

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 2050.¹ 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) (Ratsgo et al., 2021). These techniques utilize a multi-step approach to

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

⁷⁷ 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.

⁸⁶ 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 (Ratsgo 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.²

¹¹³ 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

¹²³ fine-tuning, these are the models sizes trainable on more cost-effective commercially available resources.³ To evaluate the results of these experiments, this research will leverage the BLEU score an 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.

¹³⁸ 2 Related Work

¹³⁹ 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.

¹⁴⁹ 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),⁴ 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

²

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

³

<https://medium.com/@ccibeekeoc42/unlocking-low-resource-language-understanding-enhancing-translation-with-llama-3-fine-tuning-df8f1d04d206>

⁴ <http://www.signbank.org/signpuddle>

166 vocabulary sharing achieving BLEU score 22.06 216 this contains spoken language text along with its
 167 on Gloss-to-Text and 45.89 on Text-to-Gloss. 217 accompanying FSW notation for ASL, German
 168 Other research efforts include (Fang et al., 218 Sign Language and others. Table 1., shows the
 169 2024), which developed SignLLM, a 219 details of the dataset used, with the following
 170 comprehensive LLM pipeline, which integrates 220 language pairs; pt-bzs as Brazilian Portuguese &
 171 two transformer approaches, the Multi-Language 221 Brazilian Sign Language, en-ase as American
 172 Switching Framework and Prompt2LangGloss, 222 English & American Sign Language, de-gsg as
 173 further incorporating reinforcement learning 223 Standard German & German Sign Language, fr-fcs
 174 techniques with a Priority Learning Channel. These 224 as Canadian French & Quebec Sign Language, en-
 175 methods are employed for the text-to-pose SLT 225 bfi as British English & British Sign Language, and
 176 task, achieving a 23.25 BLEU-4 score and 49.52 226 ga-isg as Irish & Irish Sign Language.
 177 ROUGE score. Their research also provides a new
 178 dataset for training multilingual sign language
 179 tasks called prompt2sign, containing prompts, text,
 180 video frames and pose data key points. Another
 181 research includes the incorporation of an LLM in
 182 SLR to improve the gloss sentence generation in
 183 video-to-gloss, as demonstrated in (Sincan et al.,
 184 2024). Jung-Ho et al., 2024, also contributed to the
 185 field with their paper on SLT evaluation, where
 186 they utilized LLMs to assess their newly proposed
 187 gloss multi-channel evaluation metric, SignBLEU.
 188 Following the increased application of LLMs on
 189 MT tasks, increasing research into improving the
 190 performance of LLMs on these tasks have been
 191 published. Some of this include the works of (Xu
 192 et al., 2024), which propose two-step fine-tuning.
 193 The first involving fine-tuning of the base model
 194 on monolingual data for all the languages covered
 195 in the task and the second involving SFT,
 196 leveraging high quality parallel data, resulting in
 197 new models referred to as Advanced Language
 198 Models based translators (ALMA). Following their
 199 research, (Xu et al., 2024), propose Contrastive
 200 Preference Optimization (CPO) a policy
 201 optimization approach which leverages preference
 202 data further improving the results of (Xu et al.,
 203 2024) creating new models, ALMA-R, achieving
 204 SOT performance on MT tasks with medium sized
 205 LLMs. Other research into improving LLM
 206 performance in MT tasks include (Moslem and
 207 Haque, 2023; Jiao et al., 2023; Zeng et al., 2024),
 208 however, the research by (Xu et al., 2024) and (Xu
 209 et al., 2024) will adapted and used in the
 210 experimentation for this paper, due to their SOT
 211 performance.

212 3 Methodology

213 3.1 Data Collection and Preprocessing

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

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

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

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

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

258 0.030 and 0.026 for FSW-to-text and text-to-FSW 303 a sliding window attention, GQA with a byte-
259 tasks respectively, leading to further review of 304 fallback BPE tokenizer. The model is also
260 techniques to improve the performance. Of which, 305 pretrained on publicly available data and on
261 further training the model tokenizer with the 306 publicly available chat datasets.⁷ The previous
262 prompts containing both English text and FSW 307 generation of the model has also been used in some
263 notation to allow the tokenizer to better tokenize 308 machine translation research such as (Moslem and
264 the FSW notation for ASL was utilized,⁵ improving 309 Haque, 2023).
265 the model performance at 5000 steps to 0.520 and
266 0.095 respectively.

267 3.2 Model Selection

268 The model selection is constrained by budget 310 limitations, therefore models utilized in these 311 experiments will be limited to open-source models.
269 Additional selection criteria considered, include 312 further training an LLM on an instruction and
270 the model's performance on low-resource 313 output pair. The method improves the model's
271 translation tasks.

272 Based on the LLM rankings in translation task,
273 the pre-trained model ‘‘Meta-Llama-3-8B-Instruct’’
274 was selected for experimentation. This model was
275 selected not only for its performance in SFT, but
276 also for its capability in, human alignment through
277 RLHF and policy optimization.² The model is an
278 auto-regressive decoder only large language model
279 which uses an optimized transformer architecture,
280 with 8 billion parameters, with an 8000 token
281 context window. The model was tuned with SFT
282 and RLHF to improve its alignment with human
283 preferences for helpfulness and safety. The model
284 also utilizes Grouped-Query Attention (GQA) for
285 improved inference scalability. The previous
286 generation of the model has been popularly used in
287 LLM machine translation research, as can be seen
288 in the following papers (Jiao et al., 2023; Zeng et
289 al., 2024; Xu et al., 2024; Xu et al., 2024). The
290 model was pretrained on over 15 trillion tokens of
291 publicly available data and fine-tuned on publicly
292 available instruction dataset and over 10 million
293 human annotated examples.⁶

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

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

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

327 3.3 Experiment Setup

328 This paper explores the performance of different
329 LLM domain adaptation methods in SLT.
330 Comparing the performance of zero-shot
331 prompting on the base model, with few-shot
332 prompting on the base model, with SFT of the base
333 model with parallel data and then CPO (Xu et al.,
334 2024) combined with Simple Preference
335 Optimization (SimPO) (Meng et al., 2024) on the
336 SFT tuned model. Below the steps taken to set up
337 each of these experiments are described, starting
338 with the selection of the prompt to be used.

339 The prompts were engineered by prompting the
340 base models, until settling on an efficient prompt
341 which enables the LLMs return preferable
342 responses to the task without requiring further
343 context. The prompts were then formatted
344 leveraging the advised template from the model
345 documentation pages as can be seen in Appendix
346 A., for the llama⁸ and mistral⁹ models respectively.

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

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

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

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

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

347 **Zero-Shot Learning setup:** In the default setup 396 resource given the lower memory requirements of
348 the base model is tasked with translating samples 397 the mistral model compared to the llama model,
349 for both text-to-FSW and FSW-to-text, without any 398 leading to 49,210 steps to train one epoch. The
350 prior sample input. 399 same lora parameters were used, however only

351 **Few-Shot Learning setup:** An instance of the 400 0.06% of the available parameters were trained.
352 base-model is provided with a prompt and 401 The model was trained on the Google collab
353 randomly selected examples from the training set 402 environment using the T4 compute instance with
354 as can be seen in Appendix B. and is evaluated with 403 13.4 /15 GB of GPU ram used.
355 samples from the test set.

356 **SFT setup:** The models were loaded using 404 **SFT + CPO with SimPO setup:** the SFT tuned
357 Hugging Face transformer library and were fine- 405 models were further fine-tuned on a combination
358 tuned using SFT, chosen for its effective 406 of CPO and SimPO both of which are
359 performance on low-resource datasets ([Medium](#), 407 improvements to the Direct Preference
360 2024). The setup utilizes, Parameter Efficient Fine- 408 Optimization (DPO) technique for (RLHF)
361 Tuning (PEFT) with Quantitized Low-Rank 409 ([Rafailov et al., 2024](#)). The CPO technique
362 Adaptation (QLoRa), leveraging double 410 improves upon SFT by providing a means for a
363 quantization to 4-bit, to improve computational 411 model to learn to reject poor translations, by
364 efficiency during model fine-tuning, maximizing 412 incorporating a BC-Regularizer with the
365 the limited budget ([Dettmers et al. 2023](#)). The 413 approximated uniform reference model from DPO
366 unsloth¹⁰ library is further leveraged to improve 414 to steer the model towards the preferred data
367 model training speed and reduce GPU ram usage, 415 distribution ([Xu et al., 2024](#)), resulting in more
368 improving training speed by 2.2 times and reducing 416 performant and memory efficient training and
369 GPU ram usage by up to 73%.

370 For the meta model, hyperparameter tuning 418 $LCPO(\pi\theta; U) = -E(x, yw, yl) \sim$
371 was carried out over a short training period of 5000 419 $D[\log \sigma(\beta \log \pi\theta(yw|x) - \beta \log \pi\theta(yl|x)) +$
372 steps adjusting the model training parameters 420 $\log \pi\theta(yw|x)]$ (1)

373 based on the performance and finally settling on the 421 The SimPO Technique, improves upon DPO
374 parameters which provided the best model results 422 technique by removing the requirement for a
375 of 0.52 BLUE score for the text-to-FSW task. 423 reference model, using the average log probability
376 These parameters are a learning rate of 2e-4 for 424 of a sequence as the reward, and additionally
377 training stability, the “paged_adamw_8bit” 425 adding a target reward margin to the objective
378 optimizer to leverage the efficiency provided by the 426 maximizing the margin between the preferred
379 quantization, a lora rank of 64 and a lora alpha of 427 response and dis-preferred response, further
380 32 as recommended by ([Dettmers et al. 2023](#)), 428 improving the performance. This technique also
381 resulting in the training of 0.15% of available 429 incorporates length normalization, reducing the
382 parameters and a lora drop out of 0.1 to account for 430 likelihood for the model to generate inaccurate and
383 overfitting to the dataset. The batch size used in 431 lengthy responses ([Meng et al., 2024](#)).

384 training was limited to 1 due to limited memory, 432 $LSimPO(\pi\theta; U) = -E(x, yw, yl) \sim$
385 resulting in 196,800 steps to train one epoch. 433 $D[\log \sigma(\beta|yw|\log \pi\theta(yw|x) -$
386 Validation is carried out every 1000 steps with 434 $\beta|yl|\log \pi\theta(yl|x) - \gamma)]$ (2)

387 10,360 parallel samples in both translation 435 The combination of both CPO and SimPO can
388 directions. The model was trained on the Google 436 be leveraged by incorporating the length
389 collab environment with the L4 compute instance 437 normalization and target reward margin from the
390 with 19.3/22 GB of GPU ram used. 438 SimPO technique within the CPO technique which
391 For the mistral model, to reduce cost overhead, 439 has been explored by ([CPO_SIMPO, 2024](#)).

392 the same parameters used for the llama model was 440 $LCPO - SimPO(\pi\theta; U) = -E(x, yw, yl) \sim$
393 utilized for the mistral model, with a few 441 $D[\log \sigma(\beta|yw|\log \pi\theta(yw|x) -$
394 exceptions such as the batch size, which was 442 $\beta|yl|\log \pi\theta(yl|x) - \gamma) + \alpha \log \pi\theta(yw|x)]$ (3)

¹⁰ <https://github.com/unslotha/unsloth>

443 The hyperparameters used in these experiments, 482 the maximum n-gram order to consider (Papineni
 444 were selected from (Xu et al., 2024; Meng et al., 483 et al., 2002).
 445 2024) and with some modifications added from
 446 (CPO_SIMPO, 2024), these can be seen in Table 2.

Model	Learni 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

447 The CPO-SimPO specific hyperparameters as 448 seen in Table 3. represent, the α , β and γ as seen in
 449 the CPO-SimPO loss function above. Other
 450 hyperparameters not yet listed include, the lora
 451 rank which is 16 across both models, the lora alpha
 452 which is 32 for both models and the optimizer used,
 453 which was the paged_adamw_8bit optimizer in
 454 both models, to avoid out of memory (OOM)
 455 problems. Gradient checkpointing was also
 456 employed in both experiments to reduce memory
 457 requirements with a gradient accumulation step of
 458 8 being selected. Both models were trained over 1
 459 epoch on a dataset of 42k parallel translations in
 460 both directions utilizing the reference translation as
 461 the preferred and the randomly selected data in the
 462 preprocessing steps earlier as the dis-preferred
 463 translation. The models were fine-tuned with a
 464 batch size of 2 and 4 resulting in 2363 steps for the
 465 meta model and 1318 steps for the mistral model.
 466 The models were fine-tuned on a vast.ai
 467 environment with the A100 SXM4 80 GB GPU
 468 with a max usage of 59.2 GB and 70.6 of GPU
 469 memory being observed respectively.

470 3.4 Performance Evaluation Metric

471 Evaluation will be conducted using metrics
 472 previously employed by researchers, specifically
 473 those utilized in Signwriting studies (Jiang et al.,
 474 2023; Moryossef and Jiang, 2024). Of these
 475 metrics the BLUE score and chrF2++ will be used
 476 to evaluate text-to-FSW and FSW-to-text tasks.
 477 With their respective formulae shown below.

$$478 BLEU = BP \cdot \exp(\sum_{n=1}^N w_n \log p_n) \quad (4)$$

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

$$484 ngrF\beta = (1 + \beta^2) \frac{ngrP \cdot ngrR}{\beta^2 \cdot ngrP + ngrR} \quad (5)$$

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

489 The G-Eval framework proposed by Liu et al.,
 490 2023, is used to evaluate the output of the fine-
 491 tuned results from both models with gpt-4o-mini
 492 released by OpenAI¹¹ as the evaluation LLM. The
 493 criterion for evaluation is adapted from Freitag et
 494 al., 2021, which suggests a MQM Framework,
 495 adapted specifically for human evaluation of
 496 translations. This paper adapts the category of the
 497 MQM topology proposed by Freitag et al., 2021,
 498 absorbing subcategories into the main categories
 499 and adapting the criteria to sign language in FSW
 500 notation. Resulting in the categories; accuracy,
 501 fluency, terminology, style and locale convention,
 502 in respective order of topology. The scoring system
 503 is also modified due to the difference is assigned
 504 score from G-Eval implemented using Deepval,¹²
 505 returning scores between 0-1 for each criterion.
 506 The total max score of 25 is retained with, however
 507 instead of non-translation as a category, it is added
 508 as a severity when accuracy score is near zero from
 509 G-Eval. The adapted MQM category topology with
 510 their criteria and adapted scoring model leveraging
 511 the MQM weighting from (Freitag et al., 2021), can
 512 be seen in Appendix C.

513 4 Results and Discussion

514 The different experiments were evaluated using the
 515 BLEU score and chrF2++, Table 3., below shows
 516 the average score of 500 samples taken from the
 517 test dataset. Whilst the G-Eval-MQM metric
 518 described earlier will be carried out for 100 of those
 519 samples, on the fine-tuning experiments. The SFT
 520 experiment was early stopped for both the meta and
 521 the mistral model at 140,000 steps and 29,000
 522 steps, with a validation loss of 0.42 and 0.28
 523 respectively, due to computation resource
 524 constraints, and to facilitate the execution of the
 525 CPO-SimPO experiments in a timely manner.
 526 The CPO-SimPO tuned experiment showcased
 527 the highest performance across both models on the

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

¹² <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

528 Text-to-FSW task, on all metrics, with the meta 559 model on 21 language pair, with a BLEU score of
 529 model showcasing the best performance with 18.07 560 25.0, showcasing a 1.57 score increase, however it
 530 BLEU score, 55.44 chrF2++ score and 14.26 G- 561 falls short on the chrF2++ score, which was 47.0.
 531 Eval-MQM score. Although not at the level of 562 As with the previous task, these results still
 532 previous research leveraging the Sockeye Model, 563 showcase the significant performance gain over the
 533 with a BLEU score of 35.7, however, it showcases 564 base model with an increase of 25.10 and 25.7
 534 a comparable chrF2++ score to 58.4 from (Jiang et 565 BLEU scores over few-shot and zero-shot
 535 al., 2023). These results showcase the possibility of 566 prompting respectively. Possible consideration for
 536 teaching an LLM to carry out SLT with FSW as the 567 improvement of the translation for these
 537 intermediary text. Improving the base model’s 568 experiments is the inclusion of the monolingual
 538 performance on the task to 18.07 BLEU score and 569 data fine-tuning (Xu et al., 2024), as this was
 539 55.44 chrF2++ score from 0.00 and 1.07 on zero- 570 substituted for training the tokenizer on the dataset
 540 shot learning, and 0.00 and 6.86 on few-shot 571 improving the model’s ability to embed FSW.
 541 learning respectively. These results also showcase 572 Across all fine-tuning experiments the mistral
 542 the benefits of RHLF leveraging CPO-SimPO, 573 model fails to show similar improvements as the
 543 improving the performance over SFT by 1.2 BLEU 574 llama model. With a minimal improvement of 0.07
 544 score and 0.07 chrF2++ score, albeit minimal. This 575 BLEU score and 0.04 chRF2++ score over the base
 545 minimal improvement over the SFT experiment 576 model with zero-shot prompting on the FSW-to-
 546 can be attributed to the dataset generation which 577 Text task. Although it showcases a bit better
 547 utilized random data from the same dataset for the 578 improvement in the Text-to-FSW task with an
 548 dis-preferred data as opposed to utilizing outputs 579 improvement of 3.00 BLEU score and 31.46
 549 from the SFT models as the dispreferred data (Xu 580 chrF2++ score. This poor performance could be
 550 et al., 2024).

551 For the FSW-to-Text task, the meta model 585 **5 Conclusion**
 552 performs best, with the SFT experiment 586 In this study, the application of LLMs in SLT was
 553 showcasing a BLEU score of 26.57 slightly over 587 explored utilizing the signbank+ dataset, a
 554 the CPO-SimPo experiment by 0.54, however, on
 555 chrF2++ metric the CPO-SimPO performs slightly
 556 better by 0.35. The G-Eval-MQM score on this task
 557 are 19.18 and 18.46 respectively. The experiment
 558 performs better than Jiang et al., 2023, multilingual

588 multilingual dataset, of which the ASL and English
589 sentence pairs, was selected and prepared.
590 Different LLM training methods were set up as
591 experiments, with the selected models Meta-
592 Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.3,
593 to evaluate the performance of LLMs in SLT tasks,
594 text-to-FSW and FSW-to-text. The combined CPO
595 and SimPO (CPO-SimPO) experiment on SFT
596 tuned meta model resulted in the highest
597 performance for the text-to-FSW task with a BLEU
598 score of 18.07, chrF2++ score of 55.44 and G-Eval-
599 MQM score of 14.26. While on the signwriting-to-
600 text task the SFT experiment and the CPO-SimPO
601 experiments on the meta model, perform similarly
602 with BLEU scores of 26.57 and 26.03, chRF2++
603 scores of 37.30 and 37.38, and G-Eval-MQM score
604 of 19.18 and 18.46 respectively. The results of
605 these experiments are comparable to previous
606 research in SLT leveraging FSW notation as an
607 intermediary translation. These results show the
608 capability to train an LLM to learn the patterns of
609 FSW and map them to English text and vice versa.
610 Showcasing the significant role LLMs can play in
611 improving the performance of multi-step SLT
612 solutions, by leveraging these models to perform
613 the intermediary translation from sign language in
614 FSW to English texts and vice versa. The paper
615 also proposes an adapted evaluation framework
616 combining G-Eval framework and the MQM
617 framework, termed G-Eval-MQM which shows
618 decent correlation with other automatic metrics like
619 BLEU score and chrF2++ score, while adequately
620 evaluating FSW notations leveraging gpt-4o-mini
621 as the evaluation model.

6 Ethical Considerations

623 The following considerations were made during
624 the development of the experiments, to tackle some
625 biases that could be inherent in the data as these
626 were curated from public entries. Different models
627 were selected to address social biases that may
628 have been present in the pretraining data for the
629 models. The models were also selected based on
630 their adherence and transparency on responsible AI
631 as can be seen from llama 3 responsible use
632 guidelines,¹³ and self-reflection guardrails for
633 mistral models¹⁴ and the results of responsible Ai
634 safeguard evaluation for llama models (Wang et al,
635 2024).

7 Limitations

7.1 SignWriting Translation Evaluation

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

7.2 Resource Limitations

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

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

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

683 References

- 684 Bragg D., Koller O., Bellard M., Berke L., Boudreault 737
685 P., Braffort A., Caselli N., Huenerfauth M., Kacorri 738
686 H., Verhoef T., Vogler C., and Ringel Morris M. 740
687 2019. Sign Language Recognition, Generation, and 741
688 Translation: An Interdisciplinary Perspective. In 742
689 *Proceedings of the 21st International ACM 743*
690 *SIGACCESS Conference on Computers and 744*
691 *Accessibility (ASSETS '19)*, pp. 16–31. Association 745
692 for Computing Machinery, New York, NY, USA. 746
693 Dettmers, T., Pagnoni, A., Holtzman, A. and 747
694 Zettlemoyer, L., 2023. QLoRA: Efficient 748
695 Finetuning of Quantized LLMs. In Oh, A., 749
696 Naumann, T., Globerson, A., Saenko, K., Hardt, M. 750
697 and Levine, S. (eds.) *Advances in Neural 751*
698 *Information Processing Systems*, vol. 36, Curran 752
699 Associates, Inc., pp. 10088–10115. 753
700 Fang, S., Wang, L., Zheng, C., Tian, Y., & Chen, C. 754
701 (2024/05/17). SignLLM: Sign Languages 755
702 Production Large Language Models, 756
703 <https://doi.org/10.48550/arXiv.2405.10718>, last 757
704 accessed 2024/05/25. 758
705 Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, 759
706 Q., and Macherey, W., 2021. Experts, Errors, and 760
707 Context: A Large-Scale Study of Human Evaluation 761
708 for Machine Translation. *Transactions of the 762*
709 Association for Computational Linguistics, 9, 763
710 pp.1460–1474. Cambridge, MA: MIT Press. 764
711 Jiang, Z., Moryossef, A., Mülle, M. & Ebling, S.: 765
712 Machine Translation between Spoken Languages 766
713 and Signed Languages Represented in SignWriting. 767
714 In: *Findings of the Association for Computational 768*
715 *Linguistics: EACL 2023*, Dubrovnik, Croatia, p. 769
716 1706–1724. Association for Computational 770
717 Linguistics. 771
718 Jiao, W., Huang, J., Wang, W., He, Z., Liang, T., Wang, 772
719 X., Shi, S., Tu, Z., 2023. ParroT: Translating during 773
720 Chat using Large Language Models tuned with 774
721 Human Translation and Feedback. In: Bouamor, H., 775
722 Pino, J., Bali, K. (Eds.), *Findings of the Association 776*
723 for Computational Linguistics: EMNLP 2023, 777
724 Singapore, pp. 15009–15020. Association for 778
725 Computational Linguistics. 779
726 Jung-Ho, K., John, H.-E. M., Changyong, K., & Hui, 780
727 L. D.: SignBLEU: Automatic Evaluation of Multi- 781
728 channel Sign Language Translation. In: N. 782
729 Calzolari, M.-Y. Kan, V. Hoste, A. Lenc, S. Sakti, & 783
730 N. Xue (Ed.), *Proceedings of the 2024 Joint 784*
731 International Conference on Computational 785
732 Linguistics, Language Resources and Evaluation 786
733 (LREC-COLING 2024), Torino, Italia, pp. 14796– 787
734 14811. ELRA and ICCL. 788
735 Kocmi, T., Zouhar, V., Federmann, C., Post, M., 2024. 789
736 Navigating the Metrics Maze: Reconciling Score 790
737 Magnitudes and Accuracies. In: Ku, L.W., Martins, 738
A., Srikumar, V. (Eds.), *Proceedings of the 62nd 739*
Annual Meeting of the Association for 740
Computational Linguistics (Volume 1: Long 741
Papers), Bangkok, Thailand, pp. 1999–2014. 742
Association for Computational Linguistics.
743 Lee, H., Kim, J.-H., Jun, H. E., Kim, J., & Park, J. C.: 744
745 Leveraging Large Language Models with 746
Vocabulary Sharing For Sign Language Translation. 747
In: 2023 IEEE International Conference on 748
Acoustics, Speech, and Signal Processing 749
Workshops (ICASSPW), Rhodes Island, Greece. 750
IEEE.
751 Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C., 752
2023. G-Eval: NLG Evaluation using GPT-4 with 753
Better Human Alignment. In Bouamor, H., Pino, J. 754
and Bali, K. (eds.) *Proceedings of the 2023 755*
Conference on Empirical Methods in Natural 756
Language Processing, Singapore, pp. 2511–2522. 757
Association for Computational Linguistics.
758 Medium, Difference between Trainer class and 759
SFTTrainer (Supervised Fine tuning trainer) in 760
Hugging 761
Face?, 762
<https://medium.com/@sujathamudadla1213/difference-between-trainer-class-and-sfttrainer-supervised-fine-tuning-trainer-in-hugging-face-d295344d73f7>, 763
last accessed 2024/06/28.
764 Medium, Exploring Mistral 7B for Low-Resource 765
Language Understanding: A Fine-Tuning Approach 766
to 767
Machine 768
Translation. 769
<https://medium.com/@ccibeekeoc42/exploring-mistral-7b-for-low-resource-language-understanding-a-fine-tuning-approach-to-machine-98a0b11fc085>, last accessed 2024/08/20.
770 Meng, Y., Xia, M. and Chen, D., 2024. SimPO: Simple 771
Preference Optimization with a Reference-Free 772
Reward. ArXiv, 773
2405.14734. 774
<https://arxiv.org/abs/2405.14734>, last accessed 775
2024/08/29.
776 Moryossef, A. and Jiang, Z. SignBank+: Preparing a 777
Multilingual Sign Language Dataset for Machine 778
Translation Using Large Language Models, 779
<https://doi.org/10.48550/arXiv.2309.11566>, last 780
accessed 2024/06/03.
781 Moslem, Y., Haque, R., Way, A., 2023. Fine-tuning 782
Large Language Models for Adaptive Machine 783
Translation. <https://arxiv.org/abs/2312.12740>, last 784
accessed 2024/08/20.
785 Papineni, K., Roukos, S., Ward, T. and Zhu, W., 786
2002. Bleu: a Method for Automatic Evaluation of 787
Machine Translation. In P. Isabelle, E. Charniak, 788
and D. Lin, eds. *Proceedings of the 40th Annual 789*
Meeting of the Association for Computational 790
Linguistics, Philadelphia, Pennsylvania, USA pp.
9

- 791 311-318. Association for Computational 846 *Oliver, N., Scarlett, J., Berkenkamp, F. (Eds.),*
 792 Linguistics. 847 *Proceedings of the 41st International Conference on*
 793 Popović, M., Bojar, O., Buck, C., Chatterjee, R., 848 *Machine Learning.* PMLR, pp. 55204-55224.
 794 Federmann, C., Graham, Y., Haddow, B., Huck, M., 849 Zeng, J., Meng, F., Yin, Y. and Zhou, J., 2024. Teaching
 795 Yepes, A.J., Koehn, P. and Kreutzer, J., 2017. 850 Large Language Models to Translate with
 796 chrF++: words helping character n-grams. In 851 Comparison, arXiv:2307.04408, 2023, last accessed
 797 *Proceedings of the Second Conference on Machine* 852 2024/08/20.
 798 *Translation*, Copenhagen, Denmark, pp. 612-618.
 799 Association for Computational Linguistics. 853 Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang,
 800 Post, M., 2018. A Call for Clarity in Reporting BLEU 854 S., Li, J., Hu, R., Zhang, T., Wu, F., Wang, G., 2023.
 801 Scores. In: *O. Bojar, R. Chatterjee, C. Federmann,* 855 *Instruction Tuning for Large Language Models: A*
 802 *M. Fishel, Y. Graham, B. Haddow, M. Huck, A.* 856 *Survey.*
 803 *Jimeno Yepes, P. Koehn, C. Monz, M. Negri, A.* 857 <https://api.semanticscholar.org/CorpusID:2610491>
 804 Névéol, M., Neves, M., Post, L., Specia, M., Turchi and 858 52, last accessed 2024/08/23.
 805 K. Verspoor, eds. *Proceedings of the Third*
 806 *Conference on Machine Translation: Research*
 807 *Papers*, Brussels, Belgium, pp.186-191.
 808 Association for Computational Linguistics.
 809 Rafailov, R., Sharma, A., Mitchell, E., Ermon, S.,
 810 Manning, C.D. and Finn, C., 2023. Direct
 811 Preference Optimization: Your Language Model is
 812 Secretly a Reward Model. ArXiv, abs/2305.18290.
 813 <https://api.semanticscholar.org/CorpusID:2589593>
 814 21, last accessed 2024/08/29.
 815 Rastgoo, R., Kiani, K., Escalera, S. and Sabokrou, M.:
 816 Sign Language Production: A Revie. In: *2021*
 817 *IEEE/CVF Conference on Computer Vision and*
 818 *Pattern Recognition Workshops (CVPRW)*,
 819 Nashville, TN, USA (2021). pp. 3446-3456.
 820 Sincan, O. M., Camgoz, N. C., & Bowden, R,
 821 (2024/03/15). Using an LLM to Turn Sign Spottings
 822 into Spoken Language Sentences,
 823 <https://doi.org/10.48550/arXiv.2403.10434>, last
 824 accessed 2024/05/20.
 825 The joint of Contrastive Preference Optimization
 826 (CPO) & Simple Preference Optimization (SimPO),
 827 https://github.com/fe1ixxu/CPO_SIMPO, last
 828 accessed 2024/08/29.
 829 Wang, Y., Li, H., Han, X., Nakov, P., & Baldwin, T.
 830 (2024). Do-Not-Answer: Evaluating Safeguards in
 831 LLMs. In *Y. Graham & M. Purver (Eds.), Findings*
 832 *of the Association for Computational Linguistics:*
 833 *EACL 2024*, St. Julian's, Malta (pp. 896-911).
 834 Association for Computational Linguistics.
 835 Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H,
 836 2024. A paradigm shift in machine translation:
 837 Boosting translation performance of large language
 838 models,
 839 <https://doi.org/10.48550/arXiv.2309.11674>, last
 840 accessed 2024/08/20.
 841 Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van
 842 Durme, B., Murray, K., Kim, Y.J., 2024. Contrastive
 843 Preference Optimization: Pushing the Boundaries of
 844 LLM Performance in Machine Translation. In:
 845 *Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A.,*

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>	M521x519S17910453x504S36d00479x481S26502431x505S26502394x505
<i>Translate the Formal SignWriting Notation "M519x515S1a520480x487S10020504x485" in American Sign Language to English Text.</i>	71

Table 4. Manually Curated Few-Shot examples

⁸⁹⁵ C G-Eval-MQM Categories and Criteria

Error Category	Criteria
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

Error Category	Criteria
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

Severity	G-Eval Score range	Categories	Weight
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