

# Advancing Sign Language Translation through Large Language Models

Anonymous COLING 2025 submission

## Abstract

This paper explores multi-step Sign Language Translation (SLT) using Large Language Models (LLMs) and Formal SignWriting (FSW), a system of writing sign languages for the intermediate step between the written form of both languages. Recent research for multi-step SLT represents the sign languages using glosses, which has limitations of not capturing all non-manual signals. The paper focuses on the translation between American Sign Language (ASL) and English. We use Large Language Models with only a decoder architecture, to learn the mappings between FSW notation and the English words, utilizing Supervised Fine-Tuning (SFT), with a combination of Contrastive Policy Optimization (CPO) and Simple Preference Optimization (SimPO) to fine-tune the LLMs, on the SignBank + dataset. The results are evaluated on a benchmark dataset, with the automated metrics BLEU, chrF2++ score and G-Eval-MQM metric. Achieving comparable performance with existing research, with a BLEU score of 18.07 and chrF2++ score of 55.44 for the text-to-FSW task and a BLEU score of 26.03 and chrF2++ score of 37.38 for the FSW-to-text task. The paper also proposes a combination of the G-Eval framework and the defined Multidimensional Quality Metric (MQM) framework for machine translation, G-Eval-MQM, which has been further adapted to facilitate SLT evaluation for FSW notation, with scores 14.26 and 18.46 for both tasks respectively. We show that our approach achieves comparable performance to existing research which

utilize Neural Machine Translation (NMT) frameworks and showcases the potential for LLMs to be utilized as the model for the intermediary step for SLT, advancing the performance of multi-step SLT. While also highlighting techniques that can be utilized in advancing Machine Translation (MT) for low-resource languages.

## 1 Introduction

Over 5% of the world's population have disabling hearing loss, this amounts to 430 million people. These numbers are projected to rise to 750 million people by the year 2050.<sup>1</sup> The hearing-impaired communities communicate over sign language, which is different from spoken languages, with its own vocabulary, grammatical rules and syntax. However, majority of digital media only provide output modalities for the hearing community, it can be difficult for members of the hearing disabled community to gain the benefits that those without hearing impairments do (Bragg et al., 2019).

In recent years with the progress made in machine learning, significant research has gone into the translation of sign language into spoken languages known as Sign Language Recognition (SLR), and while more research is going into spoken language to sign language translation known as Sign Language Production (SLP), they are not as prominent. There have been recent developments in Sign Language Translation (SLT) leveraging techniques such as Avatar Approaches, Neural Machine Translation (NMT), Motion Graph Approaches, Conditional image/video generation approaches and more recently Large Language Models (LLMs) (Ratsgoo et al., 2021). These techniques utilize a multi-step approach to

---

<sup>1</sup> <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>

achieve the translation, translating from the spoken language in written form to sign language in written form before converting the sign language in written form to pose, video or animation and vice versa (Jiang et al., 2023). This paper focuses on researching the translation of spoken language in written form to sign language in written form and vice versa, to facilitate SLT in both directions, utilizing LLMs.

LLMs have shown remarkable application in the translation between spoken languages, which has encouraged researchers to apply these techniques to SLT. For these translations, different methods for representing sign language in written form are used, such as glosses, skeletal poses, and notations such as SignWriting and HamNoSys. Majority of the research in spoken language to SLT leverage glosses as the sign language representation, which unlike other notations like SignWriting omit non manual signals from a signer (Ratsgoo et al., 2021). Sign Writing also has other advantages such as being universal, being extensively documented, computer-supported and is used by researchers in the hearing-impaired community (Jiang et al., 2023). This paper aims to explore the impact of different LLMs, along with different techniques for training the LLMs on a machine translation task, on SLT performance. The experiment will leverage the SignBank + dataset, to train LLMs using techniques such as few-shot learning, zero-shot-learning, and fine-tuning, on SignWriting to Text and Text to SignWriting, comparing the performance with current state-of-the-art (SOT) SLT leveraging FSW, the ASCII representation of SignWriting. This paper focusses its efforts on American Sign Language (ASL) Translation.<sup>2</sup>

For this research the large language models Meta-Llama-3-8B-Instruct to be referred to as the meta model and Mistral-7B-Instruct-v0.3 to be referred to as the mistral model henceforth, are selected due to their performance in translation tasks as medium sized large language models and their performance in Supervised Fine-Tuning (SFT) tasks. The model selection is restricted to medium sized models as even with model quantization techniques and parameter efficient

fine-tuning, these are the models sizes trainable on more cost-effective commercially available resources.<sup>3</sup> To evaluate the results of these experiments, this research will leverage the BLEU score and n-gram overlap evaluation technique (Post, 2018), along with CHRF2++, which captures word-level and character level statistics (Popović et al., 2017). In addition, the G-Eval framework proposed by Liu et al., 2023, which leverages LLMs with Chain of Thought (CoT) prompting and a form filling paradigm will be adapted with the Multidimensional Quality Metric (MQM) framework for Machine Translation (MT) proposed by Freitag et al., 2021, to evaluate the results.

## 2 Related Work

As of the time of writing, the paper is only aware of the efforts of Jiang et al., 2023 and Moryossef and Jiang, 2024 on the application of NMT Techniques on SLT utilizing FSW notation. Utilizing open source NMT frameworks such as Sockeye, Fairseq, OpenNMT, and the mT5 model. Achieving comparable performance across both papers with a BLEU score 24.65 and 24.33, and chrF2++ score of 31.22 and 27.88 with the Sockeye model on the text-to-FSW task respectively.

With the advent of LLMs, more research has gone into leveraging them in achieving SLT tasks, such as in (Moryossef and Jiang, 2024), an LLM, the gpt-3.5-turbo was utilized to clean and expand the SignBank dataset (a dataset which contains sentence-pairs for spoken language and FSW),<sup>4</sup> creating a new dataset SignBank +. Most recent research, however, has utilized LLMs in SLT by incorporating glosses as the representation for sign language. For instance, Lee et al., 2023, exploring the application of vocabulary sharing with LLM on SLT utilizing the NIASL2021 dataset for Korean Language to Korean Sign Language translation. The research explores different LLM architectures such as Encoder-Decoder, and Decoder only Transformers, both with and without pre-training. With the Ko-GPT-Trinity-1.2B model with

<sup>2</sup>

<https://datatracker.ietf.org/doc/pdf/draft-slevinski-formal-signwriting-09.pdf>

<sup>3</sup>

<https://medium.com/@ccibeekeoc42/unl>

[locking-low-resource-language-understanding-enhancing-translation-with-llama-3-fine-tuning-df8f1d04d206](https://arxiv.org/abs/2308.04420)

<sup>4</sup>

<http://www.signbank.org/signpuddle>

vocabulary sharing achieving BLEU score 22.06 on Gloss-to-Text and 45.89 on Text-to-Gloss.

Other research efforts include (Fang et al., 2024), which developed SignLLM, a comprehensive LLM pipeline, which integrates two transformer approaches, the Multi-Language Switching Framework and Prompt2LangGloss, further incorporating reinforcement learning techniques with a Priority Learning Channel. These methods are employed for the text-to-pose SLT task, achieving a 23.25 BLEU-4 score and 49.52 ROUGE score. Their research also provides a new dataset for training multilingual sign language tasks called prompt2sign, containing prompts, text, video frames and pose data key points. Another research includes the incorporation of an LLM in SLR to improve the gloss sentence generation in video-to-gloss, as demonstrated in (Sincan et al., 2024). Jung-Ho et al., 2024, also contributed to the field with their paper on SLT evaluation, where they utilized LLMs to assess their newly proposed gloss multi-channel evaluation metric, SignBLEU.

Following the increased application of LLMs on MT tasks, increasing research into improving the performance of LLMs on these tasks have been published. Some of this include the works of (Xu et al., 2024), which propose two-step fine-tuning. The first involving fine-tuning of the base model on monolingual data for all the languages covered in the task and the second involving SFT, leveraging high quality parallel data, resulting in new models referred to as Advanced Language Models based translators (ALMA). Following their research, (Xu et al., 2024), propose Contrastive Preference Optimization (CPO) a policy optimization approach which leverages preference data further improving the results of (Xu et al., 2024) creating new models, ALMA-R, achieving SOT performance on MT tasks with medium sized LLMs. Other research into improving LLM performance in MT tasks include (Moslem and Haque, 2023; Jiao et al., 2023; Zeng et al., 2024), however, the research by (Xu et al., 2024) and (Xu et al., 2024) will be adapted and used in the experimentation for this paper, due to their SOT performance.

### 3 Methodology

#### 3.1 Data Collection and Preprocessing

The dataset used in this project, is the cleaned SignBank+ dataset (Moryossef and Jiang, 2024),

this contains spoken language text along with its accompanying FSW notation for ASL, German Sign Language and others. Table 1., shows the details of the dataset used, with the following language pairs; pt-bzs as Brazilian Portuguese & Brazilian Sign Language, en-ase as American English & American Sign Language, de-gsg as Standard German & German Sign Language, fr-fcs as Canadian French & Quebec Sign Language, en-bfi as British English & British Sign Language, and ga-isg as Irish & Irish Sign Language.

SubDataset	Language pair	Samples
gpt-3.5-cleaned	pt-bzs	52,221
gpt-3.5-cleaned	en-ase	30,202
gpt-3.5-cleaned	de-gsg	24,656
gpt-3.5-cleaned	fr-fcs	11,119
fingerspelling	en-ase	28,122
fingerspelling	en-bfi	23,771
fingerspelling	ga-isg	23,716
bible	en-us	13,320

Table 1: Different sub-datasets in the SignBank+ dataset with over 10K sample language

In Table 1., gpt-3.5-cleaned refers to the subset cleaned using gpt-3.5-turbo, using prompt engineering. Fingerspelling refers to a subset where the words do not have specific signs, therefore are signed letter-by-letter, e.g. nouns not typically found in a dictionary. And lastly the bible represents the ASL translation of the English bible.

For preprocessing, the dataset is transformed to prompts, for both text-to-FSW and FSW-to-text, utilizing the prompt format recommended for the models used in the experiments. These prompts will then be split into train, validation, and test, which will be used in finetuning the models on the dataset and evaluating them, 10 samples will be selected at random from the training set to provide few-shot examples. The ASL dataset in signbank+ is separated by the • (U+16EB) character, separating terms which translate to the same FSW notation [4]. After which a randomization method is used to select from within the dataset dis-preferred examples of both English text and FSW notations for the Reinforcement Learning from Human Feedback (RLHF) experiment. Finally, the prompts are then filtered to reduce the sequence length to under 2500.

For tokenization, although (Jiang et al., 2023) have provided a tokenization technique for signwriting, the tokenizer for the selected LLMs was used to tokenize the prompts. After experimenting with a few training steps (5000), the model performance was subpar with BLEU scores

0.030 and 0.026 for FSW-to-text and text-to-FSW tasks respectively, leading to further review of techniques to improve the performance. Of which, further training the model tokenizer with the prompts containing both English text and FSW notation to allow the tokenizer to better tokenize the FSW notation for ASL was utilized,<sup>5</sup> improving the model performance at 5000 steps to 0.520 and 0.095 respectively.

### 3.2 Model Selection

The model selection is constrained by budget limitations, therefore models utilized in these experiments will be limited to open-source models. Additional selection criteria considered, include the model’s performance on low-resource translation tasks.

Based on the LLM rankings in translation task, the pre-trained model “Meta-Llama-3-8B-Instruct” was selected for experimentation. This model was selected not only for its performance in SFT, but also for its capability in, human alignment through RLHF and policy optimization.<sup>2</sup> The model is an auto-regressive decoder only large language model which uses an optimized transformer architecture, with 8 billion parameters, with an 8000 token context window. The model was tuned with SFT and RLHF to improve its alignment with human preferences for helpfulness and safety. The model also utilizes Grouped-Query Attention (GQA) for improved inference scalability. The previous generation of the model has been popularly used in LLM machine translation research, as can be seen in the following papers (Jiao et al., 2023; Zeng et al., 2024; Xu et al., 2024; Xu et al., 2024). The model was pretrained on over 15 trillion tokens of publicly available data and fine-tuned on publicly available instruction dataset and over 10 million human annotated examples.<sup>6</sup>

The second model used for experimentation is the mistral AI model “Mistral-7B-Instruct-v0.3”, also selected due to its performance on low-resource translation task and its competitive performance to the llama model (Medium, 2024). The model is an auto regressing decoder only model as well, with 7.3 billion parameters, utilizing

a sliding window attention, GQA with a byte-fallback BPE tokenizer. The model is also pretrained on publicly available data and on publicly available chat datasets.<sup>7</sup> The previous generation of the model has also been used in some machine translation research such as (Moslem and Haque, 2023).

The model selection was also impacted by research into instruction tuning which involves further training an LLM on an instruction and output pair. The method improves the model’s controllability, improves predictability and constrain the output in alignment with the desired patterns and or domain knowledge. This method of tuning is also computationally efficient and supports the adaptation of the model to a domain, without the need for an architecture change or extensive retraining (Zhang et al., 2023). The instruct versions of the llama and mistral models facilitates the paper to apply instruction tuning techniques to SFT of the models, leveraging the advantages of instruction tuning without any additional overhead to train the models to understand and follow instructions.

### 3.3 Experiment Setup

This paper explores the performance of different LLM domain adaptation methods in SLT. Comparing the performance of zero-shot prompting on the base model, with few-shot prompting on the base model, with SFT of the base model with parallel data and then CPO (Xu et al., 2024) combined with Simple Preference Optimization (SimPO) (Meng et al., 2024) on the SFT tuned model. Below the steps taken to set up each of these experiments are described, starting with the selection of the prompt to be used.

The prompts were engineered by prompting the base models, until settling on an efficient prompt which enables the LLMs return preferable responses to the task without requiring further context. The prompts were then formatted leveraging the advised template from the model documentation pages as can be seen in Appendix A., for the llama<sup>8</sup> and mistral<sup>9</sup> models respectively.

<sup>5</sup> Training a new tokenizer from an old one: <https://huggingface.co/learn/nlp-course/chapter6/2>

<sup>6</sup> <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

<sup>7</sup> <https://mistral.ai/news/announcing-mistral-7b/>

<sup>8</sup> <https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-3>

<sup>9</sup> <https://www.promptingguide.ai/models/mistral-7b>

**Zero-Shot Learning setup:** In the default setup the base model is tasked with translating samples for both text-to-FSW and FSW-to-text, without any prior sample input.

**Few-Shot Learning setup:** An instance of the base-model is provided with a prompt and randomly selected examples from the training set as can be seen in Appendix B. and is evaluated with samples from the test set.

**SFT setup:** The models were loaded using Hugging Face transformer library and were fine-tuned using SFT, chosen for its effective performance on low-resource datasets (Medium, 2024). The setup utilizes, Parameter Efficient Fine-Tuning (PEFT) with Quantized Low-Rank Adaptation (QLoRa), leveraging double quantization to 4-bit, to improve computational efficiency during model fine-tuning, maximizing the limited budget (Detrmers et al. 2023). The unsloth<sup>10</sup> library is further leveraged to improve model training speed and reduce GPU ram usage, improving training speed by 2.2 times and reducing GPU ram usage by up to 73%.

For the meta model, hyperparameter tuning was carried out over a short training period of 5000 steps adjusting the model training parameters based on the performance and finally settling on the parameters which provided the best model results of 0.52 BLUE score for the text-to-FSW task. These parameters are a learning rate of 2e-4 for training stability, the “paged\_adamw\_8bit” optimizer to leverage the efficiency provided by the quantization, a lora rank of 64 and a lora alpha of 32 as recommended by (Detrmers et al. 2023), resulting in the training of 0.15% of available parameters and a lora drop out of 0.1 to account for overfitting to the dataset. The batch size used in training was limited to 1 due to limited memory, resulting in 196,800 steps to train one epoch. Validation is carried out every 1000 steps with 10,360 parallel samples in both translation directions. The model was trained on the Google collab environment with the L4 compute instance with 19.3/22 GB of GPU ram used.

For the mistral model, to reduce cost overhead, the same parameters used for the llama model was utilized for the mistral model, with a few exceptions such as the batch size, which was increased to 4 to fully leverage the compute

resource given the lower memory requirements of the mistral model compared to the llama model, leading to 49,210 steps to train one epoch. The same lora parameters were used, however only 0.06% of the available parameters were trained. The model was trained on the Google collab environment using the T4 compute instance with 13.4 /15 GB of GPU ram used.

**SFT + CPO with SimPO setup:** the SFT tuned models were further fine-tuned on a combination of CPO and SimPO both of which are improvements to the Direct Preference Optimization (DPO) technique for (RLHF) (Rafailov et al., 2024). The CPO technique improves upon SFT by providing a means for a model to learn to reject poor translations, by incorporating a BC-Regularizer with the approximated uniform reference model from DPO to steer the model towards the preferred data distribution (Xu et al., 2024), resulting in more performant and memory efficient training and improved performance.

$$LCPO(\pi\theta; U) = -E(x, yw, yl) \sim D[\log \sigma(\beta \log \pi\theta(yw|x) - \beta \log \pi\theta(yl|x)) + \log \pi\theta(yw|x)] \quad (1)$$

The SimPO Technique, improves upon DPO technique by removing the requirement for a reference model, using the average log probability of a sequence as the reward, and additionally adding a target reward margin to the objective maximizing the margin between the preferred response and dis-preferred response, further improving the performance. This technique also incorporates length normalization, reducing the likelihood for the model to generate inaccurate and lengthy responses (Meng et al., 2024).

$$LSimPO(\pi\theta; U) = -E(x, yw, yl) \sim D[\log \sigma(\beta |yw| \log \pi\theta(yw|x) - \beta |yl| \log \pi\theta(yl|x) - \gamma)] \quad (2)$$

The combination of both CPO and SimPO can be leveraged by incorporating the length normalization and target reward margin from the SimPO technique within the CPO technique which has been explored by (CPO\_SIMPO, 2024).

$$LCPO - SimPO(\pi\theta; U) = -E(x, yw, yl) \sim D[\log \sigma(\beta |yw| \log \pi\theta(yw|x) - \beta |yl| \log \pi\theta(yl|x) - \gamma) + a \log \pi\theta(yw|x)] \quad (3)$$

<sup>10</sup> <https://github.com/unslothai/unsloth>

The hyperparameters used in these experiments, were selected from (Xu et al., 2024; Meng et al., 2024) and with some modifications added from (CPO\_SIMPO, 2024), these can be seen in Table 2.

Model	Learn ng rate	Alpha	Beta	Gamma
Meta model	1e-6	0.05	10	5.4
Mistral model	5e-7	0.05	2.5	0.25

Table 2: Hyperparameters for CPO-SimPO fine-tuning

The CPO-SimPO specific hyperparameters as seen in Table 3. represent, the  $\alpha$ ,  $\beta$  and  $\gamma$  as seen in the CPO-SimPO loss function above. Other hyperparameters not yet listed include, the lora rank which is 16 across both models, the lora alpha which is 32 for both models and the optimizer used, which was the `paged_adamw_8bit` optimizer in both models, to avoid out of memory (OOM) problems. Gradient checkpointing was also employed in both experiments to reduce memory requirements with a gradient accumulation step of 8 being selected. Both models were trained over 1 epoch on a dataset of 42k parallel translations in both directions utilizing the reference translation as the preferred and the randomly selected data in the preprocessing steps earlier as the dis-preferred translation. The models were fine-tuned with a batch size of 2 and 4 resulting in 2363 steps for the meta model and 1318 steps for the mistal model. The models were fine-tuned on a vast.ai environment with the A100 SXM4 80 GB GPU with a max usage of 59.2 GB and 70.6 of GPU memory being observed respectively.

### 3.4 Performance Evaluation Metric

Evaluation will be conducted using metrics previously employed by researchers, specifically those utilized in Signwriting studies (Jiang et al., 2023; Moryossef and Jiang, 2024). Of these metrics the BLUE score and chrF2++ will be used to evaluate text-to-FSW and FSW-to-text tasks. With their respective formulae shown below.

$$BLEU = BP \cdot \exp(\sum_{n=1}^N w_n \log p_n) \quad (4)$$

Were  $BP$  stands for Brevity Penalty,  $w_n$  is the weight for n-gram precision of order  $n$ ,  $p_n$  is the n-gram modified precision score of order  $n$ , and  $N$  is

the maximum n-gram order to consider (Papineni et al., 2002).

$$ngrF\beta = (1 + \beta^2) \frac{ngrP \cdot ngrR}{\beta^2 \cdot ngrP + ngrR} \quad (5)$$

Were  $ngrP$  and  $ngrR$  are n-gram precision and recall averaged over all n-grams and  $\beta$  is a parameter which assigns  $\beta$  times more weight to recall than precision (Popović et al., 2017).

The G-Eval framework proposed by Liu et al., 2023, is used to evaluate the output of the fine-tuned results from both models with gpt-4o-mini released by OpenAI<sup>11</sup> as the evaluation LLM. The criterion for evaluation is adapted from Freitag et al., 2021, which suggests a MQM Framework, adapted specifically for human evaluation of translations. This paper adapts the category of the MQM topology proposed by Freitag et al., 2021, absorbing subcategories into the main categories and adapting the criteria to sign language in FSW notation. Resulting in the categories; accuracy, fluency, terminology, style and locale convention, in respective order of topology. The scoring system is also modified due to the difference is assigned score from G-Eval implemented using Deepval,<sup>12</sup> returning scores between 0-1 for each criterion. The total max score of 25 is retained with, however instead of non-translation as a category, it is added as a severity when accuracy score is near zero from G-Eval. The adapted MQM category topology with their criteria and adapted scoring model leveraging the MQM weighting from (Freitag et al., 2021), can be seen in Appendix C.

## 4 Results and Discussion

The different experiments were evaluated using the BLEU score and chrF2++, Table 3., below shows the average score of 500 samples taken from the test dataset. Whilst the G-Eval-MQM metric described earlier will be carried out for 100 of those samples, on the fine-tuning experiments. The SFT experiment was early stopped for both the meta and the mistral model at 140,000 steps and 29,000 steps, with a validation loss of 0.42 and 0.28 respectively, due to computation resource constraints, and to facilitate the execution of the CPO-SimPO experiments in a timely manner.

The CPO-SimPO tuned experiment showcased the highest performance across both models on the

<sup>11</sup> <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence>

<sup>12</sup> <https://docs.confident-ai.com/docs/metrics-llm-evals>

Experiments	Model	Task	BLEU	chrF2++	G-Eval-MQM
SFT + CPO-SimPO	Meta-Llama-3-8B-Instruct	SignWriting-to-Text	26.03	37.38	18.46
SFT + CPO-SimPO	Meta-Llama-3-8B-Instruct	Text-to-SignWriting	18.07	55.44	14.26
Supervised Fine-Tuning	Meta-Llama-3-8B-Instruct	SignWriting-to-Text	26.57	37.30	19.18
Supervised Fine-Tuning	Meta-Llama-3-8B-Instruct	Text-to-SignWriting	16.87	55.37	14.11
Few-Shot Prompting	Meta-Llama-3-8B-Instruct	SignWriting-to-Text	0.92	6.86	N/A
Few-Shot Prompting	Meta-Llama-3-8B-Instruct	Text-to-SignWriting	0.00	17.71	N/A
Zero-Shot Prompting	Meta-Llama-3-8B-Instruct	SignWriting-to-Text	0.33	5.70	N/A
Zero-Shot prompting	Meta-Llama-3-8B-Instruct	Text-to-SignWriting	0.00	1.07	N/A
SFT + CPO-SimPO	Mistral-7B-Instruct-v0.3	SignWriting-to-Text	3.00	3.81	24.73
SFT + CPO-SimPO	Mistral-7B-Instruct-v0.3	Text-to-SignWriting	2.31	33.45	20.8
Supervised Fine-Tuning	Mistral-7B-Instruct-v0.3	SignWriting-to-Text	3.00	3.77	25
Supervised Fine-Tuning	Mistral-7B-Instruct-v0.3	Text-to-SignWriting	2.31	33.45	20.16
Few-Shot Prompting	Mistral-7B-Instruct-v0.3	SignWriting-to-Text	0.11	4.47	N/A
Few-Shot Prompting	Mistral-7B-Instruct-v0.3	Text-to-SignWriting	0.00	13.86	N/A
Zero-Shot prompting	Mistral-7B-Instruct-v0.3	SignWriting-to-Text	0.23	3.73	N/A
Zero-Shot prompting	Mistral-7B-Instruct-v0.3	Text-to-SignWriting	0.00	1.99	N/A

Table 3: Translation quality of text- signwriting and signwriting-to-text

Text-to-FSW task, on all metrics, with the meta model showcasing the best performance with 18.07 BLEU score, 55.44 chrF2++ score and 14.26 G-Eval-MQM score. Although not at the level of previous research leveraging the Sockeye Model, with a BLEU score of 35.7, however, it showcases a comparable chrF2++ score to 58.4 from (Jiang et al., 2023). These results showcase the possibility of teaching an LLM to carry out SLT with FSW as the intermediary text. Improving the base model’s performance on the task to 18.07 BLEU score and 55.44 chrF2++ score from 0.00 and 1.07 on zero-shot learning, and 0.00 and 6.86 on few-shot learning respectively. These results also showcase the benefits of RHLF leveraging CPO-SimPO, improving the performance over SFT by 1.2 BLEU score and 0.07 chrF2++ score, albeit minimal. This minimal improvement over the SFT experiment can be attributed to the dataset generation which utilized random data from the same dataset for the dis-preferred data as opposed to utilizing outputs from the SFT models as the dispreferred data (Xu et al., 2024).

For the FSW-to-Text task, the meta model performs best, with the SFT experiment showcasing a BLEU score of 26.57 slightly over the CPO-SimPo experiment by 0.54, however, on chrF2++ metric the CPO-SimPO performs slightly better by 0.35. The G-Eval-MQM score on this task are 19.18 and 18.46 respectively. The experiment performs better than Jiang et al., 2023, multilingual model on 21 language pair, with a BLEU score of 25.0, showcasing a 1.57 score increase, however it falls short on the chrF2++ score, which was 47.0. As with the previous task, these results still showcase the significant performance gain over the base model with an increase of 25.10 and 25.7 BLEU scores over few-shot and zero-shot prompting respectively. Possible consideration for improvement of the translation for these experiments is the inclusion of the monolingual data fine-tuning (Xu et al., 2024), as this was substituted for training the tokenizer on the dataset improving the model’s ability to embed FSW. Across all fine-tuning experiments the mistral model fails to show similar improvements as the llama model. With a minimal improvement of 0.07 BLEU score and 0.04 chrF2++ score over the base model with zero-shot prompting on the FSW-to-Text task. Although it showcases a bit better improvement in the Text-to-FSW task with an improvement of 3.00 BLEU score and 31.46 chrF2++ score. This poor performance could be attributed to the lack of hyperparameter optimization, utilizing the hyperparameters from tuning the meta model, and lack of fine-tuning on a monolingual dataset (Xu et al., 2024).

## 5 Conclusion

In this study, the application of LLMs in SLT was explored utilizing the signbank+ dataset, a

multilingual dataset, of which the ASL and English sentence pairs, was selected and prepared. Different LLM training methods were set up as experiments, with the selected models Meta-Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.3, to evaluate the performance of LLMs in SLT tasks, text-to-FSW and FSW-to-text. The combined CPO and SimPO (CPO-SimPO) experiment on SFT tuned meta model resulted in the highest performance for the text-to-FSW task with a BLEU score of 18.07, chrF2++ score of 55.44 and G-Eval-MQM score of 14.26. While on the signwriting-to-text task the SFT experiment and the CPO-SimPO experiments on the meta model, perform similarly with BLEU scores of 26.57 and 26.03, chrF2++ scores of 37.30 and 37.38, and G-Eval-MQM score of 19.18 and 18.46 respectively. The results of these experiments are comparable to previous research in SLT leveraging FSW notation as an intermediary translation. These results show the capability to train an LLM to learn the patterns of FSW and map them to English text and vice versa. Showcasing the significant role LLMs can play in improving the performance of multi-step SLT solutions, by leveraging these models to perform the intermediary translation from sign language in FSW to English texts and vice versa. The paper also proposes an adapted evaluation framework combining G-Eval framework and the MQM framework, termed G-Eval-MQM which shows decent correlation with other automatic metrics like BLEU score and chrF2++ score, while adequately evaluating FSW notations leveraging gpt-4o-mini as the evaluation model.

## 6 Ethical Considerations

The following considerations were made during the development of the experiments, to tackle some biases that could be inherent in the data as these were curated from public entries. Different models were selected to address social biases that may have been present in the pretraining data for the models. The models were also selected based on their adherence and transparency on responsible AI as can be seen from llama 3 responsible use guidelines,<sup>13</sup> and self-reflection guardrails for mistral models<sup>14</sup> and the results of responsible AI safeguard evaluation for llama models (Wang et al, 2024).

<sup>13</sup> <https://ai.meta.com/static-resource/july-responsible-use-guide>

## 7 Limitations

### 7.1 SignWriting Translation Evaluation

The proposed G-Eval-MQM metric adequately accesses the performance of the translations showing correlations with metrics like BLEU and chrF2++. However, some limitations exist, such as consistency, as the framework can return different scores given the same input, however, these scores are often within the same error severity. Another limitation is the overlap of other metrics on accuracy, as during fine-tuning of the prompts it was noticed the model will often still use accuracy as part of the criterion in other categories e.g with reasons such as “The Actual Output is grammatically correct, but it does not match the Expected Output, which contains different names”. Given a translation output of “my name is patrick cliff” and an expected output of “my name is phillip clark”. These limitations can be improved by employing better prompt-tuning techniques, as experimentation which led to the current prompts improved the original performance. Researchers could also enhance the proposed G-Eval-MQM metric by refining the criteria prompts, the error categories and the scoring model, creating an improved framework for automatic evaluation of SLT. Given the advent of model-based evaluation, this paper proposes further research into creating a model-based metric for SLT automatic evaluation building upon existing metrics such as BLEURT and COMET (Kocmi et al., 2024).

### 7.2 Resource Limitations

Due to resource limitations the study, could not carry out certain tasks which could have improved the results of the training, such as; fine-tuning of the LLMs with monolingual dataset from the different sign languages in FSW notation and spoken languages, before SFT, hyperparameter optimization for the mistral model and CPO-SimPO experiments, and curation of the preference dataset by inferencing the SFT models as opposed to randomization, which could have improved the performance of the CPO-SimPO experiment. The models selected and hyperparameters values used during experiments were also limited due to resource limitations. Finally, the SFT experiments had to be early stopped due to resource limitations.

<sup>14</sup> <https://docs.mistral.ai/capabilities/guardrailing>

## References

- Bragg D., Koller O., Bellard M., Berke L., Boudreaault P., Braffort A., Caselli N., Huenerfauth M., Kacorri H., Verhoef T., Vogler C., and Ringel Morris M. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*, pp. 16–31. Association for Computing Machinery, New York, NY, USA.
- Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L., 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. and Levine, S. (eds.) Advances in Neural Information Processing Systems*, vol. 36, Curran Associates, Inc., pp. 10088–10115.
- Fang, S., Wang, L., Zheng, C., Tian, Y., & Chen, C. (2024/05/17). SignLLM: Sign Languages Production Large Language Models, <https://doi.org/10.48550/arXiv.2405.10718>, last accessed 2024/05/25.
- Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W., 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9, pp.1460–1474. Cambridge, MA: MIT Press.
- Jiang, Z., Moryossef, A., Mülle, M. & Ebling, S.: Machine Translation between Spoken Languages and Signed Languages Represented in SignWriting. In: *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia, p. 1706–1724. Association for Computational Linguistics.
- Jiao, W., Huang, J., Wang, W., He, Z., Liang, T., Wang, X., Shi, S., Tu, Z., 2023. ParroT: Translating during Chat using Large Language Models tuned with Human Translation and Feedback. In: *Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, pp. 15009–15020. Association for Computational Linguistics.
- Jung-Ho, K., John, H.-E. M., Changyong, K., & Hui, L. D.: SignBLEU: Automatic Evaluation of Multi-channel Sign Language Translation. In: *N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenc, S. Sakti, & N. Xue (Ed.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, pp. 14796–14811. ELRA and ICCL.
- Kocmi, T., Zouhar, V., Federmann, C., Post, M., 2024. Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. In: *Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, pp. 1999–2014. Association for Computational Linguistics.
- Lee, H, Kim, J.-H., Jun, H. E., Kim, J., & Park, J. C.: Leveraging Large Language Models with Vocabulary Sharing For Sign Language Translation. In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, Rhodes Island, Greece. IEEE.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C., 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Bouamor, H., Pino, J. and Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 2511–2522. Association for Computational Linguistics.
- Medium, Difference between Trainer class and SFTTrainer (Supervised Fine tuning trainer) in Hugging Face?, <https://medium.com/@sujathamudadla1213/difference-between-trainer-class-and-sfttrainer-supervised-fine-tuning-trainer-in-hugging-face-d295344d73f7>, last accessed 2024/06/28.
- Medium, Exploring Mistral 7B for Low-Resource Language Understanding: A Fine-Tuning Approach to Machine Translation. <https://medium.com/@ccibekeoc42/exploring-mistral-7b-for-low-resource-language-understanding-a-fine-tuning-approach-to-machine-98a0b11fc085>, last accessed 2024/08/20.
- Meng, Y., Xia, M. and Chen, D., 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward. *ArXiv*, 2405.14734. <https://arxiv.org/abs/2405.14734>, last accessed 2024/08/29.
- Moryossef, A. and Jiang, Z. SignBank+: Preparing a Multilingual Sign Language Dataset for Machine Translation Using Large Language Models, <https://doi.org/10.48550/arXiv.2309.11566>, last accessed 2024/06/03.
- Moslem, Y., Haque, R., Way, A., 2023. Fine-tuning Large Language Models for Adaptive Machine Translation. <https://arxiv.org/abs/2312.12740>, last accessed 2024/08/20.
- Papineni, K., Roukos, S., Ward, T. and Zhu, W., 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *P. Isabelle, E. Charniak, and D. Lin, eds. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA pp.

311-318. Association for Computational Linguistics.

Popović, M., Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P. and Kreutzer, J., 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, pp. 612-618. Association for Computational Linguistics.

Post, M., 2018. A Call for Clarity in Reporting BLEU Scores. In: *O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, C. Monz, M. Negri, A. Névél, M. Neves, M. Post, L. Specia, M. Turchi and K. Verspoor, eds. Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium, pp.186-191. Association for Computational Linguistics.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D. and Finn, C., 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. ArXiv, abs/2305.18290. <https://api.semanticscholar.org/CorpusID:258959321>, last accessed 2024/08/29.

Rastgoo, R., Kiani, K., Escalera, S. and Sabokrou, M.: Sign Language Production: A Review. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA (2021). pp. 3446-3456.

Sincan, O. M., Camgoz, N. C., & Bowden, R. (2024/03/15). Using an LLM to Turn Sign Spotting into Spoken Language Sentences, <https://doi.org/10.48550/arXiv.2403.10434>, last accessed 2024/05/20.

The joint of Contrastive Preference Optimization (CPO) & Simple Preference Optimization (SimPO), [https://github.com/felixxu/CPO\\_SIMPO](https://github.com/felixxu/CPO_SIMPO), last accessed 2024/08/29.

Wang, Y., Li, H., Han, X., Nakov, P., & Baldwin, T. (2024). Do-Not-Answer: Evaluating Safeguards in LLMs. In *Y. Graham & M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024*, St. Julian's, Malta (pp. 896-911). Association for Computational Linguistics.

Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H., 2024. A paradigm shift in machine translation: Boosting translation performance of large language models, <https://doi.org/10.48550/arXiv.2309.11674>, last accessed 2024/08/20.

Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., Kim, Y.J., 2024. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In: *Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (Eds.), Proceedings of the 41st International Conference on Machine Learning*. PMLR, pp. 55204-55224.

Zeng, J., Meng, F., Yin, Y. and Zhou, J., 2024. Teaching Large Language Models to Translate with Comparison, arXiv:2307.04408, 2023, last accessed 2024/08/20.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., Wang, G., 2023. *Instruction Tuning for Large Language Models: A Survey*. <https://api.semanticscholar.org/CorpusID:261049152>, last accessed 2024/08/23.

## 859 **A Prompt format**

860 Below is the prompt formatting used for the  
861 llama model:

```
862 <|begin_of_text|><|start_header_id|>system<|e
863 nd_header_id|>
864 Given English text, translate it into American
865 Sign Language using Formal SignWriting
866 Notation. Return only the translated Fomal
867 SignWriting, do not provide an explanation.
868 <|eot_id|>
869 <|start_header_id|>user<|end_header_id|>ask
870 <|eot_id|>
871 <|start_header_id|>assistant<|end_header_id|>
872 M521x519S17910453x504S36d00479x481S26
873 502431x505S26502394x505<|eot_id|>
```

874  
875 Below is the prompt formatting used for the  
876 mistral model:

```
877 <s>[INST]Given English text, translate it into
878 American Sign Language using Formal
879 SignWriting Notation. Return only the translated
880 Formal SignWriting, do not provide an
881 explanation.
882 ask[/INST]
883 M521x519S17910453x504S36d00479x481S2650
884 2431x505S26502394x505
885 </s>
```

## 886 **B Few-shot prompt**

887 You are an expert in American Sign Language  
888 Translation. I will require you to carry out  
889 translation from American Sign Language in the  
890 Formal SignWriting notation in ASCII to English  
891 text and will also require you to carry out  
892 translation from English text to American Sign  
893 Language in the Formal SignWriting Notation in  
894 ASCII. Only return the translation. No explanation.

<i>user</i>	<i>assistant</i>
<i>Translate the english text "ask" to Ameri-can Sign Language in the Formal Signwriting Notation.</i>	<i>M521x519S17910453x504S36d00479x481S26502431x505S26502394x505</i>
<i>Translate the Formal SignWriting Notation "M519x515S1a520480x487S10020504x485" in American Sign Language to English Text.</i>	<i>71</i>

Table 4. Manually Curated Few-Shot examples

<i>Error Category</i>	<i>Criteria</i>
Accuracy	Does the translation convey the same meaning as the source text? Are there any omissions, additions, mistranslations or untranslated text?
Fluency	Is the translation grammatically correct and natural in American Sign Language using the Formal Sign Writing Notation? Are transitions between signs smooth and logical? Is there any wrong grammatical register? Are there any internal inconsistencies not related to terminology? Are the characters garbles due to incorrect encoding?
Terminology	Are the specific terms and phrases translated correctly to American Sign Language in Formal Sign Writing Notation? Is the terminology used standard and appropriate for the context? Are terminologies used consistently?
Style	Does the translation fit the style of American Sign Language in Formal Sign Writing? Is the translation consistent with the style and tone of the source text? Is the translation free of awkward phrasing, repetition, and verbosity?
Locale Convention	If translated text does not contain addresses, currency, dates, names or telephone numbers ignore these criteria and return a high score. Otherwise confirm if the translated text conforms with established signs or appropriate fingerspelling for addresses, currency, dates, names, telephone number and time expressions in American Sign Language in Formal Sign Writing Notation? Are these translated in the correct format?

Table 5. G-Eval-MQM Categories and Criteria for SignWriting Evaluation

<i>Error Category</i>	<i>Criteria</i>
Accuracy	Does the translation convey the same meaning as the source text? Are there any omissions, additions, mistranslations or untranslated text?
Fluency	Is the translation grammatically correct and natural in English language? Is there any wrong grammatical register? Are there any internal inconsistencies not related to terminology? Are the characters garbles due to incorrect encoding?
Terminology	Are the specific terms and phrases translated correctly to English? Is the terminology used standard and appropriate for the context? Are terminologies used consistently?
Style	Does the translation fit the style of English language? Is the translation consistent with the style and tone of the source text? Is the translation free of awkward phrasing, repetition, and verbosity?
Locale Convention	If translated text does not contain addresses, currency, dates, names or telephone numbers ignore these criteria and return a high score. Otherwise confirm if the translated text conforms with the correct format for addresses, currency, dates, names, telephone number and time expressions in English, if it is present in the translated text?

Table 6. G-Eval-MQM Categories and Criteria for English Evaluation

<b>Severity</b>	<b>G-Eval Score range</b>	<b>Categories</b>	<b>Weight</b>
Non-Translation	0.2 – 0.0	accuracy	25
		all others	5
Major	0.4 – 0.21	all	5
Minor	0.89 – 0.41	all	1
Neutral	1.0 – 0.9	all	0

Table 7. G-Eval-MQM Scoring model