Large Language Models as *Sign Language Interfaces*: Mitigating the Requests of Deaf Users of LLMs in a Hearing-Centric World

Anonymous ACL submission

Abstract

Deaf or Hard-of-Hearing (DHH) individuals use Large Language Models (LLMs) in unique ways and request to incorporate sign language grammar and Deaf culture during the training of these models, in addition to video-based sign language input capabilities. Yet, developers of instruct-tuned LLMs have not paid attention to these requests. Instead, special translation models are developed for sign languages (SLs), diminishing the needs of signers to a simple lack of communication between hearing and Deaf communities. In this paper, we take an orthogonal approach to these traditional methods of studying SLs. To meet the requests of Deaf users of LLMs, we look at the sign language processing (SLP) algorithm from a theoretical lens, then introduce the first text-based and multimodal LLMs. We propose new prompting and fine-tuning strategies for text-based and multimodal SLP, incorporating sign linguistic rules and conventions. We test the generalization of these models to other SLP tasks, showing LLMs can process signs while still being adept at spoken language tasks. Our code and model checkpoints will be open-source. We will update our model suite as newer open-source LLMs, datasets, and SLP tasks become available.

1 Introduction

Most Deaf and Hard-of-Hearing individuals have well-developed methods to navigate a hearing-centric world, and they adapt these strategies to newly emerging technologies in unique ways (Desai et al., 2024). With the recent prevalence of public-facing LLMs, DHH users voice their distinct challenges using these systems–such as confronting misconceptions and biases in English language use (Swisher, 1989), and Information Deprivation Trauma (Schild and Dalenberg, 2012)–that are mostly left unaddressed by developers of LLMs. According to a recent seminal study by Huffman et al. (2024), 44.1% of Deaf or Hard of Hearing (DHH) individuals who use LLMs say that they have challenges in asking questions to LLMs, and 22.1% are unsatisfied due to limited sign language support in LLMs. In this paper, we acknowledge these concerns and address them by first introducing a theoretical perspective to the sign language processing algorithm, and introducing a family of fine-tuned LLMs (both text-based and video-based models) that are grounded in sign language linguistics.

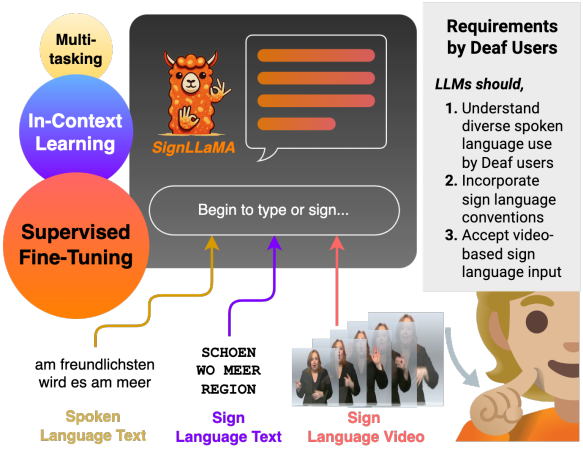

Figure 1: Deaf users have specific requests pertaining to the development of LLMs, as shown above. We show that text-based and multimodal open LLMs when prompted or fine-tuned, can learn to perform sign language processing tasks, and these requests can be mitigated. Further, multitasking fine-tuning on both spoken (OpenOrca) and signed (PHOENIX-14T) corpora alleviates forgetting of spoken language capabilities (e.g., QA tasks in English).

In more detail, our contributions are described as follows.

1. With a user-centered design approach for DHH (Potter et al., 2014), we identify the needs and requirements of Deaf users of LLMs and identify ways of incorporating sign language knowledge into them.

2. We theoretically and empirically study the problem of catastrophic forgetting during fine-tuning on sign language data, providing solutions to resolve this issue.
3. We introduce multimodal and text-based LLMs fine-tuned on SLP tasks and analyze in detail whether they satisfy the requirements set by signers.

Our results show that fine-tuning large, pre-trained models offers new generalization capabilities compared to previous sign recognition training strategies, e.g., via in-context learning. All code, data, and model checkpoints will be publicly available and will be regularly updated to reflect new developments in LLMs and SLP data & tasks.

2 Needs & Requests of the Deaf Community

From the personal interviews presented in Huffman et al. (2024), there are three major areas that DHH would like to see improvements with LLMs: ***1) LLMs should understand diverse spoken language use by the Deaf***, ***2) LLMs should have a deep understanding of the DHH community***, and ***3) LLMs should accept visual sign language as input***. Essentially, signers want LLMs to understand SL grammar order, or at least the gloss notation –an intermediary textual representation for signs–, (e.g., most of the time when Deaf individuals write in American Sign Language, hearing people can misunderstand it as "Deaf" English). Furthermore, signers want sign language-specific datasets to be used in the training of the LLMs. Also, video-based sign understanding is requested to be able to input a model in sign languages.

Here, it is necessary to distinguish reading and writing in spoken *versus* sign languages. Most bilingual signers default to reading and writing in spoken languages or modified versions of them instead of SLs while interacting with LLMs due to lack of effective interfaces (Desai et al., 2024; Inan et al., 2024; Bragg et al., 2020; Glasser et al., 2020; Hariharan et al., 2018). We are specifically interested in the problem of interfacing with signers using text-based or multimodal LLMs, which helps signers to read and write in SLs while also enhancing their reading and writing capabilities in spoken languages (Samuel J. Supalla, 2021). As a concrete example, we are aiming to create an LLM that DGS-bilingual signers use to converse with using text, or videos, instead of using German to chat with LLMs.

These concrete requirements by DHH motivate our work and open up the following several important scientific questions and research areas:

1. *How can LLMs understand signers better?*
2. *What are some possible ways of including Deaf knowledge and contexts into LLMs?*
3. *Does in-context learning or supervised fine-tuning make LLMs more capable of understanding Deaf culture and signing?*
4. *Does pretraining with sign language knowledge affect spoken language capabilities of LLMs?*
5. *Can these effects be mitigated in post hoc model training?*

To answer these questions in more detail, inspired from all of the prior work, we first look at the problem from a theoretical lens, and then we apply large pre-trained language models to tasks in SLP. To represent SLs in a textual environment, we experiment with glosses (intermediary textual representations of signs), which are also found to be helpful with the spoken language reading skills of signers (Luft, 2023a; Supalla, 2017)[1]. For the visual modality of SLs, we use LLaVA-based models. This allows us to cover all modalities signers use as input to an LLM.

Our results point to a future where language models can also be pre-trained on SLs *without significant degradation of their spoken language capabilities*, marking an essential step for the wider adoption of SLs into LLM pipelines. This has broader implications for creating LLM-based tools that meet the requests of signers in a hearing-centric world.

3 A Theoretical Sign Language Learning Algorithm

Many current proprietary or open-source LLMs do not consider sign language data during their training process (e.g., due to lack of signers or expertise in Deaf culture). This is also noticed by Deaf users and is requested to be mediated in (Huffman et al., 2024). We believe this lack of accessibility can be mitigated in two ways: 1) including SL-specific data in pretraining or 2) using techniques such

[1]Even though the Sign Language Translation community does not recommended using glosses for model development as it can lead to information loss, pedagogical literature in SL suggests using glosses as an interface for signers (Heather Gibson, 2021) is advantageous. For further discussion of the limitations of glosses, please refer to §9, and (Müller et al., 2023))

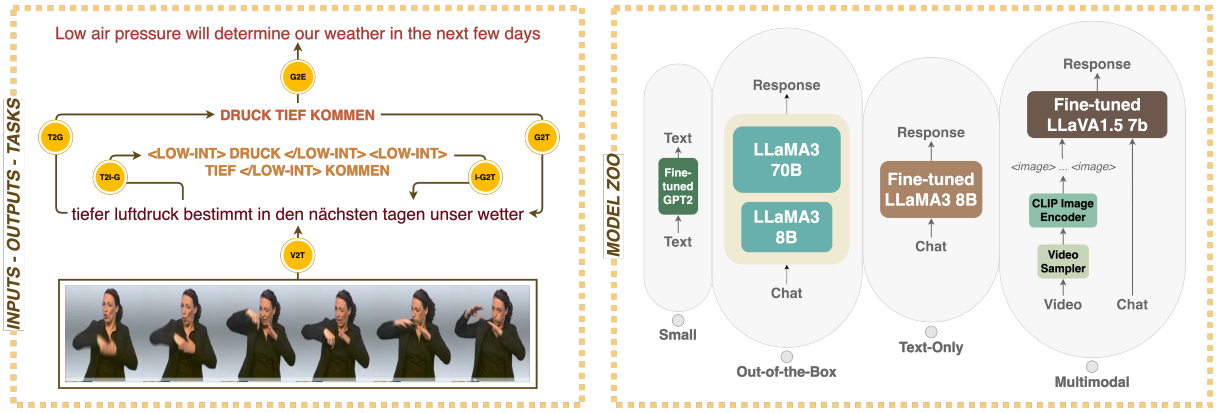

Figure 2: This figure presents a summary of all the inputs, outputs, tasks, and models we are using and introducing in this paper. The box on the left contains a sample from the RWTH-PHOENIX-14T dataset. From top to bottom, the sentences are English text, DGS glosses, intensified DGS glosses, and German text. Yellow knobs represent tasks, in which the acronyms of the tasks are inlaid (please refer to Section §4 for detailed task names).

as prompt-tuning or fine-tuning with various SLP tasks. We present these two ways as an SLP learning algorithm, for the first time from a theoretical perspective, to incorporate sign-language-specific information into LLMs.

Specifically, in the case of LLMs learning how to process SLs, this algorithm contains two specific steps:

1. **Pre-Training:** LLMs are trained on multiple tasks that do not include (many or any) sign-language-specific tasks. This step is not unique to SLP and is common for any state-of-the-art LLM that is used by both signed and spoken language users. Using the terminology of Sicilia and Alikhani (2022), this process picks weights to minimize a *test divergence* or "error" $\mathbf{TD}_{PT}$ where PT is the pre-training data distribution:

$$\mathbf{TD}_{PT}(\theta) = \mathbf{E}[|\ell(D, \hat{D})|] \\ D \sim \text{LM}(X;\theta), \hat{D} \sim \text{ANOT}(X) \quad (1)$$

   where LM is the language model, ANOT is a human annotation provided the same context X (e.g., a prompt), and X ranges over the dataset PT. The test ℓ compares any measure of the quality or other properties of the generated text between the LLM and the human; e.g., BLEU, ROUGE as well as human preference scores. Even though we advocate for an accessible training setup for LLMs, it is not yet feasible to require all proprietary and open-source models to have SL data during pretraining for this step.
2. **Fine-Tuning:** Only in this latter stage, an LLM can be fine-tuned on SLP tasks such as translation. For the *sign-only fine-tuning*, we call this data distribution DGS. So, abstractly, our sign-only fine-tuning process described previously attempts to minimize $\mathbf{TD}_{DGS}(\theta)$.

Given this learning setup, we follow up empirically with experiments using multiple strategies to inject SL knowledge into LMs (e.g. in-context learning, supervised fine-tuning) during pre-training and fine-tuning stages. We also experiment with and discuss the consequences of introducing SL-specific data during fine-tuning with a spoken-language pre-trained model and how they can be mitigated.

4 Methods

In this section, we introduce the details of the data, tasks, and the text-based and multimodal LLMs we use in the experiments (see Figure 2).

DGS Data Due to widespread adoption as a benchmark in the SLP community, we use the RWTH-PHOENIX-14T[2] corpus of weather forecast signs in German Sign Language (DGS). This dataset contains around 7000 training samples, 500 validation samples, and 600 test samples. Each sample has a video, a text in spoken German, and a gloss – which is an intermediary textual representation of signs – in German Sign Language. Video samples consist of frames of multiple signers sampled at 25 fps, with a size of 210 by 260 pixels. We also include an enhanced version of this dataset, which contains *intensifier* information in its gloss representations as introduced by (Inan et al., 2022). Intensifiers in SLs are depicted through non-manual markers and can change the

[2] https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/

meaning of a sign, and this dataset contains additional tokens to capture intensifier information. We also translate the German text to English text to provide data for a cross-lingual task (discussed next). We use Google Translate.[3]

Tasks As RWTH-PHOENIX-14T is a parallel corpus between spoken German and DGS, most previous research has focused on translation tasks between these languages. In this paper, we focus on translating DGS to German (broadly considered as a sign understanding or recognition task) and German to DGS (broadly considered as sign generation). In addition to these, we introduce additional tasks to test generalization. Specifically, we consider:

- **(G2T)** ***DGS Gloss to German Text*****:** a text-based translation task from textual intermediary representations of DGS (glosses) to German text.
- **(T2G)** ***German Text to DGS Gloss*****:** the inverse problem of the above and is text-based.
- **(V2T)** ***DGS Videos to German Text*****:** a multimodal task where the input is a video of a signer signing in DGS, and the output is German text.
- **(I-G2T)** ***Intensified DGS Gloss to German Text*****:** a text-based task with augmented DGS tokens. Additional symbols <HIGH-INT> and <LOW-INT> are wrapped around glosses to depict intensity in the video that is not depicted in traditional gloss representations (Inan et al., 2022).
- **(T2I-G)** ***German Text to Intensified DGS Gloss*****:** the inverse problem of (I-G2T), still text-based.
- **(G2E)** ***DGS Gloss to English Text*****:** a novel task of cross-modal translation, where DGS glosses from the German Sign Language family are translated to English text from the spoken Indo-European language family. Without any pretraining, this is a difficult test of generalization and composition of contextualized meanings across traditional and SLs.

To test generalizability and in-context learning, G2T is the only DGS task we use for any fine-tuning (see § 5.2). All the other tasks are used to evaluate the models' performance.

Models In this paper, we use two main foundation models: LLaMA-3 8B Chat (Touvron et al., 2023b) for text-based inputs and LLaVA 1.5 7B (Liu et al., 2023a,b) for multimodal inputs. To compare with traditional SLP approaches, which use smaller language models *sans any foundational pre-training*, we also use a randomly initialized GPT-2 model (Radford et al., 2019) trained on the G2T task of the RWTH-PHOENIX-14T dataset. This controlled difference allows us to quantify the utility of concepts learned during foundational training (e.g., in LLama and LLaVA) on SLP. Lastly, for G2T task, we use LLaMA-3 70B with 4-bit quantization[4] to show how the number of parameters affects the results.

5 Experiments

5.1 In-Context Learning

Our initial set of experiments test whether SL-specific information can be included in LLMs using in-context learning. For this, we prompt language models using linguistic and cognitive science rules of glossing and signing. To evaluate their performance, we use the tasks described in § 4. We incorporate the following linguistic rules of SLs into the design of the prompts that we provide to the models:

- **zero-shot prompt:** The prompt is structured as, "This is a sentence in German Sign Language glosses: <glosses>. You MUST translate these to spoken German. You MUST give the answer directly without any other text." It does not contain any linguistic rules.
- **rule-based prompt:** The prompt is structured as five rules of glossing semantics. These rules are described in (Hanke et al., 2020).
- **notation prompt:** This is structured as a set of rules about gloss morphologies. These rules are borrowed from Stein et al. (2010).
- **one-shot prompt:** This prompt gives a single example of a DGS gloss and a corresponding German text. This example is formatted following the semantic and morphological rules above.

All prompts are given in Appendix B.

For the multimodal foundation model, we provide a single chat template. We use a mixed prompting strategy, where the video of signers is sampled at 50 frame intervals, fed into a CLIP-based Image Encoder (Radford et al., 2019), and then incorporated into the prompt tokenization by the use of <*image*> for each frame. Then, the image portion of the prompt is succeeded by the text-based prompt *"This video is in German Sign Language. What is the sentence being signed in German?"*

[3] https://cloud.google.com/translate/

[4] https://ollama.com/library/llama2:70b

5.2 Supervised Fine-Tuning with LoRA

Besides in-context learning via few-shot prompts, we also consider fine-tuning LLaMA3 and LLaVA1.5 models using Supervised Fine-Tuning [5], which is a supervised training method in addition to the RLHF algorithm (Ouyang et al., 2022) for chat-based model training, which aligns the models' representations with human judgments. In this case, the human annotations are either glosses or text. For fast model training and reduced memory consumption, we use Low-Rank Adaptation of Language Models (LoRA) as introduced by Hu et al. (2022). We give details of model hyperparameters and training details in Appendix A.

Sign-Only Fine-Tuning As noted, for text-based models we fine-tune on the G2T task from § 4, and for multimodal we fine-tune on the V2T task. This provides the model a simple introduction to the meaning of signed glosses by grounding them to their parallel German language context. We discuss the results of these experiments, in detail, in § 6.

5.3 Multi-Tasking Mitigates Forgetting

As the last set of experiments, we conduct theoretical analyses and describe their empirical implications. We hypothesize that the sign-only tuning strategy can lead to catastrophic forgetting. Due to the shared token vocabulary, the model may overwrite existing knowledge and semantics in the contextualized representations of traditional language tokens. However, signers want SL information incorporated into the models. Then how can we realize that *post hoc*, without forgetting the spoken language tasks? Intuitively, we expect that forcing the model to "replay" spoken language tasks from pre-training will prevent forgetting.

Motivated by neuroscience, *experience replay* has been suggested as a strategy to reduce forgetting in machine learning, with positive results (Rolnick et al., 2019). Moreover, replay has been studied in mathematical theories of how language models learn with similar success (Sicilia and Alikhani, 2022).

We re-frame our learning environment given in § 3 to motivate our hypothesis. Namely, we show that multi-task fine-tuning (i.e., replay) can help mitigate forgetting in shared-vocabulary sign processing with LLMs.

Problem When we write out the pre-training and fine-tuning objectives in 1, it is clear that the two processes optimize *different* objectives (e.g., over different datasets). There is no way to ensure that picking θ to minimize $\mathbf{TD}_{DGS}$ will not have a negative impact (i.e., increase) $\mathbf{TD}_{PT}$. This potential for increase in error on the pre-training tasks characterizes the behavior we call "forgetting."

Solution As mentioned, we consider a *multi-tasking fine-tuning* strategy where *DGS* data and tasks similar to the pre-training data are mixed. This multi-tasking data can be represented by a mixture distribution:

$$\text{MIX} = \alpha\ \text{PT} + (1-\alpha)\ \text{FT} \tag{2}$$

where $\alpha \in (0, 1)$ is a weighing factor between the probabilities assigned by two datasets. Instead of sampling X from only PT or only FT, we flip an α-weighted coin to pick from which we sample. Holding all else constant, this implies the equality:

$$\text{TD}_{MIX} = \alpha\ \text{TD}_{PT} + (1-\alpha)\ \text{TD}_{FT}. \tag{3}$$

By this choice, we can see:

$$|\text{TD}_{MIX} - \text{TD}_{PT}| \tag{4}$$
$$= (1-\alpha)|\text{TD}_{FT} - \text{TD}_{PT}| \tag{5}$$
$$< |\text{TD}_{FT} - \text{TD}_{PT}|. \tag{6}$$

Since TD_{MIX} is always closer in magnitude to TD_{PT} than TD_{FT}, we can see that minimizing TD_{MIX} can better prevent large increases TD_{PT}, or "forgetting." This simple inequality provides a theoretical motivation for our multi-tasking suggestion in § 5.2. Our empirical results in § 6 also confirm our theoretical hypotheses.

Implementation To test the implications of this theoretical analysis, in practice, we also train on an additional dataset (OpenOrca[6]) randomly mixing the sign and traditional data during tuning. This dataset consists of system prompts, questions, and responses, augmented from the FLAN collection (Longpre et al., 2023). Our multi-tasking strategy can be viewed as a type of experience replay since many tasks from OpenOrca are presumed to be similar to prior experience during pre-training.[7] It is commonly used to fine-tune smaller open models such as LLaMA for better task success, surpassing proprietary models such as GPT-3.5. The

[5] https://huggingface.co/docs/trl/main/en/sft_trainer

[6] https://huggingface.co/datasets/Open-Orca/OpenOrca

[7] Most open-source models do not share training data.

dataset is mainly in English and consists of multiple tasks: entailment and semantic understanding, temporal and spatial reasoning, causal judgment, multilingual understanding, world knowledge, logical and geometric reasoning, and similar other tasks (Mukherjee et al., 2023). While the original dataset contains around 3 million samples, we use the same split sizes as RWTH-PHOENIX-14T to ensure balance in sign/traditional task prioritization.

6 Findings

In this section, we present our results and discuss our findings under five research questions. We outline all of these questions in the following sections and give answers to them with our findings. For further discussion of these findings and their position in the SLP research literature, please refer to Appendix § F.

6.1 Automatic Metrics

For all the tasks, to compare the generated text with the ground truth, we make use of automatic metrics. We use both traditional n-gram metrics of BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and also use learned generation metrics such as BERTScore (Zhang* et al., 2020). To implement all of these, we use the Huggingface evaluate library[8]. We do not include classification-based metrics, as our language models generate full-textual responses rather than classes.

Prompt Strategy	BLEU_1	ROUGE_1	BS-F1
zero-shot prompt	24.5	0.277	0.841
rule-based prompt	22.8	0.255	0.836
notation prompt	24.3	0.277	0.840
one-shot prompt	27.1	0.309	0.851

Table 1: Preliminary evaluation of prompting strategies on the validation set of RWTH-PHOENIX-14T using LLaMA-3 8B. The prompts are given in Appendix § B. BS-F1 refers to BERTScore-F1.

How do different prompting strategies affect the performance? When we look at the results of the in-context learning experiments evaluated using the non-finetuned LLaMA-3 8B model in Table 1, we see that an in-context example of sign knowledge performs best. These show that rule-based prompts and notation-based prompts perform similarly to or less than zero-shot prompts. This is an insightful finding, pointing to the realization that just providing sign language grammar rules is not necessarily enough to teach the model to understand better sign language, but an example can be more effective. This influences the designing of off-the-shelf LLM-based systems for the use of the DHH community.

Gloss to Text Translation (G2T)				
Models	*TEST SET*			
	$\text{B}_1 \uparrow$	$\text{B}_2 \uparrow$	$\text{R}_{LSum} \uparrow$	$\text{BS}_{F1} \uparrow$
one-shot GPT2	3.14	0.04	0.067	0.798
ft-LLaMA3 8B	**27.1**	**11.4**	0.275	**0.851**
multi-LLaMA3 8B	22.7	9.46	**0.294**	0.851

Table 2: This table shows the comparison of small fine-tuned models with Large Language Models and multitasking Large Language Models. It can be seen that the performance of the larger LLaMA-based models is higher overall compared to a smaller model (GPT2). Also, multitasking to prevent forgetting does not affect model performance.

How does supervised fine-tuning LLMs affect the performance compared to a fine-tuned small model? Inherently, most of the SLT models use small transformer-based architectures[9], and it is important to investigate whether these models are still viable given the presence of LLMs. To understand the performance difference, we present results comparing the baseline of a small GPT-2 model fine-tuned on the G2T task with our larger models LLaMA-3 8B and Multitasking LLaMA-3 8B in Table 2. As is evident from the scores, LLaMA-3 outperforms fine-tuned GPT2 by a large margin. This implies that using larger models as backbones for SLP instead of smaller transformer-based models is an encouraging future direction, as they contain more pretrained semantic information that is helpful in sign language tasks as well.

How does the fine-tuned video-based model perform compared to a text-based model? We show the performance differences between fine-tuned and non-finetuned video-based LLMs in Table 4. Here, unsurprisingly, the fine-tuned model is performing the best across all metrics. Yet, if we compare results in Table 2 and 4, the video-based model is performing lower in a task that has

[8] https://huggingface.co/docs/evaluate/

[9] some newer models exist that use LLMs such as (Wong et al., 2024) and (Fang et al., 2024). Still, these are yet to be accepted by the community as stable models.

Multimodal Sign Understanding (SignVideo2Text)				
Models	*TEST SET*			
	$B_1 \uparrow$	$B_2 \uparrow$	$R_{LSum} \uparrow$	$BS_{F1} \uparrow$
LLaVA1.5 7B	2.140	0.006	0.022	0.658
ft-LLaVA1.5 7B	**12.776**	**2.404**	**0.103**	**0.779**

Table 4: This table shows the automatic metric results for the translation task of German Sign Language video to German Text. ft-LLaVA1.5 7B is the fine-tuned model.

the same output. This shows that Deaf users' requests for video input capabilities are not yet met and may require better modality modeling efforts. The main bottleneck of improving video LLMs in the task of sign understanding is the lack of high-quality data. However, human annotations for sign language glosses can also be costly to collect. We discuss more on this matter in the Appendix section §E. The implications of using videos rather than glosses mean that in the absence of signer annotations on the glosses, videos can be used as input as well, with a decrease in the overall performance. This opens the possibility of partially satisfying the requests of the signers.

Given the theoretical background of forgetting, how does including multiple tasks during fine-tuning affect spoken-language performance? To answer this question, we use generic spoken language benchmarks by EleutherAI Evaluation Harness (Gao et al., 2023) and test the performance difference between the multitasking, fine-tuned, and non-finetuned models. We show the results in the bar plot in Figure 3. We can empirically observe that there is a drop in performance between non-finetuned and fine-tuned LLaMA3 models. This shows the data shift that we have outlined in Section §3 due to the differences in data distribution between the pretrained LLaMA3 and the sign-finetuned LLaMA3. This strongly suggests that there is forgetting of the original capabilities of the pretrained model. This verifies our theoretical hypotheses, and the increase in performance during multitasking suggests that signed and spoken languages can be introduced to models *post hoc* with minor forgetting of the original spoken language tasks.

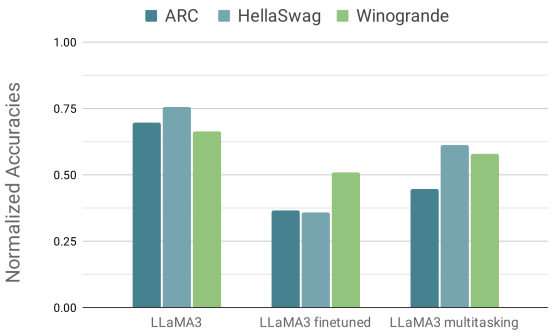

Figure 3: This is the bar plot showing the ablation study on the multitasking/mixing model on the Open Language Model Benchmarks of ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and WinoGrande (Sakaguchi et al., 2019), all degrade (forgetting) when LLaMA3 is fine-tuned on the sign language tasks, and when trained on multiple tasks, it performs better.

Can the performance in G2T generalize to other SLP tasks? To answer, we show the results for all the sign language tasks in Table 3. Based on the BLEU scores, the lowest-performing task is T2G (the reverse of G2T, the task on which the model was fine-tuned), and the best-performing task is

Performance of All Models on All Tasks										
Task	**Prompt Strategy**	Finetuned GPT2			Not Finetuned LLaMA3 8B			Multitasking LLaMA3 8B		
		B_1	R_{LSum}	BS_{F1}	B_1	R_{LSum}	BS_{F1}	B_1	R_{LSum}	BS_{F1}
T2G	one-shot	1.419	0.027	0.798	8.556	0.127	**0.818**	**10.921**	**0.165**	0.794
T2G	zero-shot	1.879	0.030	0.810	8.335	0.122	**0.802**	**10.485**	**0.161**	0.794
G2E	one-shot	3.604	0.066	0.822	**9.226**	0.084	0.807	3.104	0.034	0.828
G2E	zero-shot	3.931	0.056	0.808	**12.369**	0.103	0.816	5.442	0.064	0.83
I-G2T	one-shot	2.242	0.048	0.791	9.573	0.111	0.691	**17.637**	0.155	0.524
I-G2T	zero-shot	1.642	0.043	0.768	11.589	0.143	0.769	**21.157**	0.279	0.845
T2I-G	one-shot	1.305	0.054	0.815	42.277	0.576	0.897	**43.636**	0.156	0.778
T2I-G	zero-shot	0.050	0.062	0.802	**56.128**	0.704	0.910	43.229	0.155	0.778

Table 3: This table shows the performance of all the models for all the tasks that we introduce in Section §4 for the test set. The one-shot strategy contains an example for the task. B_1 corresponds to BLEU-1, R_{LSum} corresponds to ROUGE, and BS_{F1} corresponds to BERTScore.

T2I-G. It can be seen that, to a certain degree, there is some generalizability to different tasks, but most tasks do not reach the same level of performance as 27.1 in the G2T task (Table 2). Curiously, T2I-G performs much better than the G2T task, which may indicate the importance of prosody and how LLMs can recognize intensifications better than they can generate translations directly. Another interesting observation is that the multitasking model performs better in all tasks except G2E than the non-finetuned model. This shows that forgetting of SLP and spoken language tasks is mitigated mostly, but sometimes forgetting may still occur. All in all, this analysis shows us that fine-tuning LLMs on an SLP task leads to better measurable performance and some generalization in similar SLP tasks. This is an encouraging result showing that the requests of signers can be satisfied by including sign language tasks in the fine-tuning stage without losing measurable performance on SLP or spoken language tasks.

7 Related Work

Besides text-based models like LLaMA (Touvron et al., 2023a), Mixtral (Jiang et al., 2024), QWEN (Bai et al., 2023), Orca (Mukherjee et al., 2023), Phi (Gunasekar et al., 2023), multimodal models have been gaining popularity, especially in computer vision communities. Large Vision-Language models such as LLaVA (Liu et al., 2023c), Video-LLaMA (Zhang et al., 2023), Video-LLaVA (Lin et al., 2023), LanguageBind (Zhu et al., 2024), MultiModal-GPT (Gong et al., 2023), Mirasol3B (Piergiovanni et al., 2023), LAVIS (Li et al., 2023), LaViLa (Zhao et al., 2023), and UniVL (Luo et al., 2020) propose to align representations of combinations of images, videos, text, and/or speech signals with human judgments. Further details of these and similar models have been discussed in a survey paper by Yin et al. (2023). However, none of these models claim to include SLP tasks in their pre-training or fine-tuning data. Through our theoretical and empirical studies, this paper aims to address this gap.

The absence of literature using large models for SLP is mainly due to the low-resource nature of SLs (Yin et al., 2021). However, there have been several lines of research applying transformer-based language models to sign language translation (Camgoz et al., 2018; Yin and Read, 2020; Chen et al., 2023b), sign language understanding (Hu et al., 2023; Moryossef et al., 2021), sign generation (Stoll et al., 2020), SignWriting translation (Jiang et al., 2023), incorporating facial expressions (Viegas et al., 2023), modeling prosody (Inan et al., 2022), and sign language segmentation (Moryossef et al., 2023). Lee et al. provides an early work that leverages (smaller, but still large) language models with shared vocabularies for SLP. They focus on older models (without RLHF, Ouyang et al., 2022). Further, Gong et al. (2024); Wong et al. (2024) give a more recent application of LLMs as part of a translation pipeline, and Fang et al. (2024) fine-tunes diffusion-based LLMs for sign avatar generation. However, none involves instruct-tuning large language models (text-based or multimodal) with both spoken and signed capabilities, which we introduce in this paper for the first time.

In addition to the SLP and LLM literature, SL education works are important for this work. In the SL pedagogy literature, some works focus on case studies of gloss-based intermediary textual constructs as ways of ASL to English literacy (Cripps et al., 2020), a formal distinction between sign and spoken language reading (Supalla, 2017), and reading assessments for DHH signers (Luft, 2023b). These works have influenced our choice of glosses as intermediary representations for text-based LLMs. We believe that text-based and video-based language models can be helpful as reading and writing companions that use glosses or videos to interface with signers.

8 Conclusion

In this paper, we have prompted, fine-tuned, and compared text-only and multimodal language models for sign language processing tasks, as requested by Deaf users. We have provided theoretical grounding and analyzed our results with implications on how much LLMs can meet the needs of signers without losing capabilities in spoken languages. From our findings, it can be claimed that LLMs can be fine-tuned to SLs, and in-context learning can help to create an off-the-shelf LLM tailored towards the Deaf and Hard-of-Hearing community, which can be accomplished without forgetting spoken language capabilities.

Moving forward, training bigger models with larger multilingual corpora is a promising next step for a broader set of novel sign language processing tasks. We will make our code, data, and model weights publicly available upon acceptance.

9 Limitations

The major limitation of our work has been the computing power required to fine-tune, test, and carry out inference. Even with the smallest large language models, it becomes quickly infeasible to test multiple independent variables. Hence, our techniques have been tested on the smaller end of the large language family of models. Larger models can have higher performance gains.

An additional limitation of our models is the context length. With long linguistic rules added to the prompt, certain samples of glosses made the inference lengthy. The maximum number of generated tokens has been a limiting factor of the output of models as well, which resulted in poor performance metrics. These can be alleviated with higher computing powers.

Another major limitation is the dataset size and number of available tasks in SLP. The SLP community has focused on translation tasks so far, and not many other task definitions and datasets exist that can be useful for signers. This affects our benchmarking, as the only tasks we can test the generalization on are either other translation tasks or traditional NLP tasks that are non-specific to SLs. Having diverse tasks and accompanying datasets is needed for the future of SLP.

Certain other SL datasets exist, such as How2Sign (Duarte et al., 2021), CSLDaily (Zhou et al., 2021), and BOBSL (Albanie et al.). These datasets are larger and have diverse domains compared to RWTH-PHOENIX-14T that we have used in this work. The main reason that we chose to focus on RWTH-PHOENIX-14T is because the glosses in it are annotated manually by signers while in other datasets automated ways are used or glosses are not available. Glossing is a core part of our paper, as we are focusing on new ways of interfacing with signers using LLMs instead of just translation. This currently can be accomplished by reading and writing in glosses.

10 Ethical Statement

We are using LLaMA3-based models for both our text-only and multimodal setups, which are trained on data acquired by Meta and are not made publicly available; even though the model itself is open-source, the pretraining dataset is not open. This leads to unaccountable biases that have been collected during the dataset formation and in the pretraining, our models may have inherent biases passed down from these pretraining setups. Our RWTH-PHOENIX-14-T dataset contains the faces of the signers, which is a piece of private information. This private information is used in accordance with the original dataset creator's directions and privacy concerns. Furthermore, sign language processing can be a sensitive topic, especially when the community-centric approach is not taken for the design of systems. For this, we collaborate with the Deaf and Hard-of-Hearing communities or signers in general while developing such systems as this one.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Danielle Bragg, Meredith Ringel Morris, Christian Vogler, Raja Kushalnagar, Matt Huenerfauth, and Hernisa Kacorri. 2020. Sign language interfaces: Discussing the field's biggest challenges. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–5.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Emanuela Campisi, Anita Slonimska, and Asli Özyürek. 2023. Cross-linguistic differences in the use of iconicity as a communicative strategy. In *the 8th Gesture and Speech in Interaction (GESPIN 2023)*.

Xuanyi Chen, Junfei Hu, Falk Huettig, and Asli Özyürek. 2023a. The effect of iconic gestures on linguistic prediction in Mandarin Chinese: a. [Online; accessed 14. Feb. 2024].

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2023b. Two-stream network for sign language recognition and translation. *Preprint*, arXiv:2211.01367.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Jody H. Cripps, Samuel J. Supalla, and Laura A. Blackburn. 2020. A Case Study on Accessible Reading with Deaf Children. *ODU Digital Commons*, 4(1).

Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 54–65, Torino, Italia. ELRA and ICCL.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sen Fang, Lei Wang, Ce Zheng, Yapeng Tian, and Chen Chen. 2024. Signllm: Sign languages production large language models. *Preprint*, arXiv:2405.10718.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. 2012. RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3785–3789, Istanbul, Turkey. European Language Resources Association (ELRA).

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Abraham Glasser, Vaishnavi Mande, and Matt Huenerfauth. 2020. Accessibility for deaf and hard of hearing users: Sign language conversational user interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pages 1–3.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. *Preprint*, arXiv:2404.00925.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *Preprint*, arXiv:2305.04790.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. *arXiv*.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Dhananjai Hariharan, Sedeeq Al-khazraji, and Matt Huenerfauth. 2018. Evaluation of an english word look-up tool for web-browsing with sign language video for deaf readers. In *Universal Access in Human-Computer Interaction. Methods, Technologies, and Users: 12th International Conference, UAHCI 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I 12*, pages 205–215. Springer.

Jenelle Rouse Heather Gibson, Shelley Potma. 2021. An Innovative Pedagogical Approach: American Sign Language (ASL) Gloss Reading Program. [Online; accessed 14. Sep. 2024].

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11221–11239.

Shuxu Huffman, Si Chen, Kelly Avery Mack, Haotian Su, Qi Wang, and Raja Kushalnagar. 2024. "we do use it, but not how hearing people think": How the deaf and hard of hearing community uses large language model tools. *Preprint*, arXiv:2410.21358.

Mert Inan, Katherine Atwell, Anthony Sicilia, Lorna Quandt, and Malihe Alikhani. 2024. Generating signed language instructions in large-scale dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 140–154, Mexico City, Mexico. Association for Computational Linguistics.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational

approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Dilay Z. Karadöller, David Peeters, Francie Manhardt, Asli Özyürek, and Gerardo Ortega. 2023. Iconicity and gesture jointly facilitate learning of L2 signs at first exposure. *Language Learning*.

Emily Kubicek and Lorna C. Quandt. 2019. Sensorimotor system engagement during ASL sign perception: An EEG study in deaf signers and hearing non-signers. *Cortex*, 119:457–469.

Emily Kubicek and Lorna C. Quandt. 2021. A Positive Relationship Between Sign Language Comprehension and Mental Rotation Abilities. *J. Deaf Stud. Deaf Educ.*, 26(1):1–12.

Huije Lee, Jung-Ho Kim, Eui Jun Hwang, Jaewoo Kim, and Jong C. Park. Leveraging Large Language Models With Vocabulary Sharing For Sign Language Translation. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 04–10. IEEE.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. 2023. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada. Association for Computational Linguistics.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *Preprint*, arXiv:2311.10122.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *Preprint*, arXiv:2301.13688.

Pamela Luft. 2023a. Promoting Independent Literacy for ASL Readers With Disabilities. In *Strategies for Promoting Independence and Literacy for Deaf Learners With Disabilities*, pages 20–70. IGI Global.

Pamela Luft. 2023b. Using Comprehensive Observational Data to Improve Reading Instruction: Case Studies of DHH Student Readers. In *Cases on Teacher Preparation in Deaf Education*, pages 102–145. IGI Global.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2021. Real-Time Sign Language Detection Using Human Pose Estimation. In *Computer Vision – ECCV 2020 Workshops*, pages 237–248. Springer, Cham, Switzerland.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,

Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

AJ Piergiovanni, Isaac Noble, Dahun Kim, Michael S. Ryoo, Victor Gomes, and Anelia Angelova. 2023. Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities. *Preprint*, arXiv:2311.05698.

Leigh Ellen Potter, Jessica Korte, and Sue Nielsen. 2014. Design with the deaf: do deaf children need their own approach when designing technology? In *ACM Conferences*, pages 249–252. Association for Computing Machinery, New York, NY, USA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Laura Blackburn Samuel J. Supalla. 2021. Why Signed Language Reading is Important. [Online; accessed 14. Sep. 2024].

Sven Schild and Constance J Dalenberg. 2012. Trauma exposure and traumatic symptoms in deaf adults. *Psychological trauma: Theory, research, practice, and policy*, 4(1):117.

Anthony Sicilia and Malihe Alikhani. 2022. LEATHER: A framework for learning to generate human-like text in dialogue. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 30–53, Online only. Association for Computational Linguistics.

Daniel Stein, Jens Forster, Uwe Zelle, Philippe Dreuw, and Hermann Ney. 2010. Rwth-phoenix: Analysis of the german sign language weather forecast corpus. In *sign-lang@ LREC 2010*, pages 225–230. European Language Resources Association (ELRA).

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *Int. J. Comput. Vision*, 128(4):891–908.

Samuel J. Supalla. 2017. A Sketch on Reading Methodology for Deaf Children. [Online; accessed 13. Sep. 2024].

M Virginia Swisher. 1989. The language-learning situation of deaf students. *Tesol Quarterly*, 23(2):239–257.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*.

Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. 2023. Including facial expressions in contextual embeddings for sign language generation. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 1–10, Toronto, Canada. Association for Computational Linguistics.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2gpt: Leveraging large language models for gloss-free sign language translation. *Preprint*, arXiv:2405.04164.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceed-*

ings of the 28th International Conference on Computational Linguistics, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *Preprint*, arXiv:2306.13549.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *Preprint*, arXiv:2306.02858.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *CVPR*.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *Preprint*, arXiv:2310.01852.

A Hyperparameters & Training Implementation Details

We trained all of the models on an Apple MacBook Pro with an M3 Max chip. Libraries used were PyTorch, Huggingface TRL, Transformers, Datasets, Evaluate, and W&B. The hyperparameters for the LLaMA models are: learning rate of 1e-3, lr scheduler type: "reduce lr on the plateau", per device training batch size of 2, number of epochs of 5, and weight decay of 0.01, and maximum sequence length of 300 tokens. LoRA configuration for the LLaMA model is: rank of 8, LoRA alpha of 32, and LoRA dropout of 0.1. For the LLaVA model: mm projector learning rate of 2e-5, one epoch, batch size of 2, learning rate of 5e-5, linear lr scheduler type, maximum sequence length of 2048. LoRA configuration for LLaVA model: LoRA rank: 128, and LoRA alpha: 256.

B All Prompt Types

Here we present all the prompt types that have been used in the experiments:

- **zero-shot prompt:** This is a sentence in German Sign Language glosses: <glosses>. You MUST translate these to spoken German. You MUST give the answer directly without any other text.

- **rule-based prompt:** "Instructions Here are some basic rules of German GLOSSES: 1) German signs correspond to meanings not to words. 2) Some GLOSSes are formed from more than one German word. In this case the words are joined by a hyphen. The hyphen indicates one single sign that is labeled with two or more German words. 3) Glosses combined with a plus sign are two separate signs that are joined together to make what appears to be a single sign 4) In DGS, some signs are repeated for specific meaning. for instance LEARN + LEARN changes the sign from the VERB “To Learn” to the NOUN “Learning.” 5) Words that are to be Fingerspelled are indicated in one of two ways: - Separated by hyphens between each Fingerspelled letter: G-L-A-D-Y-S - Preceded by the initials FS in parenthesis: (fs) GLADYS. Task You MUST translate <glosses> of DGS to German without using any special characters, according to these rules."

- **notation-based prompt:** "Instruction Below is a list of common symbols used in the writing of DGS Glosses: - The Crosshatch: This symbol indicates a loan sign, a sign originating from the fingerspelling of an English word. - Parentheses: ( ) Additional information about the production of a sign is can added to the written gloss between a set of parentheses. Such information can be abbreviated as in (2h)DO++, or it may appear as German instructions to add information to a sign: GIVE (left), or to a Classifier CL:1 (man hurries past). - CL: The abbreviation CL: indicates a classifier. The information following the colon indicates the hand shape and number of hands. - The Umlaut (two dots above a

Ground Truth Text	Intensified Gloss	Generated Text
woher soll die wärme derzeit auch kommen	WOHER <HIGH-INT>WARM </HIGH-INT>	wo es auch am wochenende warmer wird 🔥🔥🔥🔥
where is the heat supposed to come from?	WHERE WARM	where it gets warmer on the weekend too 🔥🔥🔥🔥
morgen gibt es mal sonne mal wolken in der nordhälfte zeigt sich die sonne häufiger	MORGEN SONNE WOLKE WECHSELHAFT NORD <HIGH-INT>SONNE </HIGH-INT>OFT	morgen wird es auch wieder sehr sonnig und teils auch wolkig und auch **hochintensiv** 🌞
Tomorrow there will be clouds, in the north half the sun will be more common	TOMORROW SUN CLOUD VARIABLE NORTH <HIGH-INT>SUN </HIGH-INT>OFTEN	Tomorrow it will be very sunny and sometimes also cloudy and also **highly intensive** 🌞

Table 5: This table shows three samples of German Text, DGS Gloss, and the generated text by the LLaMA2 7b+ model. Each sample includes a translation in English as well. LLaMA learns to depict intensifier tokens as emojis without any instructions or training data examples.

given hand shape) (indicate the bending of the fingers of that hand. The 3 (called the "bent three") is the hand shape used in the sign "INSECT". This technique is only used in reference to a specific handshape such as a classifier.

Task You MUST translate <glosses> to German according to these symbols."

- **one-shot prompt:** "Example ""Here's a sample DGS gloss: "ORT REGEN DURCH REGEN KOENNEN UEBERSCHWEMMUNG KOENNEN" which translates to ""mancherorts regnet es auch länger und ergiebig auch lokale überschwemmungen sind wieder möglich"" in German

  Task You MUST translate <glosses> to German according to this example. "

C The Effect of Model Size

TEST SET				
Models	$B_1 \uparrow$	$B_2 \uparrow$	$R_{LSum} \uparrow$	$BS_{F1} \uparrow$
LLaMA3 8B	**12.057**	1.968	0.144	0.764
LLaMA3 70B	11.281	**2.054**	**0.175**	**0.798**

Table 6: This table shows the performance differences between LLaMA3 8B, and LLaMA3 70B variants. The bigger model generates more intelligible sentences, yet fails to carry out the translation task.

***RQ2:* How does the number of parameters affect the performance of the model in text-based SLP tasks?** We show the effects of the number of parameters of the text-based model for the G2T task in Table 6. A higher number of parameters does not always correlate with better automatic metric results. A higher number of parameters also increases the fine-tuning duration.

D Towards Prosodic, Iconic and Semantically-Rich Sign Language Representations via LLMs

SLs and the current machine learning setups for SLP systems have been constrained to multimodal translation systems mostly, as can be seen from our tasks as well. However, sign interpretation and production by humans are not translation-based processes between modalities. Cognitive science, neuroscience, and linguistics research into the SLs by Kubicek and Quandt (2019, 2021) show that prosody during signing affects interpretation and action recognition, and Karadöller et al. (2023); Chen et al. (2023a); Campisi et al. (2023) show that different SLs use different levels of iconicity and iconic signs can facilitate interpretation. In this section, we present a case study on the current iconicity characteristics that are developed during the fine-tuning of the LLaMA3 model by using emojis as placeholders for intensifiers.

D.1 Iconicity Case Study: Emojis as Intensifiers

During the fine-tuning of the LLaMA3 8b+ model, it has been observed in the generated outputs for the intensified tasks there are emojis, even though the model is not instructed to include emojis, and the training set does not contain emoji tokens for the RWTH-PHOENIX-14-T. Some samples are shown in Table 5. Here, it is observed that the model is mapping the intensifier tokens that exist in the intensified dataset to emojis. However, this is not a one-to-one mapping, and it is more so using the iconicity of the emoji to depict semantics that does not exist in the textual glosses.

It can be claimed that iconicity, which is normally depicted in the spatial modality during the signing, is now depicted with a different modality in a semantically rich textual form. Also, in the last sample, the generation directly includes "highly

intensive," which shows that sometimes the model does not map the intensifier tokens directly to emojis. Overall, it can be qualitatively claimed that this mapping of semantics to icons via emojis is a property of LLMs fine-tuned on multiple tasks. This provides a paradigm shift in SLP, where including prosodically-rich tasks of SLs can be accomplished with the help of large foundation models instead of seeing them as translation problems. Yet, new task definitions and datasets specific to SLs should be made available for further investigations of these capabilities.

E The Glossing Trade-Off

This section presents a trade-off between using textual representations of signs such as glosses or SignWriting that are linguistically-backed or directly using video of signers. This trade-off may not be an option most of the time, as having access to intermediary textual representations such as glosses as part of the sign corpora is not prevalent across all datasets available online. To decide whether to use glosses or videos, we can use insights from the linguistics literature and data collection experience from the RWTH-PHOENIX-14-T dataset.

In the original data collection effort as described by Forster et al. (2012) and Stein et al. (2010), the annotations of glosses are done by a congenitally Deaf person with no previous annotation experience. On average, they report that it took the annotator 24 hours to annotate 15 minutes of footage. When we compare these statistics to the fine-tuning statistics of the text-based and multimodal models, we can observe the trade-offs better. This is presented in Table 7. It can be seen that the text-based model has nearly double the performance of the multimodal, and it needs less storage space and leads to less carbon emissions, even though it takes longer to annotate.

Trade-off Statistics						
	$\mathbf{T}_A$ (h)	$\mathbf{T}_{FT}$ (h)	$\mathbf{T}_I$ (s/tok)	**S** (GB)	**Carbon Emissions** (kg)	**Perf.** (B_1)
Annotator + Text-Based	2400	8	4	0.1	0.211	22.85
Multimodal	0	8	8	50	0.240	13.62

Table 7: This table shows different statistics comparing the human annotation with the text-based model and video-based multimodal model. Carbon emissions are calculated using the US EPA's greenhouse gas equivalencies calculator. $\mathbf{T}_A$: average time for annotation, $\mathbf{T}_{FT}$: average time for fine-tuning, $\mathbf{T}_I$: average time for inference, **S**: storage space needed for data.

F Discussion of LLMs in SLP Research

After these detailed analyses, in our findings section, we discuss the implications of these pretrained and fine-tuned LLMs on SLP tasks. First, it is important to note that translation is not the only area that needs attention under sign language processing. With instruct-tuned end-to-end dialogue systems like LLMs, it becomes ever more important to include SLs in the pretraining and fine-tuning if we are to claim that they are truly universal large *language* models. This can be achieved by including SLP during the pretraining and fine-tuning stages without losing performance in spoken language tasks, as we have shown in this paper.

As noted in the glossing trade-offs in section § E, SLs have multiple ways of representation (text, image sequences, graphs, skeletal position coordinates), and deciding which modalities are linguistically relevant for language models to be trained on is important. Opening up the venue of fine-tuned LLMs for SLs allows more development on signed iconicity, phonology, prosody, and dialogue for the future versions of these LLMs (please see Appendix D for a case study on the representation of iconicity of SLs with LLMs), just like some current LLMs that are capable of some those aspects for spoken languages.

The more we build separate translation systems for SLs, the more we lose the universality of LLMs, steal from the future integration of SLs into LLMs, and turn away from the needs of the Deaf and Hard-of-Hearing community. To prevent this, we presented the first universal LLM suite, which can carry out language understanding tasks independent of its modality (spoken or signed).