Instruction-Tuned Large Language Models for Mexican Sign Language Gloss Translation with Linguistic and Synthetic Data Augmentation

Obdulia Pichardo-Lagunas[1,*,†], Sabino Miranda[2,*,†], Bella Martinez-Seis[1,†] and Carlos Gómez-García [3,†]

[1]*UPIITA-IPN, Instituto Politécnico Nacional, Ciudad de México, Ciudad de México, México*
[2]*Universidad Autónoma de la Ciudad de México, Ciudad de México, México*
[3]*Khasm Labs, Seattle, USA*

Abstract

Mexican Sign Language (MSL) translation remains a low-resource challenge due to the scarcity of parallel corpora. In this work, we explore instruction-tuned large language models (LLMs) for the tasks of Bidirectional Translation between Mexican Sign Language Glosses and Spanish (MSLG-SPA 2026) at IberLEF 2026. We fine-tuned Qwen2.5 instruction models with 7B and 32B parameters using parameter-efficient fine-tuning and several data augmentation strategies, and the incorporation of linguistic information derived from part-of-speech annotations. Experimental results show that the MSLG2SPA direction substantially outperformed SPA2MSLG, achieving a best BLEU score of 64.34 with the Qwen2.5-32B-Instruct model. Our findings suggest that large instruction-tuned models are highly effective for gloss-to-text generation, while the inverse translation direction remains considerably more challenging due to the structural compression and ambiguity of gloss representations.

Keywords

Mexican Sign Language, Machine Translation, Large Language Models, Gloss Translation

1. Introduction

Sign language translation has become an important research area within natural language processing due to its potential to improve accessibility and communication for Deaf communities. However, most existing research has focused on high-resource sign languages such as American Sign Language (ASL), while Mexican Sign Language (MSL) remains comparatively underexplored. The MSLG-SPA 2026 shared task [1], organized as part of IberLEF 2026 [2], addresses bidirectional translation between Mexican Sign Language glosses (MSLG) and Spanish and poses challenges due to limited parallel corpora.

Recent advances in large language models (LLMs) have substantially improved machine translation performance, particularly in multilingual and low-resource settings through instruction tuning and parameter-efficient fine-tuning techniques [3, 4]. In prior work, researchers have also explored linguistically informed transformer architectures for sign language translation. For example, Varanasi et al. [5] proposed a framework for translating English text into American Sign Language (ASL) glosses using linguistic features and transformer-based models, achieving strong performance across general-domain and domain-specific translation tasks.

Parameter-efficient adaptation methods such as LoRA and QLoRA have enabled the fine-tuning of large language models with reduced computational requirements while maintaining competitive performance [6, 7]. Additionally, instruction-tuned multilingual models such as the Qwen family have recently demonstrated strong reasoning and generation capabilities across structurally diverse

IberLEF 2026, September 2026, León, Spain
*Corresponding authors.
†These authors contributed equally.
opichardola@ipn.mx (O. Pichardo-Lagunas); sabino.miranda@uacm.edu.mx (S. Miranda); bcmartinez@ipn.mx (B. Martinez-Seis); cgomez.egarcia@gmail.com (C. G.)
0000-0003-3577-7218 (O. Pichardo-Lagunas); 0000-0002-0595-2401 (S. Miranda)

languages [8]. These characteristics make them suitable candidates for low-resource sign language translation tasks.

In this work, we investigate instruction-tuned Qwen2.5 models for both translation directions: Mexican Sign Language gloss to Spanish (MSLG2SPA) and Spanish to Mexican Sign Language gloss (SPA2MSLG). Our approach combines QLoRA fine-tuning with linguistic augmentation based on part-of-speech (POS) annotations extracted using spaCy [9]. We additionally propose a synthetic data augmentation pipeline based on generator–evaluator agents designed to increase linguistic diversity while preserving gloss structure consistency.

Experimental results show that the MSLG2SPA direction substantially outperformed SPA2MSLG across all evaluation metrics. The best-performing configuration, based on Qwen2.5-32B-Instruct with POS-enhanced inputs and synthetic augmentation, achieved a BLEU score of 64.34. Our findings suggest that instruction-tuned LLMs combined with controlled synthetic augmentation are highly effective for low-resource gloss-to-text translation, while the inverse direction remains considerably more challenging due to the compressed and ambiguous nature of gloss representations.

2. System description

Our approach models Mexican Sign Language translation as an instruction-following sequence generation task using large language models. We fine-tuned Qwen2.5 instruction-tuned models with QLoRA in both translation directions: from Mexican Sign Language glosses to Spanish (MSLG2SPA) and from Spanish to Mexican Sign Language glosses (SPA2MSLG). In addition to raw text, we incorporated linguistic information extracted with spaCy [9], including part-of-speech tags, to provide explicit syntactic guidance during training. We also explored a synthetic data augmentation pipeline based on generator–evaluator agents to increase structural diversity while preserving gloss consistency. The final system combines parameter-efficient fine-tuning, linguistic information, and synthetic data generation for low-resource sign language translation. The following sections provide descriptions of these components.

2.1. Data

The experiments were conducted using the official training data provided by the MSLG-SPA 2026 shared task. The corpus consists of 489 parallel pairs composed of a Spanish sentence and its corresponding Mexican Sign Language (MSLG) gloss sequence. The dataset was split into training, validation, and test sets at 70/10/20. Specifically, 70% of the data was used for model training, 10% for validation during fine-tuning, and the remaining 20% for evaluation. Synthetic data augmentation was applied exclusively to the training split in order to avoid data leakage and ensure a fair evaluation protocol. This partitioning strategy allowed us to perform controlled experiments while assessing the impact of linguistic augmentation and synthetic data generation under low-resource conditions.

2.2. Models

Our approach employed mainly the instruction-tuned Qwen2.5 models [8], specifically Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct, due to their strong multilingual generation and reasoning capabilities. Both models were fine-tuned using QLoRA [4], a parameter-efficient adaptation strategy that significantly reduces GPU memory usage while preserving competitive performance.

Our experiments explored several configurations varying in model size (7B and 32B parameters), the percentage of augmented training data, and the inclusion of linguistic annotations extracted with spaCy. To evaluate translation quality, we employed multiple automatic metrics commonly used in machine translation tasks, including BLEU [10], chrF [11], METEOR [12], and COMET [13].

2.3. Linguistic Augmentation

To enrich the input representation, we incorporated linguistic annotations extracted with spaCy [9]. In particular, we focused on integrating part-of-speech (POS) information as an additional syntactic signal for both Spanish sentences and MSL glosses. The objective was to provide the model with explicit grammatical cues to disambiguate lexical roles and improve structural generation during translation.

We experimented with two input configurations: *Base*, which used only the original text or gloss sequence, and *POS*, which augmented each token with its corresponding POS tag.

For example, the Spanish sentence:

El gato negro es mi favorito.

was transformed into the POS-augmented representation:

```
El <DET> gato <NOUN> negro <ADJ> es <AUX> mi <DET> favorito <ADJ>
```

Similarly, the corresponding MSL gloss sequence:

```
GATO COLOR NEGRO MI FAVORITO
```

was represented as:

```
GATO <NOUN> COLOR <NOUN> NEGRO <ADJ> MI <DET> FAVORITO <ADJ>
```

This augmentation strategy allows the model to access shallow grammatical structure while preserving the original lexical content. The additional syntactic information could be particularly beneficial in low-resource scenarios, where explicit linguistic signals may compensate for limited training examples.

2.4. Data Augmentation Pipeline

To address the limited size of the training corpus, we developed a synthetic data augmentation pipeline based on a multi-agent generation and evaluation framework implemented with LangGraph [14]. The objective of the pipeline was to generate new training pairs that were semantically diverse while preserving the grammatical and structural properties of the original Mexican Sign Language gloss sequences.

The augmentation process starts from a verified bilingual pair composed of a Spanish sentence and its corresponding MSL gloss representation. The system then generates new synthetic examples that maintain the same gloss structure but introduce different semantic content and vocabulary.

The pipeline consists of three interconnected stages:

1. **Spanish Sentence Generation**: A generator model produces a new Spanish sentence with grammatical structure similar to the original example while avoiding semantic repetition. Previously generated examples and evaluator feedback are included in the prompt to increase diversity and reduce duplication.
2. **MSL Gloss Generation**: Given the newly generated Spanish sentence and the original bilingual pair as a structural template, the system generates a corresponding MSL gloss sequence. The generation process attempts to preserve the same gloss ordering, grammatical markers, and structural organization of the original gloss sequence.
3. **Quality Evaluation**: A dedicated evaluator model validates each generated pair according to several criteria:
    - *Semantic diversity*: vocabulary, actors, locations, and events must be clearly distinct from the original. Trivial substitutions are rejected.
    - *Syntactic equivalence*: The gloss sequence must exactly replicate the pattern of the original (same number of units, same order, same grammatical markers).

- *MSL convention compliance*: All annotation symbols must be used correctly and semantically justified.
- *Spanish–gloss consistency*: Both descriptions must refer to the same event, participants, and action.
- *Non-repetition*: The new pair must not differ from previously approved examples in only one or two content words.

Only examples that successfully satisfied all evaluation criteria were incorporated into the augmented training corpus. When a generated example failed validation, the evaluation pipeline's feedback was used to selectively regenerate either the Spanish sentence or the MSL gloss sequence, while preserving the components that had already been validated successfully.

To prevent infinite correction cycles and maintain computational efficiency, the regeneration process was limited to a maximum of two refinement iterations per example.

The generation stage of the pipeline employed a GPT-based model (GPT-5.2) configured for low-effort reasoning in order to efficiently produce diverse synthetic examples, while the evaluation stage relied on a Qwen-based model (Qwen3 14B) using deterministic decoding to ensure consistent validation and feedback generation. This generator–evaluator framework enabled the creation of synthetic parallel data with controlled structural consistency, improved linguistic diversity, and iterative quality refinement throughout the augmentation process.

An example of the data augmentation process is shown below:

Original Spanish sentence: Mi abuela hace un mole muy rico.
Original MSLG: MI ABUELO+MUJER HACER RICO MOLE

Generated Spanish sentence: Mi tía cocina un pozole delicioso.
Generated MSLG: MI TÍO+MUJER COCINAR DELICIOSO POZOLE

In this example, the semantic content changes while preserving the overall gloss structure, grammatical markers, and token ordering.

Using this pipeline, we generated 3843 synthetic pairs. The optimal configuration (75% augmentation) improved MSLG2SPA translation performance, as shown in Section 3.

2.5. Instruction-Tuning and Fine-Tuning Setup

The translation task was formulated as an instruction-following generation problem. For both translation directions, training examples were converted into chat-style interactions using the native Qwen chat template. Each example consisted of a system instruction describing the translation task and the conventions of the Mexican Sign Language gloss notation, followed by a user message containing the source sequence and an assistant message containing the target translation.

2.5.1. Prompt Design

Since Qwen2.5 is an instruction-tuned model, translation was formulated as a supervised instruction-following task. For each translation direction, we designed a task-specific system prompt describing the conventions used in Mexican Sign Language glosses and the expected output format. The prompts explicitly defined corpus-specific annotations such as compound signs (+), lexicalized expressions (-), acronym spelling (#), and fingerspelled proper names (dm-). Additional linguistic guidance was provided regarding pronoun omission, tense markers, and contextual glosses such as LUGAR. When linguistic augmentation was enabled, the prompt also instructed the model to leverage the provided POS annotations to improve sentence structure. During both training and inference, examples were formatted using the Qwen2.5 chat template, consisting of a system instruction followed by a user message containing the source sequence. The model was trained to generate only the target translation. Table 1 illustrates an example prompt used for the MSLG2SPA task.

Table 1
Example instruction prompt used for MSLG2SPA translation with POS augmentation.

System: Translate Mexican Sign Language glosses (MSLG) into natural Spanish. Conventions: + combination of signs (HERMANO+MUJER = hermana) - lexicalized expressions (CREMA-DE-COSMÉTICO = crema facial) # spelling of acronyms or foreign words (#SEP, #TV) dm- fingerspelling of proper names (dm-LUIS) Linguistic rules: - Pronouns in MSLG are not translated when they are implicit in the verb (TÚ ENCENDER = enciende) - Temporal markers affect verb conjugation (HACE-TIEMPO YO COMER = comí) - Some glosses indicate contextual information without direct translation (LUGAR CANCÚN YO IR = fui a Cancún) Use the POS structure to correctly align the sentence. Generate only a fluent, natural, and grammatically correct Spanish sentence.
User: `Glosas: LUGAR <NOUN> CANCÚN <PROPN> YO <PRON> IR <VERB>`
Assistant: `Fui a Cancún.`

2.5.2. Experimental setup

All models were fine-tuned using the Hugging Face Transformers framework [15] together with QLoRA for parameter-efficient adaptation. During training, we used a learning rate of 1×10^{-4}, a per-device batch size of 2, and gradient accumulation over 16 steps, resulting in a batch size of 32 examples. Models were trained for 4 epochs using the Paged AdamW 32-bit optimizer. Mixed-precision training (FP16) was enabled, and the best checkpoint was selected according to the validation loss at the end of training.

For QLoRA, the base models were loaded with 4-bit quantization and double-quantization enabled. Fine-tuning employed a LoRA rank of $r = 16$, scaling factor $\alpha = 32$, and a dropout rate of 0.05. Gradient checkpointing was enabled during training to reduce memory consumption.

During inference, translation from MSLG to Spanish employed beam search with 4 beams, a maximum generation length of 64 tokens, and a length penalty of 1.1. For Spanish-to-MSLG generation, we used a maximum length of 96 tokens, a beam size of 2, a temperature of 0.3, and top-p sampling with $p = 0.9$.

The fine-tuning experiments were conducted on a server equipped with four NVIDIA L40S GPUs, each providing 48 GB of VRAM. The synthetic data augmentation pipeline was executed on a separate workstation with an NVIDIA RTX 4060 Ti GPU with 16 GB of VRAM, an AMD Ryzen 5 8500G processor, and 32 GB of system memory.

3. Experimental Results

This section presents the experimental results obtained for both translation directions: Mexican Sign Language gloss to Spanish (MSLG2SPA) and Spanish to Mexican Sign Language gloss (SPA2MSLG). We evaluated the impact of three main factors: model size, linguistic information via POS annotations, and the percentage of synthetic data augmentation in the training set.

3.1. MSLG2SPA Results

Tables 2 and 3 present the experimental results for the MSL gloss to Spanish translation task (MSLG2SPA), ordered by BLEU score. Overall, this translation direction achieved substantially better performance than SPA2MSLG across all evaluation metrics, indicating that generating natural Spanish sentences from gloss representations is comparatively easier than producing compressed gloss sequences from Spanish text.

In our experiments, *Base* refers to fine-tuning using only the original training data provided by the shared task, while *POS* refers to fine-tuning using the training/augmented data with additional part-of-speech information in the input representation described in Section 2.3. Data augmentation percentages (*Augmentation*) refer to fine-tuning using the proportion of synthetic examples generated using the pipeline described in Section 2.4.

The best overall configuration was obtained using Qwen2.5-32B-Instruct with POS information and 75% synthetic data augmentation, achieving a BLEU score of 64.34, a chrF score of 78.19, a METEOR score of 0.782, and a COMET score of 0.862.

For the Qwen-32B model, POS-enhanced representations generally improved performance when combined with moderate or high levels of data augmentation. In particular, the POS + 75% augmentation configuration clearly outperformed all other variants. However, the benefits of POS information were less consistent at lower augmentation levels and with smaller models, as shown in Table 3.

The experiments also show that synthetic augmentation substantially improved performance compared to the baseline systems without augmentation. Configurations using 75% and 100% augmented data consistently produced the strongest results, especially for the 32B model. This suggests that the proposed generator–evaluator augmentation pipeline successfully increased linguistic diversity while preserving structural consistency in the generated training examples.

Additionally, larger models consistently outperformed the 7B variants across nearly all configurations, confirming that model capacity plays an important role in low-resource sign language translation tasks. While the 7B models achieved competitive results under high augmentation settings, the 32B models demonstrated superior robustness and better overall generalization.

Table 2
Best MSLG2SPA results using Qwen2.5-32B-Instruct ordered by BLEU score.

Configuration	BLEU	chrF	METEOR	COMET
POS + 75% Augmentation	64.34	78.19	0.782	0.862
POS + 100% Augmentation	60.63	76.73	0.768	0.850
Base + 75% Augmentation	59.44	74.39	0.762	0.844
Base + 100% Augmentation	58.04	76.99	0.769	0.847
POS + 35% Augmentation	38.73	66.82	0.669	0.829
Base + 50% Augmentation	38.73	67.86	0.670	0.841
Base + 35% Augmentation	38.67	66.74	0.667	0.844
POS + 50% Augmentation	38.56	67.23	0.665	0.837
POS + 20% Augmentation	37.93	68.11	0.629	0.806
Base + 20% Augmentation	34.27	68.33	0.631	0.812
POS + 10% Augmentation	33.01	65.41	0.584	0.793
Base + 5% Augmentation	31.20	63.74	0.562	0.800
Base + 10% Augmentation	31.20	63.74	0.562	0.800
Base	31.08	64.22	0.559	0.811
POS	29.23	55.40	0.489	0.738
POS + 5% Augmentation	29.23	55.40	0.489	0.738

3.2. SPA2MSLG Results

Tables 4 and 5 present the experimental results (top-5 best results) for the Spanish-to-Mexican Sign Language Gloss translation task (SPA2MSLG), ordered by BLEU score. In contrast to the MSLG2SPA direction, the overall performance in SPA2MSLG was substantially lower across all evaluation metrics, confirming the greater difficulty of gloss generation. Nevertheless, data augmentation consistently improved translation quality compared to the baseline systems, particularly for the 32B model.

The best-performing configuration was obtained with Qwen2.5-32B-Instruct and the Base representation, with 50% synthetic augmentation, achieving a BLEU score of 36.89, a chrF score of 70.13, a METEOR score of 0.610, and a COMET score of 0.745. Unlike in the MSLG2SPA direction, incorporating

Table 3
Best MSLG2SPA results using Qwen2.5-7B-Instruct ordered by BLEU score.

Configuration	BLEU	chrF	METEOR	COMET
POS + 100% Augmentation	42.96	68.60	0.681	0.853
Base + 100% Augmentation	42.96	68.60	0.681	0.873
Base + 75% Augmentation	42.96	68.60	0.681	0.873
POS + 75% Augmentation	42.89	69.26	0.686	0.837
Base + 20% Augmentation	40.78	67.76	0.663	0.861
Base + 35% Augmentation	40.43	66.91	0.642	0.854
Base + 50% Augmentation	40.43	66.91	0.642	0.854
POS	39.28	68.78	0.675	0.817
POS + 35% Augmentation	35.70	65.36	0.615	0.829
POS + 50% Augmentation	35.61	64.38	0.571	0.799
Base	34.71	60.49	0.570	0.809
POS + 5% Augmentation	31.76	59.79	0.627	0.813
POS + 10% Augmentation	31.35	59.23	0.605	0.813
Base + 5% Augmentation	30.96	57.48	0.586	0.821
Base + 10% Augmentation	30.96	57.48	0.586	0.821
POS + 20% Augmentation	26.20	57.70	0.578	0.810

POS annotations did not consistently improve performance in SPA2MSLG and, in several cases, slightly degraded BLEU scores. These results suggest that gloss generation may rely more heavily on semantic compression and sequence abstraction than on explicit shallow syntactic information.

Additionally, larger models consistently outperformed the 7B variants, indicating that increased model capacity remains beneficial for low-resource sign language generation tasks despite the inherent ambiguity of gloss representations.

Table 4
Best SPA2MSLG results using Qwen2.5-32B-Instruct ordered by BLEU score.

Configuration	BLEU	chrF	METEOR	COMET
Base + 50% Augmentation	36.89	70.13	0.610	0.745
POS + 50% Augmentation	24.78	60.71	0.449	0.718
Base + 20% Augmentation	19.65	62.84	0.492	0.722
Base + 35% Augmentation	19.48	65.57	0.575	0.718
POS + 35% Augmentation	19.26	64.40	0.445	0.714

Table 5
Best SPA2MSLG results using Qwen2.5-7B-Instruct ordered by BLEU score.

Configuration	BLEU	chrF	METEOR	COMET
Base + 20% Augmentation	12.92	51.82	0.284	0.658
Base	7.58	44.11	0.294	0.623
POS	7.29	43.39	0.267	0.613
POS + 10% Augmentation	4.80	50.07	0.283	0.694
Base + 10% Augmentation	4.55	50.19	0.247	0.679

The results indicate that the MSLG2SPA direction consistently outperformed SPA2MSL across all metrics. The best system achieved a BLEU score of 64.34 using the Qwen2.5-32B-Instruct model with POS information and 75% of data augmentation. The incorporation of linguistic information generally improved performance in the MSLG2SPA task, particularly for larger models. However, the impact was less consistent in the SPA2MSLG task.

Additionally, larger models consistently outperformed smaller ones, suggesting that model capacity plays an important role in low-resource sign language translation tasks.

Given the substantially better development performance in MSLG2SPA, we submitted our results for this task only to the shared evaluation.

In the official IberLEF 2026 MSLG2SPA evaluation, our team `Ludwig` achieved competitive results among all participating teams. The submitted model *Ludwig_QW-POS75DA_MSL2SPA*, based on Qwen2.5-32B-Instruct with POS augmentation and 75% synthetic data augmentation, obtained a BLEU score of 45.49, a METEOR score of 0.6473, a chrF score of 70.66, and a COMET score of 0.8804. This system ranked 14th overall among all submitted runs, placing the Ludwig team in the 6th position in the official team ranking for the MSLG2SPA shared task.

Additionally, our second submitted system, *Ludwig_QW-75DA_MSL2SPA*, achieved the 15th overall position with a BLEU score of 45.38. These results demonstrate that the proposed instruction-tuned LLM approach with linguistic and synthetic data augmentation is competitive for low-resource Mexican Sign Language gloss translation.

4. Conclusions

In this work, we explored instruction-tuned large language models for bidirectional translation between Mexican Sign Language glosses and Spanish in the IberLEF 2026 shared task. Our experiments demonstrated that the MSLG2SPA direction is substantially easier than SPA2MSLG, achieving significantly higher scores across all evaluation metrics. The best configuration, based on Qwen2.5-32B-Instruct with POS-enhanced inputs and 75% synthetic data augmentation, achieved a BLEU score of 64.34 in the MSLG2SPA task.

The results also showed that synthetic data augmentation was highly beneficial in low-resource settings. In particular, moderate and high augmentation levels consistently improved performance, suggesting that the proposed generator–evaluator pipeline successfully increased linguistic diversity while preserving structural consistency in the generated examples. Additionally, linguistic augmentation with POS information improved performance primarily for the larger 32B model and the MSLG2SPA direction, while its effect was less consistent for the SPA2MSLG direction.

Across both translation directions, larger models consistently outperformed smaller ones, indicating that model capacity plays an important role in sign language translation tasks involving structurally divergent representations. However, the SPA2MSLG task remained considerably more challenging, likely due to the compressed and ambiguous nature of gloss sequences.

In the official IberLEF 2026 MSLG2SPA evaluation, our proposed approach (Ludwig team) achieved competitive performance among participating systems. The best submitted model, based on Qwen2.5-32B-Instruct with POS augmentation and 75% synthetic data augmentation, ranked 14th overall and positioned the team in 6th place in the official team ranking. These results confirm that instruction-tuned large language models combined with linguistic and controlled synthetic data augmentation constitute an effective strategy for low-resource Mexican Sign Language gloss translation.

Future work includes exploring larger, more diverse synthetic corpora, incorporating additional linguistic features such as dependency relations and dependency trees, and evaluating multimodal approaches that integrate visual sign information with gloss representations.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT based on GPT-5.5 and DeepSeek-V4 in order to support language refinement, drafting, and editing of the manuscript. The authors reviewed, edited, and validated all generated content and take full responsibility for the publication's content.

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] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: efficient finetuning of quantized llms, 2023.

[5] A. B. Varanasi, M. Sinha, T. Dasgupta, C. Jadhav, Linguistically informed transformers for text to American Sign Language translation, in: A. K. Ojha, C.-h. Liu, E. Vylomova, F. Pirinen, J. Abbott, J. Washington, N. Oco, V. Malykh, V. Logacheva, X. Zhao (Eds.), Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 50–56. URL: https://aclanthology.org/2024.loresmt-1.5/. doi:10.18653/v1/2024.loresmt-1.5.

[6] 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., 2022. URL: http://dblp.uni-trier.de/db/conf/iclr/iclr2022.html#HuSWALWWC22.

[7] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: efficient finetuning of quantized llms, 2023.

[8] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, 2025. URL: https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

[9] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python, 2020. doi:10.5281/zenodo.1212303.

[10] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[11] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049/. doi:10.18653/v1/W15-3049.

[12] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[13] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online,

2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213/. doi:10.18653/v1/2020.emnlp-main.213.

[14] LangChain, Langgraph: Build resilient language agents as graphs, 2025. URL: https://github.com/langchain-ai/langgraph, accessed: 2026-04-20.

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/. doi:10.18653/v1/2020.emnlp-demos.6.