. Neubig, Y. Goldberg, Data augmentation for sign language gloss translation, in: Proceedings of the 1st international workshop on automatic translation for signed and spoken languages (AT4SSL), 2021, pp. 1–11.

[6] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the nlp world, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 6282–6293.

[7] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[9] NLLB Team, Scaling neural machine translation to 200 languages, Nature 630 (2024) 841–846.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[11] L. Chiruzzo, B. Pérez Rodríguez, J. Bianchi-Madera, M. Dutra, R. Aguirre, Machine Translation between Uruguayan Sign Language (LSU) Glosses and Spanish, Procesamiento del Lenguaje Natural 77 (2026).

[12] B. Chen, C. Cherry, A systematic comparison of smoothing techniques for sentence-level BLEU, in: O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, L. Specia (Eds.), Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 362–367. URL: https://aclanthology.org/W14-3346/. doi:10.3115/v1/W14-3346.

[13] L. Chiruzzo, E. McGill, S. Egea-Gómez, H. Saggion, Translating spanish into spanish sign language: Combining rules and data-driven approaches, in: Proceedings of the fifth workshop on technologies for machine translation of low-resource languages (LoResMT 2022), 2022, pp. 75–83.

[14] B. Alastruey, N. Bafna, A. Caciolai, K. Heffernan, A. Kozhevnikov, C. Ropers, E. Sánchez, C.-E. Saint-James, I. Tsiamas, C. Cheng, et al., Omnilingual mt: Machine translation for 1,600 languages, arXiv preprint arXiv:2603.16309 (2026).