SSP: Self-Supervised Prompting for Cross-Lingual Transfer to Low-Resource Languages using Large Language Models

Vipul Rathore **Aniruddha Deb** **Ankish Chandresh** **Parag Singla** **Mausam**
Indian Institute of Technology
New Delhi, India
{rathorevipul28, aniruddha.deb.2002, iitdelhi24ankish}@gmail.com
{parags, mausam}@cse.iitd.ac.in

Abstract

Recently, very large language models (LLMs) have shown exceptional performance on several English NLP tasks with just in-context learning (ICL), but their utility in other languages is still underexplored. We investigate their effectiveness for NLP tasks in low-resource languages (LRLs), especially in the setting of *zero-labeled* cross-lingual transfer (0-CLT), where no labeled training data exists for the target language but data from related medium-resource languages (MRLs) and unlabeled test data for the target language are available. We introduce Self-Supervised Prompting (SSP), a novel ICL approach tailored for the 0-CLT setting.

SSP leverages the key observation that LLMs output more accurate labels if in-context exemplars are given from the target language, even if their labels are slightly noisy. To operationalize this, since target language training data is not available in 0-CLT setup, SSP operates in two stages. In Stage I, using source MRL training data, target language's test data is noisily labeled. In Stage II, these noisy test data points are used as exemplars in ICL for further improved labeling. Additionally, our implementation of SSP uses a novel Integer Linear Programming (ILP)-based exemplar selection method that balances similarity, prediction confidence and label coverage. Experimental results on three tasks and eleven LRLs (from three regions) demonstrate that SSP strongly outperforms existing SOTA fine-tuned and prompting-based baselines in the 0-CLT setting.

1 Introduction

Very large language models (LLMs) such as GPT-3.5-Turbo & GPT-4 (Ouyang et al., 2022; Achiam et al., 2023) show remarkable performance on a variety of NLP and reasoning tasks via *In-Context Learning* (ICL) (Brown et al., 2020; Chowdhery et al., 2023). ICL feeds a task-specific instruction along with a few exemplars, appended with the test input, to the LLM. As LLMs can be highly sensitive to exemplars (Zhao et al., 2021), efficient exemplar retrieval becomes essential for ICL.

While LLMs have shown excellent performance on English tasks, their effectiveness in other languages remains relatively underexplored. In this work, we study *zero-labeled cross-lingual transfer* (0-CLT) to low-resource languages (LRLs) – a setting where labeled task data from one or more related medium-resource languages (MRLs) is available, but no labeled data exists for the target LRL. We additionally leverage the available test sentences (unlabeled) in the target language. The high cost of annotating the sentences in LRLs for new tasks or domains highlights the relevance of the 0-CLT setting.

Cross-lingual transfer has been addressed through standard fine-tuning (Muller et al., 2021; Alabi et al., 2022), and language adapters (Pfeiffer et al., 2020; Üstün et al., 2020; Rathore et al., 2023), but there is limited work on cross-lingual ICL. There are two exceptions (Ahuja et al., 2023; Asai et al., 2024), where ICL is employed with exemplars from a source language, but they use uniformly random sampling for exemplar selection, resulting in performance inferior to cross-lingually fine-tuned models, such as mBERT and XLM-R (Devlin et al., 2019; Conneau et al., 2020).

In our preliminary experiments, we prompt the GPT-4 model with exemplars from source MRLs, and compare its performance with the same LLM prompted with exemplars from the target LRL. We vary the label noise on the target exemplars. Unsurprisingly, LLMs show better performance with less label noise. More interestingly, we find that a reasonably-sized noise region exists (see Figure 1), such that if the exemplar noise is within that range, then the overall performance is higher than prompting with accurate source language data.

Armed with this observation, we present Self-Supervised Prompting (SSP) – a novel ICL frame-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15081–15102
November 12-16, 2024 ©2024 Association for Computational Linguistics

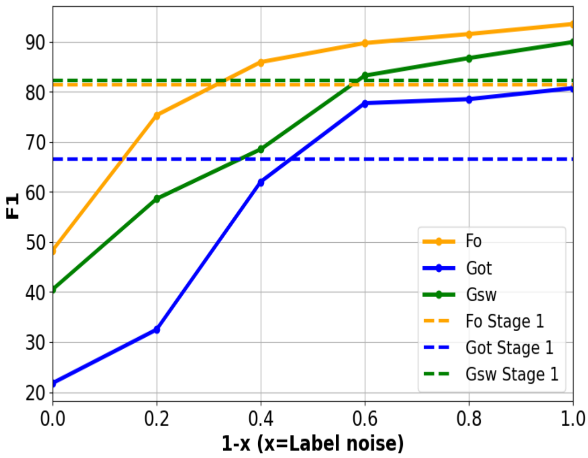

Figure 1: GPT-4, prompted with target LRL exemplars, along with artificially injected label noise (x-axis) for POS tagging task in 3 Germanic LRLs. Dashed lines represent F1 scores when prompted with source MRL exemplars (i.e. Stage 1). Label Noise means the fraction of labels in which noise is injected.

work for 0-CLT to LRLs. Since the target LRL training data is not available in 0-CLT, SSP operates in two stages. In Stage I, SSP labels all test instances of LRL using training data from MRL. This may be done by LLM prompting (as in the experiment above), or using any other existing approaches for 0-CLT, such as by fine-tuning or adapters. Once (noisy) labels on target LRL are obtained, in Stage II, SSP uses ICL using these noisy test data points (except itself) as exemplars for further performance improvement. Additionally, to select the best exemplars, we develop a novel Integer Linear Programming (ILP) based selection approach, which balances the various objectives of (1) similarity of exemplar with test sentence, (2) high confidence in label predictions, and (3) coverage of the various labels for better task understanding. Figure 2 gives an overview of our proposed pipeline.

We define 3 scenarios for our zero-labeled setup - (1) 0-CLT: Only the available test sentences of the target language are used, with no additional unlabeled data, (2) 0-CLT-U: the full wikipedia data available for target language is utilized, and (3) 0-CLT-T: a translation model supporting the target language is leveraged. The primary focus of this work is on 0-CLT (setting 1). However, we also conduct stage 1 experiments for both 0-CLT-U and 0-CLT-T settings. This enables us to comprehensively assess SSP's effectiveness across varying degrees of noise in stage I labelings.

We perform experiments on sequence labeling tasks (POS tagging and NER), and natural language inference (NLI) – a text classification task. Our datasets encompass eleven low-resource languages from typologically diverse language families and three regions: African, Germanic and American. Our experiments show consistent and substantial improvements over existing fine-tuning as well as simpler ICL-based approaches. To encourage reproducibility, we make our code and prompts publicly available.[1]

Our contributions are summarized as follows:

1. We investigate ICL strategies for zero-labeled cross-lingual transfer (0-CLT) to LRLs, using labeled data from related MRLs and unlabeled test data from the target language.
2. We propose SSP, a two-stage self-supervised prompting paradigm for this task, where the first stage may be done by an LLM or any other cross-lingually fine-tuned models.
3. We introduce a novel exemplar selection approach utilizing Integer Linear Programming (ILP). The ILP incorporates similarity to test input along with confidence of stage I predictions, and enforces label coverage constraints.
4. Experiments on 3 tasks and 11 languages show that SSP outperforms existing fine-tuning and SOTA LLM-based models in 0-CLT, 0-CLT-U (full unlabeled) as well as 0-CLT-T (translation-based) settings, hence improving labeling in the second iteration, irrespective of the initial labeling method.

2 Related Work

An ICL prompt consists of (1) task description: to facilitate the understanding of task, (2) labeled input-output pairs: Written sequentially in order of their relevance to input query, and (3) input itself.
Cross-lingual ICL: In general, cross-lingual ICL has not been systematically explored in literature. In existing works, prompting is primarily done in a high-resource language, typically English. This is called *cross-lingual (CL) prompting*. This differs from *in-language (IL) prompting*, where examples are retrieved from the candidate pool of the target language itself. This assumes the availability of labeled data for target LRL, which is not true in our zero-labeled (0-CLT) setting. In response, we develop novel techniques making use of both CL prompting and IL prompting, while not utilizing the gold labels during IL prompting stage.

[1] https://github.com/dair-iitd/SSP

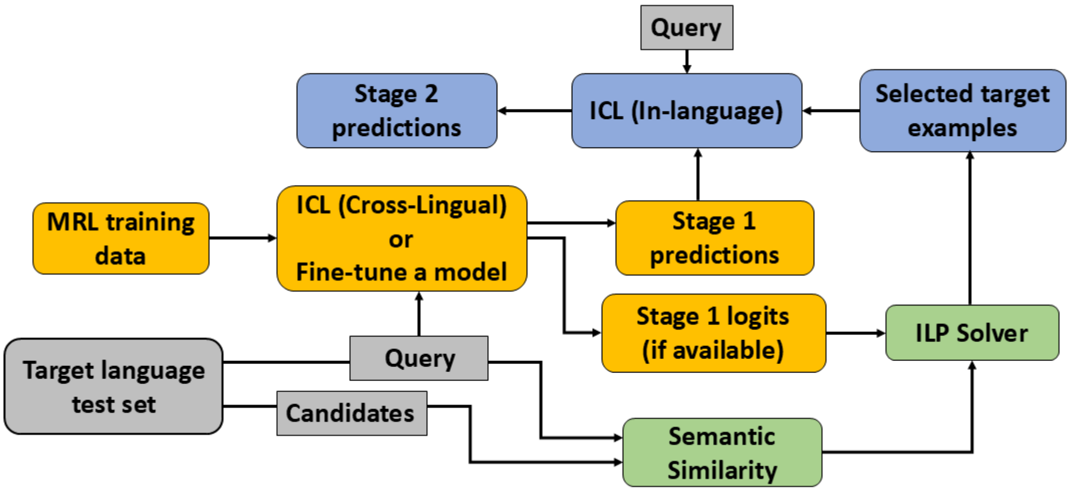

Figure 2: SSP Architecture for Cross-Lingual Transfer to Target Low-Resource Language (LRL). (1) Stage 1 (orange): Fine-tune a model or perform cross-lingual in-context learning (ICL) using medium-resource language(s) (MRL) data. (2) The ILP Solver (green) selects exemplars for Stage 2 based on semantic similarity between the query and candidates from the target language test set, also utilizing logits from Stage 1 predictions. (3) Stage 2 (blue): Perform in-language ICL for the target query using the selected exemplars along with their stage 1 labels.

Most existing cross-lingual ICL methods use uniformly random input-output pairs for exemplar selection (Zhang et al., 2022; Winata et al., 2021; Ahuja et al., 2023; Asai et al., 2024). Recent approaches (Agrawal et al., 2022; Tanwar et al., 2023) address this gap by utilizing *semantic similarity* for cross-lingual retrieval from a high-resource language's labeled data, given the target LRL's instance as query. This is facilitated by embedding-based multilingual retrievers such as multilingual sentence-transformers (Reimers and Gurevych, 2020). More recently, OpenAI-based embeddings such as Ada-002 [2] have been used effectively for cross-lingual retrieval (Nambi et al., 2023). We extend this line of work by also incorporating label confidence and label coverage in exemplar selection.

Self-Adaptive Prompting: Wan et al. (2023) proposed *Universal Self-Adaptive* (USP) framework, which has been explored for only monolingual (English) setting. USP uses an external *unlabeled* dataset of instances and labels them using LLM in Stage I. It then samples multiple Chain-of-thought (CoT) paths to estimate the logits using the same LLM, and then utilizes the entropy of logits for exemplar selection for Stage 2. Our work has similarities to USP in that both methods are two-staged prompting approaches. USP is different from SSP in that the former is much more expensive, since it requires multiple LLM calls to just estimate the logits. USP also does not use any exemplars (and only uses task description) in stage 1, which are quite important for better performance. Finally, USP has only been applied for English tasks, and has not been explored for cross-lingual tasks.

Fine-tuning approaches for Cross-lingual Transfer: Most approaches rely on fine-tuning a Pre-trained LM (PLM) such as BERT or XLM-R on the source languages (Muller et al. (2021); Alabi et al. (2022)) and deploying on an unseen target language. Recently, Language-Adapter-based approaches have been found more effective (Üstün et al., 2020) for cross-lingual transfer settings. For sequence labeling tasks (NER and POS tagging), ZGUL (Rathore et al., 2023) is a recent SOTA method that leverages ensembling Language Adapters from multiple MRLs to label each word in a target language. We leverage this in our proposed SSP pipeline.

Cross-lingual label-projection techniques: Recent methods (Chen et al., 2023a; García-Ferrero et al., 2023; Le et al., 2024) utilize an off-the-shelf translation model (NLLB Team et al., 2022) for label-projection in 2 ways – (1) *Translate-train*: translate from English to target language (X) to generate training data in X, or (2) *Translate-test*: translate test data in X to English to perform label-projection and obtain annotations in X. Although our focus is 0-CLT transfer, we also experiment with these translation models in Stage I, to assess the robustness of SSP across multiple settings.

[2] https://platform.openai.com/docs/guides/embeddings/

3 Self-Supervised Prompting

We define the setting of zero-labeled cross-lingual transfer (0-CLT) as follows. We are given source training data for a specific task: $D = \{(x_i, lg_i, y_i)\}$, where x_i is the input text in language lg_i, and the output is y_i. We are additionally given a set of unlabeled test data points $T = \{q_j\}$ from a target language lg_t. Our goal is to train a model/create a protocol, using D, T and a large pre-trained LLM, that outputs good predictions on T for the task, assuming that lg_t is a low-resource language, due to which its training data is not available, and that languages lg_i are related to lg_t.

Our solution approach, Self-Supervised Prompting (SSP), comprises two key stages as follows. In Stage I, it proposes a noisy labeling for all data points in T using source data D. This may be done in different ways, as described next. In Stage II, it uses the LLM and noisy labeling on T from Stage I as exemplars to improve the labelings. Furthermore, SSP uses a novel integer-linear programming based exemplar selection. We now describe each component of our system.

3.1 Stage I: Initial labeling using source data

To create a first labeling for all test points, SSP can use any existing approaches for 0-CLT, such as fine-tuning a multilingual language model for the task, or use of language adapters or using our LLM with in-context exemplars from source language. In our experiments, we experiment with adapters and ICL, which we briefly describe next.

Cross-Lingual ICL: In the method, we use ICL over LLM for obtaining Stage I labelings. First, we retrieve a set of top-K exemplars from D using each test instance q_j as query. This selection is based on cosine similarity between their *Ada-002* embeddings. The selected exemplars are arranged in descending order of similarity scores, and included in the prompt between the task description (TD) and the input test instance. This approach has two drawbacks. First, since the LLM will typically be a large expensive model – this will require an LLM call per test data point in Stage I. Second, generally, these LLMs do not expose their logits, hence, we will not have access to prediction confidences from Stage I labelings.

Training smaller model(s) using D: Another possibility is to fine-tune a smaller multilingual LM, such as mBERT or mDeBERTa-v3 (He et al., 2021) on D for NLI task. For sequence labeling, we can use ZGUL (Rathore et al., 2023), which trains source language adapters using D, and uses inference-time fusion of source adapters for labeling test data points. These approaches can provide Stage I labelings for T along with prediction confidences, without making any expensive LLM calls.

3.2 Stage II: in-language ICL using ILP-based exemplar selection

After Stage I predictions for target instances T are obtained, SSP prompts the LLM to label each test data point $q \in T$, but uses in-context exemplars in target language using Stage I labelings. For exemplar selection, SSP implements a novel integer linear program (ILP) that balances *semantic similarity, prediction confidence* (when available) and *label coverage*.

Our primary objective is to maximize the aggregated semantic similarity of the selected exemplars, which is obtained using cosine similarity score between their OpenAI Ada-002 embeddings. In addition, we impose two constraints:

- **Label Coverage**: The ILP tries to ensure the coverage of all labels for the given task in the selected exemplars – this has been found effective for ICL (Min et al., 2022).
- **Confidence**: In case logits for Stage I model are accessible (unlike the OpenAI LLMs), the ILP prefers selection of more confident exemplars. Our hypothesis is that confident predictions are also accurate (assuming the model is well-calibrated), and previous work has shown that performance of LLMs can be sensitive to correctness of exemplars (Wei et al., 2023)

SSP formulates these three factors into an ILP as follows. For a dataset D with n examples indexed from $\mathcal{I} = \{1 \ldots n\}$, given a test data point (query) q_j, let z_i be a binary variable denoting whether i^{th} test instance q_i is selected as an exemplar. We use a semantic similarity function sim(q_i, q_j) to get the similarity between two examples. K is the number of exemplars to be selected. Since q_j cannot be an exemplar for itself, we select exemplars from the set $\mathcal{I} \setminus \{j\}$ only.

Let the set of all labels for the given task be $\mathcal{L}$, and the multiset of all labels predicted (using argmax) for example q_i be L_i. The Stage I prediction confidence for label l in q_i is denoted as $\hat{y}_l^i$. This confidence is computed as average of probability scores across all predictions of label l in i^{th}

sentence (details in Appendix A). The ILP uses a threshold τ_l for prediction confidence for a label l. Intuitively, the ILP maximizes the semantic similarity of K chosen exemplars, subject to each label l being present at least once in the exemplars, and average prediction confidence of each data point for each label being greater than τ_l.

Formally, the ILP is formulated as

$$\max \sum_{i \in \mathcal{I} \setminus \{j\}} z_i \cdot \text{sim}(q_i, q_j) \tag{1}$$

$$\text{such that} \sum_{i \in \mathcal{I} \setminus \{j\}} z_i = K \tag{2}$$

$$z_i \cdot (\hat{y}_l^i - \tau_l) \geq 0 \, \forall \, i \in \mathcal{I} \setminus \{j\}, \forall \, l \in L_i \tag{3}$$

$$\sum_{i \in \mathcal{I} \setminus \{j\}} z_i \cdot \text{count}(L_i, l) \geq 1 \, \forall \, l \in \mathcal{L} \tag{4}$$

Here $\text{count}(L_i, l)$ denotes the number of occurrences of l in L_i. In our experiments, we set $K = 8$, and $\tau_l = 80^{th}$ percentile threshold of the set $\{\hat{y}_l^i\}_{i=1}^n$ for a particular label l. The idea is to have label-specific threshold since the fine-tuned model may not be calibrated equally for all labels.

Since logits are not accessible for OpenAI LLMs GPT-3.5 and GPT-4x, in case Stage I labeling is done by either of these models using ICL, we skip the confidence thresholding constraint of ILP. This means that for this variant of SSP, the selection is made based on only similarity and label coverage.

4 Experiments

Our main experiments assess SSP performance compared to existing state-of-the-art models for 0-CLT. We also wish to compare various SSP variants, and estimate the value of the ILP-based exemplar selection.

4.1 Tasks and Datasets

We experiment on three tasks – POS tagging, NER and Natural Language Inference (NLI). We use the UDPOS dataset (Nivre et al., 2020) for POS tagging over Germanic languages, MasakhaNER (Adelani et al., 2021) for African NER, and AmericasNLI (Ebrahimi et al., 2022) for NLI task on the indigenous languages of Americas. Overall, we use eleven low-resource test languages as target (e.g., Kinyarwanda, Faroese, and Aymara), and 2-4 source languages per dataset (e.g., Icelandic, Spanish and Swahili; always including English). Further details are in Tables 5 and 6.

Recent studies have shown sensitivity of the output to the template/format of input-output pairs written in the prompt (Sclar et al., 2023; Voronov et al., 2024). We follow the best template given in Sclar et al. (2023) for NLI, while for sequence labeling, we explore various templates on our own and report our results on the best one. We refer to Appendix B for details and the exact templates used for each of our tasks.

For obtaining test set, we randomly sample 100 test samples for each target language for NER and POS tasks. We justify this as each sentence has multiple labels, bringing the total no. of instances to be labeled per language to 2370 and 1100 for POS and NER respectively. For the NLI task, we sample 501 test samples (167 for each class: 'entailment', 'contradiction' and 'neutral'). We report statistical significance (in table captions) to justify our evaluation.

We also perform a careful contamination study, following (Ahuja et al., 2022), by asking LLMs to fill dataset card, complete sentence (and labels), given partial sentence, and generate next few instances of the dataset. As further detailed in Appendix F, we do not observe any evidence of contamination for these languages' test splits in the OpenAI LLMs.

4.2 Comparison Models

LLMs: We experiment with a series of advanced LLMs – GPT-3.5-turbo (Ouyang et al., 2022), GPT-4x (GPT-4/GPT-4-Turbo) (Achiam et al., 2023), and LLaMa-2-70b (Touvron et al., 2023) for each task. For NER and NLI, we use GPT-4-Turbo due to its superior performance compared to GPT-4. However, for POS tagging, we opt for GPT-4 instead, as GPT-4-Turbo encounters challenges in following the instructions and generating outputs compatible with the verbalizer utilized in our experiments (details in App. B). We present the exact version details of OpenAI LLMs in table 4.

Zero-shot Baselines: We compare our SSP approach with the SoTA fine tuning models, as well as LLM-based ICL methods using naive random exemplar selection. In particular, we fine-tune ZGUL – mBERT Language Adapter-based SoTA zero-shot baseline for NER and POS tagging, and mDeBERTa fine-tuned for NLI. We additionally utilize the public model mDeBERTa-v3-base-xnli (Laurer et al., 2022) for NLI evaluation. We term our own fine-tuned model as mDeBERTa^{FT} and the

Model	Hau	Ibo	Kin	Lug	Luo	Avg.	Fo	Got	Gsw	Avg
zero-labeled (0-CLT)										
Full Fine-Tuning (FFT)	49.9	54.9	55.4	56.3	40.2	51.3	77.6	17.8	62	52.5
CPG (Üstün et al., 2020)	48.6	50.4	52.6	54.3	38.6	48.9	77.3	16.9	63.9	52.7
ZGUL	52.2	56	53.7	54.5	44.4	52.2	77.2	21.1	65	54.4
ICL-Llama-2-70b	64.3	61.2	59.2	60.1	47.3	58.4	79.1	36.0	71.8	62.3
ICL-GPT-3.5-turbo	54.5	69.2	57.8	63.7	46.4	58.3	81.2	37.9	72.2	63.8
ICL-GPT-4x	64.7	80.8	64.6	71.0	53.3	66.9	81.3	66.5	82.3	76.7
SSP(ICL)-llama-2-70b	57.6	62.6	56.0	57.6	43.1	55.4	78.5	37.9	73.5	63.3
SSP(ICL)-GPT-3.5-turbo	62.8	68.4	64.0	63.8	47.6	61.3	82.4	63.2	79.4	75.0
SSP(ICL)-GPT-4x	67.2	79.6	63.3	74.1	54.4	67.7	81.8	**73.7**	85.4	**80.3**
SSP(ZGUL)-Llama-2-70b	68.4	58	56.1	54.7	42.3	55.9	79.9	39.9	72.9	64.2
SSP(ZGUL)-GPT-3.5	61.1	68.9	62.1	67.1	51.4	62.1	**82.8**	67.5	77	75.8
SSP(ZGUL)-GPT-4x	71.2	**82.4**	**71.4**	**75.4**	**55.1**	**71.1**	82.2	71.5	**85.6**	79.8
w/o Conf. thresholding	**71.3**	81.9	69.2	74.6	52.7	69.9	82.8	57	81.4	73.7
w/o Label Coverage	71.1	79.8	**71.4**	**75.4**	**55.1**	70.6	82.2	71.6	**85.6**	79.8
w/o both (sim-based)	70.3	81.8	68	74.8	51.9	69.4	82.4	55.8	82.3	73.5
w/o ILP (Random)	64.1	77.6	61.5	66.1	46.6	63.2	80.6	54.8	80.9	72.1
Translate-train (0-CLT-T)										
ZGUL	72.5	68.5	67.9	65.5	47.3	64.3	-	-	-	-
ICL-GPT-4x	68.7	78.1	58.7	76.3	53.8	67.1	-	-	-	-
SSP(ZGUL)-GPT-4x	75.1	76.7	72.3	79.9	54.4	71.7	-	-	-	-
SSP(ICL)-GPT-4x	69.9	79.8	60.6	74.7	53.8	67.8	-	-	-	-
Translate-test (0-CLT-T)										
Self-fusion (GPT-4x) (Chen et al., 2023b)	68.4	68	58.8	66.5	39.7	60.3	83	-	70	-
SSP(Self-fusion)-GPT-4x	70	78.6	64.6	77	51.3	68.3	83.7	-	83.9	-
Unlabeled data (0-CLT-U)										
AfriBERTa (Ogueji et al., 2021)	75.4	79.1	64.9	54.7	39.3	62.7	-	-	-	-
ZGUL++ (Rathore et al., 2023)	78.5	68.9	62.5	66	50.2	65.2	81.5	18.7	80.4	60.2
SSP(ZGUL++)-GPT-4x	75.6	84.7	70.3	75.4	54.6	72.1	83.9	71.7	86	80.5
Skyline (GPT-4x)	*75.5*	*85.9*	*70.7*	*73.6*	*67.2*	*74.6*	*93.5*	*80.7*	*89.9*	*88*

Table 1: Micro-F1 scores for African NER (left) and Germanic POS (right). Best 0-CLT results are bolded while overall best results are underlined. Translate-train baselines could not be run for POS tagging due to absence of label-projection models for POS. However, Translate-test was possible as label-projection is performed using GPT-4 (Exception being Gothic, as it's translation is not supported in NLLB-200). Statistical significance of bold numbers (0-CLT comparison): McNemar p-value = 0.008 and 0.0004, respectively.

public model as mDeBERTa[100], as it was trained on 100 languages (excluding our target languages). For POS and NER, we also add full parameter fine-tuning and Conditional Parameter Generation (CPG (Üstün et al., 2020)) baselines, all fine-tuned using the same underlying LM (i.e. mBERT).

SSP Variants: We implement SSP with all 3 LLMs – LLaMa-2-70b, GPT-3.5-turbo, and GPT-4x (GPT-4/GPT-4-Turbo). If Stage I uses ICL, then the same LLM is used for both stages I and II. Alternatively, ZGUL and mDeBERTa based methods are also used in Stage I of SSP.

To understand the value of the ILP, we perform three ablations on exemplar selection strategy – (a) without confidence thresholding (for fine-tuned LM), (b) without label coverage and (c) without both, i.e. pure similarity-based. The ablations are conducted with the best performing underlying LLM i.e. GPT-4x.

Leveraging Translation Models and Unlabeled Data: For a comprehensive evaluation, we use the cross-lingual label projection models *Codec* (Le et al., 2024) for translate-train and *Self-fusion* (Chen et al., 2023b) for translate-test baselines. More details are provided in Appendix A.1. Additionally, we leverage unlabeled data in the target language to establish a stronger baseline. We use the AfriBERTa encoder (Ogueji et al., 2021) for African languages and ZGUL++ (Rathore et al., 2023), which utilizes target Wikipedia data to pre-train a target language adapter, and fuses it with MRL adapters for fine-tuning on MRL data.

Skyline: To understand the current performance gap due to lack of target language training data, we also implement a skyline utilizing the gold annotated testset for target languages and perform *few-shot similarity-based* exemplar selection (using Ada-002) for *in-language* ICL to the LLM.

5 Results and Analysis

We present the results for all tasks in Tables 1, and 2. ICL-X represents ICL over an LLM

Model	Aym	Gn	Nah	Avg.
0-CLT				
mDeBERTa^{100}	34.9	43.9	48.9	42.6
mDeBERTa^{FT}	33.9	47	46.9	42.6
ICL-GPT-3.5	38.2	41.7	35.3	38.4
ICL-GPT-4-turbo	32.8	55.8	42.2	43.6
SSP(ICL)-GPT-3.5	38.4	38.8	43.2	40.1
SSP(ICL)-GPT-4-turbo	37.5	58.5	51.8	49.3
SSP(mDeBERTa^{FT})-Llama-2	36.5	37.8	41	38.4
SSP(mDeBERTa^{FT})-GPT-3.5	**43.1**	46	46.8	45.3
SSP(mDeBERTa^{FT})-GPT-4-turbo	36	**61.3**	**59.2**	**52.2**

Model	Aym	Gn	Nah*	Avg.
w/o Conf.	42.9	60.1	50.3	51.1
w/o Label	37	58.2	57.4	50.9
w/o both	34.3	59.7	57.1	50.4
w/o ILP (Random)	33.4	53.8	53.4	46.9
Translate Train				
ICL-GPT-4-turbo	42.4	49.5	-	-
SSP(ICL)-GPT-4-turbo	<u>44.4</u>	58.6	-	-
Translate Test				
ICL-GPT-4-turbo	36.4	45.5	-	-
SSP(ICL)-GPT-4-turbo	42.4	57.6	-	-
Skyline (GPT-4x)	*49.2*	*55.6*	*60*	*54.9*

Table 2: Micro-F1 scores for Americas NLI (Statistical significance of bold number (0-CLT comparison): McNemar p-value = 0.054). * Nahuatl (Nah) not supported in NLLB-200.

X with source language exemplars i.e. stage 1. SSP($model$)-X represents the use of $model$ for Stage I followed by LLM X for Stage II. Whenever ICL is used in Stage I, then the same LLM X is used for both stages.

Analyzing the results, we first observe that all ICL-X baselines perform much better than previous fine-tuning approaches for the 0-CLT task. This reaffirms the importance of studying and improving in-context learning over very large language models for our setting.

Comparing among SSP variants, it is not surprising that GPT-4x performance supercedes GPT-3.5, which is much better than Llama2 70B. We next compare ICL baselines and SSP variants, when using the same LLM. We find that SSP's two stage workflow consistently outperforms ICL by significant margins. In fact, in-language exemplars with very noisy labels from stage 1 (E.g. for Got language with GPT-3.5-Turbo) perform quite well. These observations underscore the value of target language exemplars in ICL, even at the cost of having noisy labels. Moreover, we compare SSP with Stage I via ICL over an LLM vs. via a fine-tuning baseline (ZGUL or mDeBERTa). Fine-tuning baseline for Stage I has two benefits – it is cheaper (due to no LLM calls in Stage I), and has prediction logits available that can allow ILP to select highly confident exemplars for stage II. Due to the latter, in two of the three language groups, the use of a fine-tuning baseline performs much better, and in the third group, it is marginally behind due to weaker performance in one language (Gothic). This happens because ZGUL has a particularly poor performance on this language, leading to much noisier labels in Stage II exemplars.

Finally, we experiment on SSP in 0-CLT-U (full target Wikipedia) and 0-CLT-T (Translation model) settings, as shown in Table 1. We observe that the order of stage I performance is 0-CLT-T (translate-test) < 0-CLT < 0-CLT-T (translate-train) < 0-CLT-U, and same order of performance gets translated in stage II as well, while stage II performance being consistently better than stage 1 in all scenarios.

We further investigate the effect of translation errors (noise) on Stage 1 performance within a translate-test framework and their impact on overall Stage 2 performance. Our analysis shows that translation errors negatively affect Stage 1 performance. This is illustrated in Figure 7 for the Guarani (Gn) language in the NLI task. However, the SSP model demonstrates significant robustness to this noise, achieving a 12 F1 point improvement (from 45.5 to 57.6) in Stage 2 for Guarani. This supports our hypothesis that SSP is effective under varying levels of noise in Stage 1 labelings.

Overall, our best 0-CLT SSP solution uses a fine-tuning baseline (ZGUL or mDeBERTa) for Stage I and GPT-4 for Stage II, using its novel ILP-based exemplar selection. It outperforms closest 0-CLT baselines by around 3 F1 pts, on average, establishing a new state of the art for zero-labeled cross-lingual transfer to low-resource languages. The best SSP reported 0-CLT results are statistically significant compared to the second best counterpart using McNemar's test (p-values in Tables 1 and 2 captions). We believe that our work is a significant advancement to the existing paradigm (Tanwar et al., 2023; Nambi et al., 2023), which is restricted to optimizing only 1 round of In-context learning.

5.1 Ablation Study

We now discuss the results of removing ILP components in Stage II exemplar selection. Tables 1, and 2 (last four rows) report the impact of removing confidence thresholding constraint, label coverage

Model	Neu.	Ent.	Con.	Macro-F1
mDeBERTa-FT	34.7	53	40.3	42.6
SSP(mDeBERTa^{FT})	**51.7**	**53.4**	51.4	**52.2**
(w/o Label)	42.6	52.3	**57.9**	50.9

Table 3: Labelwise F1 scores for fine-tuned model (mDeBERTa^{FT}) and SSP(mDeBERTa^{FT}) w. and w/o label coverage variants (GPT-4-Turbo)

constraint, both of these constraints (i.e., just using similarity) from the ILP. The final row removes ILP completely and presents results of random exemplars in Stage II. All these ablations are done on SSP with ZGUL/mDeBERTa for Stage I, as only those output the prediction probabilities.

Impact of label coverage: We observe an average gain of 1.3 F1 points for AmericasNLI compared to the ablation model that does not impose label coverage constraint. We further compute the average number of exemplars for each label that are covered in the selected set for both methods, along with their label-wise F1 scores (see Figure 3). We observe that the 'neutral' label is not sampled in most cases for *w/o label coverage* variant, while exactly one 'neutral' label is sampled in the SSP(mDeBERTa-FT), with label constraint. This happens as the fine-tuned model mDeBERTa-FT has very poor recall (24) for 'neutral' class and hence any selection strategy has a tendency to not sample this label, unless enforced via a constraint. The class-wise F-1 and recall for SSP(mDeBERTa-FT)-GPT4 with and w/o label coverage are presented in Tables 3 and 8 respectively. We observe a difference of 22 recall points for 'neutral' class (57.6 vs 35.6) between the two ILP variants. An example illustrating this is shown in Figure 8.

Impact of confidence thresholding: For sequence labeling tasks, confidence thresholding plays a key role. This is validated from ablation results in Table 1, wherein removing confidence thresholding from SSP leads to 5.7 points drop for POS tagging (Germanic) and 1.3 points for NER. The drop is particularly significant (around 13.5 points) for Gothic (Got), which shows that not utilizing the confidence scores can lead to drastic drop. This may be because performance of ZGUL is already poor on Gothic (21 F1 points), but confidence thresholding may have likely compensated by picking higher quality exemplars. Removing thresholding would increase noise in exemplars considerably, leading to the drop (see Figure 4).

We further study its impact by computing the quality of Stage II exemplars selected by SSP(mDeBERTa^{FT}), as well as it's ablation variants. We compute the label-wise precision over all K×N (K=8, N=no. of test instances) samples for each target language, and then report their macro-average. We observe for (Figure 3) that the macro-precision of selected exemplars by full ILP is consistently higher than it's other ablation variants, the least value being of w/o both (similarity-based) variant. This implies that the ILP is able to effectively sample high-precision (correctly labeled) exemplars which, in turn, gets translated into it's superior downstream performance on the task.

For completeness, we also show the exemplar precision (correctness) statistics for NER and POS in Figure 4. The trends hold similar in the sense-that 'w/o confidence' and 'similarity-based' variants have significantly lower precision (higher noise) than SSP. This is expected because both these eschew confidence thresholding, leading to sampling of lower-confidence predictions. This translates to worse downstream performance (see Table 1).

We also note that w/o ILP (completely random selection) ablation performs much worse than SSP, showcasing the importance of carefully selecting the exemplar set.

We present an error analysis of SSP approach in section B.2.

5.2 Scalability of SSP with candidate pool size

We explore how the size of candidate pool – used for ILP during exemplar retrieval – affects the performance of SSP(ZGUL)-GPT-4x. We progressively sample bins from the test sets with varying sizes (8, 32, 64, and 100 (the full set)), which serve as candidate sets for ILP. For a fair comparison, evaluation is performed on all 100 test samples (i.e. our original split). The avg. F1 results for African NER and Germanic POS are shown in Fig. 5.

While the performance for Germanic POS seems to scale pretty well and doesn't saturate in the given regime, for African NER it tends to plateau when pool size reaches 64. For completeness, we provide detailed language-wise results in table 10.

6 Conclusions and Future Work

We study the zero-labeled cross-lingual transfer (0-CLT) setting for low-resource languages, when task-specific training data is available for related medium resource languages, along with unlabeled test data for target language. We present Self-Supervised Prompting (SSP) – a novel two-stage

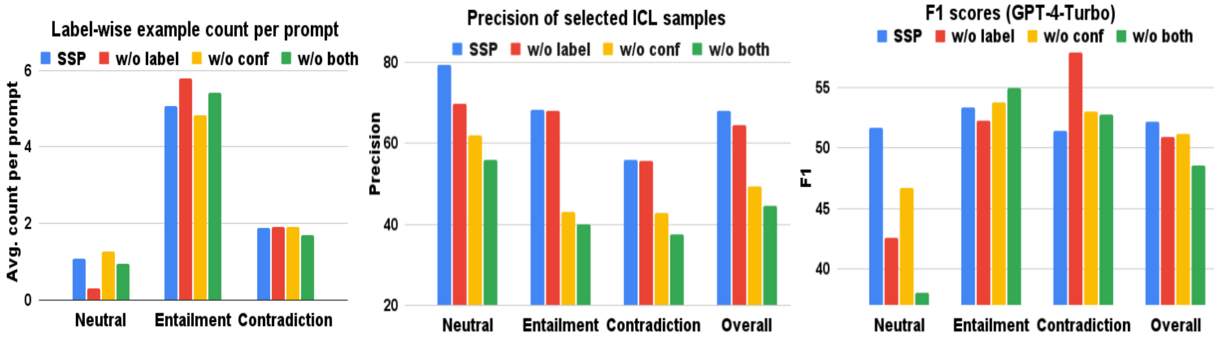

Figure 3: Label-wise statistics for AmericasNLI: Left to right - Label-wise count per prompt, Precision of ICL exemplars, and F1 results (averaged over target languages) using different selection strategies (GPT-4-Turbo)

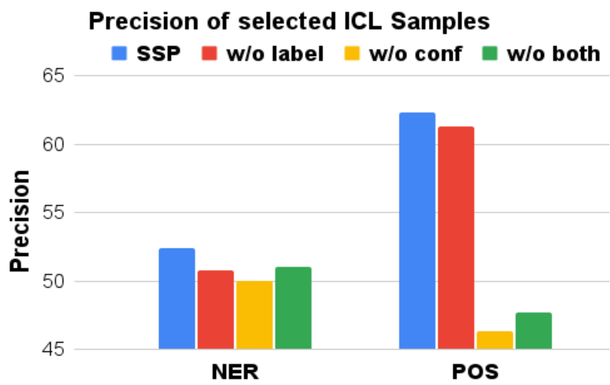

Figure 4: Precision of selected exemplars for African NER and Germanic POS

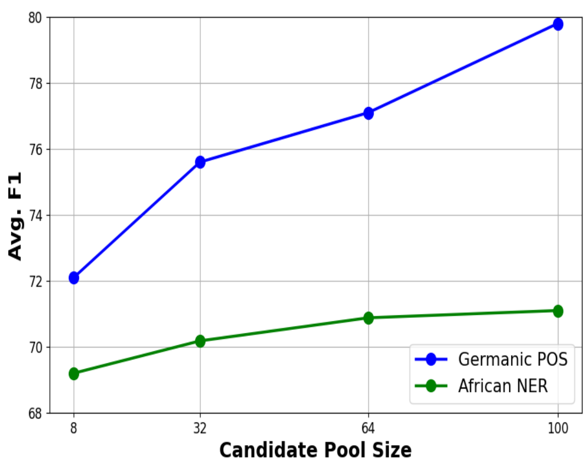

Figure 5: Avg. F1 scores for African NER and Germanic POS as a function of candidate pool size in SSP

framework for the use of in-context learning over very large language models. At a high-level, SSP first noisily labels the target test set using source training data (either by training a model/adapter) or by in-context learning over an LLM. SSP then uses these noisily labeled target data points as exemplars in in-context learning over the LLM. A key technical contribution is the use of integer-linear program that balances exemplar similarity, labeling confidence and label coverage to select the exemplars for a given test point. Thorough experiments on three NLP tasks, and eleven low-resource languages from three language groups show strongly improved performance over published baselines, obtaining a new state of the art in the setting. Ablations show the value each ILP component in downstream performance. We release our code to enable further research in the community.[3]

In the future, we seek to extend our technique to more non-trivial applications such as open generation tasks (Singh et al., 2024; Kolluru et al., 2022). We also posit that smaller fine-tuned models, when calibrated properly, can result in more efficient selection of exemplars to an LLM, as compared to poorly calibrated counterparts, in terms of SSP's downstream performance. We leave a careful and systematic investigation into this hypothesis for future work.

Acknowledgements

Mausam is supported by grants from Google and Jai Gupta Chair Professorship from IIT Delhi. Parag is supported by the IBM AI Horizon Networks (AIHN) grant and Shanthi and K Ananth Krishnan Young Faculty Chair Professorship in AI. Mausam and Parag are supported by IBM SUR awards. Vipul is grateful for a travel grant from the IBM AIHN.

We acknowledge the Microsoft AMR program for supporting our work through an Azure OpenAI grant. Any opinions, findings, conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views or official policies of the funding agencies.

[3] https://github.com/dair-iitd/SSP

Limitations

We show all our results and ablations on the recent state-of-the-art LLMs including GPT4. The inference for these LLMs is expensive, and makes the model deployment infeasible. Other potential limitations are extending our method to tasks such as fact checking, in which the LLMs suffer from *hallucinations* and overprediction issues. The reason why we don't use LLM logits in ILP framework is because they are not openly released by OpenAI and hence, there becomes a need to rely on smaller fine-tuned models - which can potentially lead to sub-optimal downstream performance, in case the fine-tuned models are poorly calibrated. Another serious implication of using LLMs for non-roman script languages is unreasonably high *fertility* (tokens per word split) of the LLM tokenizers, which increases the cost as well as strips the input prompt, which is not desirable.

We also could not evaluate our approach on open generation tasks such as summarization, since their evaluation metrics are not reliable as to obtain a fair comparison of various models. Also, human evaluation could not be done at scale. That said, we note that every task is a generative task for LLM and we pose NLI as a short-form generation, while the POS and NER tasks as a templated long-form generation in current scope of our work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, et al. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022. On the calibration of massively multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4310–4323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800.

L. Bergroth, H. Hakonen, and T. Raita. 2000. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023a. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796.

Yang Chen, Vedaant Shah, and Alan Ritter. 2023b. Better low-resource entity recognition through translation and annotation fusion. *arXiv preprint arXiv:2305.13582*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2023. T-projection: High quality annotation projection for sequence labeling tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15203–15217.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . 2022. Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462.

Akshay Nambi, Vaibhav Balloli, Mercy Ranjit, Tanuja Ganu, Kabir Ahuja, Sunayana Sitaram, and Kalika Bali. 2023. Breaking language barriers with a leap: Learning strategies for polyglot llms. *arXiv preprint arXiv:2305.17740*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siqi Ouyang, Rong Ye, and Lei Li. 2022. On the impact of noises in crowd-sourced data for speech translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 92–97, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

Vipul Rathore, Rajdeep Dhingra, Parag Singla, and Mausam. 2023. ZGUL: zero-shot generalization to unseen languages using multi-source ensembling of language adapters. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6969–6987. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *ArXiv*, abs/2310.11324.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6287–6310, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Eisenschlos, Sercan Arik, and Tomas Pfister. 2023. Universal self-adaptive prompting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7437–7462, Singapore. Association for Computational Linguistics.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Differentiable prompt makes pre-trained language models better few-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Implementation and Hyperparameter Details

We use Azure OpenAI service [4] for all experiments involving GPT-3.5-turbo and GPT-4x models. For LLama-2-70b, we use the together API [5]. We set temperature as 0.0 consistently for all our experiments, making our results directly reproducible. The max_tokens (max. no. of generated tokens) parameter is set to 1024 for POS and NER tasks, while 15 for the NLI. For all experiments, the no. of exemplars (M) is fixed to 8 for uniform comparison. The selected exemplars are arranged in decreasing order of similarity scores with query in a prompt. For ILP solver, we use Python's gurobipy [6] package. For POS and NER tagging, the avg. run-time for ILP per test query = 0.05 seconds, while that of pure similarity-based retrieval = 0.006 seconds. For NLI, avg. ILP run-time is 0.2 seconds while similarity-based run-time is 0.024 seconds.

LLM	Version
GPT-3.5-turbo	gpt-3.5-turbo-0613
GPT-4	gpt-4-0613
GPT-4-turbo	gpt-4-1106-preview

Table 4: LLMs with exact version details

A.1 Translation-based baselines

We explain both translate-train and translate-test methods as follows -

- *Translate-train*: Following (Le et al., 2024), we employ *Codec* method to generate training data in target language X, X^{train}, using MRL labeled data. We perform stage 1 using following ways -
  1. fine-tune a model on X^{train}, and infer on X^{test}
  2. perform ICL using exemplars from X^{train} for each test query in X^{test}
- *Translate-test*: Following (Chen et al., 2023b), we utilize *Self-fusion* using GPT-4, that takes input as target query, it's English translation and English translation's annotations, appended as a prompt, and outputs the annotated target query.[7]

A.2 Estimating confidence $\hat{y}^i_k$

For NLI task, the model always predicts a single-word label: 'neutral', 'contradiction' or 'entailment'. We simply apply softmax on the class logits for the predicted label to compute the confidence $\hat{y}^i_j$ (for i^{th} test instance).
In sequence labeling tasks, suppose for an input sentence having words: $\{w_1, w_2, ..., w_T\}$, the model predicts labels $\{o_1, o_2, ..., o_T\}$ with probabilities $\{\hat{p}_1, \hat{p}_2, ..., \hat{p}_T\}$. Let $LabelSet$ be $\{l_1, l_2, ..., l_N\}$. We compute confidence $\hat{y}_l$ for each label for a given test example as follows:

```
for k ← 1 to N do
    ŷ_k ← 0                    ▷ init each label's confidence
    c_k ← 0                    ▷ init each label's count
end for
for i ← 1 to T do
    for j ← 1 to N do
        if l_j == o_i then
            ŷ_j ← ŷ_j + p̂_i        ▷ Update ŷ_j
            c_j ← c_j + 1          ▷ increase counter
        end if
    end for
end for
for k ← 1 to N do
    ŷ_k = ŷ_k / c_k            ▷ average over all occurrences
end for
```

This outputs the confidence scores $\hat{y}_l$ for a given example, with those not predicted in a sequence assigned a value of 0.

A.3 Dataset Details

Family	Source languages	Source size
Germanic	{En,Is,De}	30000
African	{En,Am,Sw,Wo}	19788
American	{En,Es}	19998

Table 5: Size (No. of sentences) of Combined Source language datasets (En - English, Is - Icelandic, De - German, Am - Amharic, Sw - Swahili, Wo - Woloff, Es - Spanish)

B Prompt details

Prompts for the Named Entity Recognition (NER) and Part of Speech Tagging (POS) tasks are pre-

[4] https://azure.microsoft.com/en-in/products/ai-services/openai-service
[5] https://www.together.ai/
[6] https://pypi.org/project/gurobipy/

[7] We also tried Codec for translate-test, but could not reproduce the results reported in their paper for African languages (replicated avg. F1 = 60.5 v/s reported avg. F1 = 72).

Family	Test languages	Labels
Germanic	{Fo, Got, Gsw}	2370
African	{Hau,Ibo,Kin,Lug,Luo}	1100
American	{Aym,Gn,Nah}	501

Table 6: Size (No. of labels) of Target language datasets, *per language*, on average. (Fo - Faroese, Got - Gothic, Gsw - Swiss German, Hau - Hausa, Ibo - Igbo, Kin - Kinyarwanda, Lug - Luganda, Luo - Luo, Aym - Aymara, Gn - Guarani, Nah - Nahuatl)

sented in the tab separated format shown in B.0.2 and B.0.3 respectively.

Prompts for Natural Language Inference (NLI) initially used the framework in Ahuja et al. (2023) . To improve our performance, we changed the prompt to use Sclar et al. (2023)'s framework, where the authors performed an exhaustive search over tokens used for a prompt in order to find the prompt with optimal performance. This increased Macro F1 score by atleast 10% across all the tested languages. We use the same prompt across all models used in our experiments.

B.0.1 Natural Language Inference (NLI)

Task Description: You are an NLP assistant whose purpose is to solve Natural Language Inference (NLI) problems. NLI is the task of determining the inference relation between two (short, ordered) texts: entailment, contradiction, or neutral. Answer as concisely as possible in the same format as the examples below:
Input format:
Premise: {premise} , Hypothesis: {hypothesis} ,
Output format:
Answer: {output}
Verbalizer:
match the one-word response from the model (neutral, contradiction or entailment)

B.0.2 Named Entity Recognition (NER)

Task Description: Tag the following sentence according to the BIO scheme for the NER task, using the tags PER (person), LOC (location), ORG (organization) and DATE (date). Follow the format specified in the examples below:
Input format:
Sentence: w_1 w_2 ... w_T
Output format:
Tags:
w_1<TAB>o_1
w_2<TAB>o_2
...
w_T<TAB>o_T
Verbalizer:
Extract the sequence of labels $o_1, o_2, ...o_3$ from generated response.

B.0.3 Part of Speech (PoS) tagging

Task Description: Tag the following sentence according to the Part of Speech (POS) of each word. The valid tags are ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X. Follow the format specified in the examples below:
Input format:
Sentence: w_1 w_2 ... w_T
Output format:
Tags:
w_1<TAB>o_1
w_2<TAB>o_2
...
w_T<TAB>o_T
Verbalizer:
Extract the sequence of labels $o_1, o_2, ...o_3$ from generated response.

B.1 Verbalizer details for Tagging tasks

The verbalizer for tagging tasks requires the LLM to output the words as well as the associated labels. The LLM's output may not be perfect, as it may fail to generate all words or associate a label with each word. As a result, we find the *Longest Common Subsequence* between the words generated by the LLM and the words of the example. This is done using Dynamic Programming, as described in (Bergroth et al., 2000).

Once we have found the longest common subsequence, we assign the corresponding tags generated by the LLM to these words. If the tags are invalid, we assign a default tag (O for NER, and X for POS). Finally, for the words which don't have any tags associated with them, we assign the same default tag as before.

It is to be noted that in most cases, the sentence generated by the LLM perfectly matches the original sentence. For GPT-4, less than 1% of the words fell into the category of having an invalid tag generated, or not having the word generated.

B.2 Error Analysis

We investigate scenarios where SSP approach systematically fails compared to other methods. For NER, we find that ZGUL (fine-tuned LM) underpredicts the ‘DATE’ label. As a result, SSP almost never samples this label in stage 2 exemplars, hence hurting the performance for this label. For NLI task, we observe that in order to ensure label coverage, SSP samples the underpredicted label ‘neutral’ but while doing so, also ends up hurting the performance for ‘contradiction’ label (as seen in last plot of Figure 3).

B.3 Prompts for GSW Examples

The base SSP-SIM prompts for the GSW examples highlighted in Figure 6 are given below. Labels which are misclassified in the in-context exemplars are coloured in red, and the AUX labels which are to be flipped in the ablations are coloured in blue. It is interesting to note that examples 1 and 2 are similar, as example 1 is retrieved as an in-context exemplar for example 2.

B.3.1 Example 1

Tag the following sentence according to the Part of Speech (POS) of each word. The valid tags are ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X. Follow the format specified in the examples below:
Sentence: I main , das Ganze letscht Wuchä isch mier scho ächli iigfaarä .
Tags:
‘‘‘
I PRON
main VERB
, PUNCT
das DET
Ganze NOUN
letscht ADJ
Wuchä NOUN
isch AUX
mier PRON
scho ADV
ächli ADV
iigfaarä VERB
. PUNCT
‘‘‘
Sentence: Du gsehsch uus , wi wenn de nöime no hättisch z trinken übercho .
Tags:
‘‘‘
Du PRON
gsehsch VERB
uus PRON
, PUNCT
wi SCONJ
wenn SCONJ
de DET
nöime ADJ
no ADV
hättisch AUX
z PART
trinken VERB
übercho VERB
. PUNCT
‘‘‘
Sentence: Dir weit mer doch nid verzöue , di Wäutsche heige vo eim Tag uf en anger ufghört Chuttlen ässe .
Tags:
‘‘‘
Dir PRON
weit VERB
mer PRON
doch ADV
nid ADV
verzöue VERB
, PUNCT
di DET
Wäutsche NOUN
heige VERB
vo ADP
eim DET
Tag NOUN
uf ADP
en DET
anger ADJ
ufghört VERB
Chuttlen NOUN
ässe VERB
. PUNCT
‘‘‘
Sentence: es isch nämli echt usgstorbe gsi .
Tags:
‘‘‘
es PRON
isch AUX
nämli ADV
echt ADJ
usgstorbe VERB
gsi AUX
. PUNCT
‘‘‘
Sentence: Aso bini rächt uufgschmissä gsi und dem entschprächend fascht verzwiiflät .

	Ds	Gueten	isch	immerhin	gsi	,	dass	i	ungerdesse	söfu	müed	bi	gsi	,	dass	i	ändlech	ha	chönne	go	schlofe	.
CLT-SIM	DET	NOUN	AUX	ADV	VERB	PUNCT	SCONJ	PRON	ADV	VERB	ADJ	ADP	VERB	PUNCT	SCONJ	PRON	ADV	AUX	AUX	VERB	VERB	PUNCT
SSP-CLT-SIM	DET	NOUN	AUX	ADV	AUX	PUNCT	SCONJ	PRON	ADV	ADV	ADJ	ADP	AUX	PUNCT	SCONJ	PRON	ADV	AUX	AUX	PART	VERB	PUNCT
SSP-CLT-SIM (Half AUX->VERB)	DET	NOUN	AUX	ADV	AUX	PUNCT	SCONJ	PRON	ADV	ADV	ADJ	ADP	AUX	PUNCT	SCONJ	PRON	ADV	AUX	AUX	PART	VERB	PUNCT
SSP-CLT-SIM (All AUX->VERB)	DET	NOUN	VERB	ADV	VERB	PUNCT	SCONJ	PRON	ADV	ADV	ADJ	ADP	VERB	PUNCT	SCONJ	PRON	ADV	AUX	AUX	VERB	VERB	PUNCT
Gold	DET	NOUN	AUX	ADV	AUX	PUNCT	SCONJ	PRON	ADV	ADV	ADJ	AUX	AUX	PUNCT	SCONJ	PRON	ADV	AUX	AUX	PART	VERB	PUNCT

	I	cha	der	ihri	Telefonnummere	gä	,	de	nimmsch	mou	unverbindlech	Kontakt	uuf	.
CLT-SIM	PRON	VERB	DET	ADJ	NOUN	VERB	PUNCT	PRON	VERB	ADV	ADJ	NOUN	VERB	PUNCT
SSP-CLT-SIM	PRON	AUX	PRON	PRON	NOUN	VERB	PUNCT	PRON	VERB	ADV	ADJ	NOUN	ADP	PUNCT
SSP-CLT-SIM (Half AUX->VERB)	PRON	AUX	PRON	PRON	NOUN	VERB	PUNCT	PRON	VERB	ADV	ADJ	NOUN	ADP	PUNCT
SSP-CLT-SIM (All AUX->VERB)	PRON	VERB	PRON	PRON	NOUN	VERB	PUNCT	DET	VERB	ADV	ADJ	NOUN	ADP	PUNCT
Gold	PRON	AUX	PRON	DET	NOUN	VERB	PUNCT	ADV	VERB	ADV	ADJ	NOUN	PART	PUNCT

Figure 6: Label flips for CLT-SIM and SSP-SIM, for POS tagging in Swiss-German (gsw). Incorrect labels are marked in red. SSP-SIM ablations include flipping half/all of the AUX labels in the prompt to VERB labels. Gold labels are given for reference.

Tags:
‘‘‘
Aso ADV
bini AUX
rächt ADV
uufgschmissä VERB
gsi AUX
und CCONJ
dem PRON
entschprächend ADJ
fascht ADV
verzwiiflät VERB
. PUNCT
‘‘‘
Sentence: Der Ääschme wett nöd schaffe biin em .
Tags:
‘‘‘
Der DET
Ääschme NOUN
wett AUX
nöd ADV
schaffe VERB
biin ADP
em PRON
. PUNCT
‘‘‘
Sentence: Zerscht hends am Dani gsait , är söli dòch Hoochdütsch redä , das gängi denn grad gaar nöd , wenn är so redi , wiäner redi .
Tags:
‘‘‘
Zerscht ADV
hends PRON
am ADP
Dani PROPN
gsait VERB
, PUNCT
är PRON
söli AUX
dòch ADV
Hoochdütsch ADJ
redä VERB
, PUNCT
das PRON
gängi VERB
denn ADV
grad ADV
gaar ADV
nöd ADV
, PUNCT
wenn SCONJ
är PRON
so ADV
redi VERB
, PUNCT
wiäner PRON
redi VERB
. PUNCT
‘‘‘
Sentence: Isch das e Sach gsi , bis mer se gfunge hei gha .
Tags:
‘‘‘
Isch AUX
das PRON
e DET
Sach NOUN
gsi AUX
, PUNCT
bis SCONJ
mer PRON
se PRON
gfunge VERB
hei AUX

gha VERB
. PUNCT
“‘

Sentence: Ds Gueten isch immerhin gsi , dass i ungerdesse söfu müed bi gsi , dass i ändlech ha chönne go schlofe .
Tags:
“‘

B.3.2 Example 2

Tag the following sentence according to the Part of Speech (POS) of each word. The valid tags are ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X. Follow the format specified in the examples below:
Sentence: I ha ar Marie-Claire gseit , es sig mer chli schlächt und i mög jetz nümm liire .
Tags:
“‘
I PRON
ha AUX
ar PART
Marie-Claire PROPN
gseit VERB
, PUNCT
es PRON
sig AUX
mer PRON
chli ADV
schlächt ADJ
und CCONJ
i PRON
mög VERB
jetz ADV
nümm ADV
liire VERB
. PUNCT
“‘
Sentence: De Spanier hed de Kontakt vermettlet , d Rumäne sölled d Holländer ombrocht ha .
Tags:
“‘
De DET
Spanier NOUN
hed AUX
de DET
Kontakt NOUN
vermettlet VERB
, PUNCT
d DET
Rumäne NOUN
sölled AUX
d DET
Holländer PROPN
ombrocht VERB
ha AUX
. PUNCT
“‘
Sentence: Ds Gueten isch immerhin gsi , dass i ungerdesse söfu müed bi gsi , dass i ändlech ha chönne go schlofe .
Tags:
“‘
Ds DET
Gueten NOUN
isch AUX
immerhin ADV
gsi VERB
, PUNCT
dass SCONJ
i PRON
ungerdesse ADV
söfu VERB
müed ADJ
bi ADP
gsi VERB
, PUNCT
dass SCONJ
i PRON
ändlech ADV
ha AUX
chönne AUX
go VERB
schlofe VERB
. PUNCT
“‘
Sentence: Isch das e Sach gsi , bis mer se gfunge hei gha .
Tags:
“‘
Isch AUX
das PRON
e DET
Sach NOUN
gsi AUX
, PUNCT
bis SCONJ
mer PRON
se PRON
gfunge VERB
hei AUX
gha VERB

. PUNCT
"'
Sentence: De Dialäkt muess zu de Gschecht und zum Inhaut vonere Werbig passe .
Tags:
"'
De DET
Dialäkt NOUN
muess AUX
zu ADP
de DET
Gschecht NOUN
und CCONJ
zum ADP
Inhaut NOUN
vonere ADP
Werbig NOUN
passe VERB
. PUNCT
"'
Sentence: Mit der Zit hani mi mit mir säuber uf ei Schriibwiis pro Wort aafo einige .
Tags:
"'
Mit ADP
der DET
Zit NOUN
hani VERB
mi PRON
mit ADP
mir PRON
säuber ADJ
uf ADP
ei DET
Schriibwiis NOUN
pro ADP
Wort NOUN
aafo VERB
einige DET
. PUNCT
"'
Sentence: Mit all denä Wörter hani natürli nüt chönä aafangä .
Tags:
"'
Mit ADP
all DET
denä DET
Wörter NOUN
hani PRON
natürli ADV
nüt ADV
chönä VERB
aafangä VERB
. PUNCT
"'
Sentence: Aso bini rächt uufgschmissä gsi und dem entschprächend fascht verzwiiflät .
Tags:
"'
Aso ADV
bini AUX
rächt ADV
uufgschmissä VERB
gsi AUX
und CCONJ
dem PRON
entschprächend ADJ
fascht ADV
verzwiiflät VERB
. PUNCT
"'
Sentence: I cha der ihri Telefonnummere gä , de nimmsch mou unverbindlech Kontakt uuf .
Tags:
"'

C Source and Target Languages for each task

Code	Language
En	English
Am	Amharic
Sw	Swahili
Wo	Wolof
Hau	Hausa
Ibo	Igbo
Kin	Kinyarwanda
Lug	Luganda
Luo	Luo
Is	Icelandic
De	German
Fo	Faroese
Got	Gothic
Gsw	Swiss German
Es	Spanish
Aym	Aymara
Gn	Guarani
Nah	Nahuatl

Table 7: Languages and their codes

Model	Neu.	Ent.	Con.	Overall
mDeBERTa^{FT}	24.3	**72.7**	38.7	45.2
SSP(mDeBERTa^{FT})	**57.8**	46.5	51.5	52
(w/o Label)	35.3	43.8	68.5	49.2

Table 8: Labelwise Recall for fine-tuned model (mDeBERTa^{FT}) and ILP variants w. and w/o Label coverage (GPT-4-Turbo)

D NLI Analysis

We present an example of correct prediction made by SSP as compared to the version that doesn't ensure label coverage in Figure 8 (English translation in Fig. 9).

E Qualitative Analysis: SSP-SIM

We present the analysis for the gains obtained via SSP-SIM for Germanic POS in Figure 10. The confusion matrix difference between SSP-SIM and CLT-SIM suggests that the model misclassifies auxiliary verbs as verbs in CLT-SIM, and this is corrected in SSP-SIM. These errors are a consequence of the labels on the in-context exemplars the model receives, and not the tokens of the language itself.

We highlight this via the two Swiss-German POS examples in Figure 6. The misclassified verbs are corrected by SSP-SIM, and these labels are again misclassified when more than half of the labels in the in-context exemplars are corrupted.

F Data Contamination Analysis

Following Ahuja et al. 2023, we conduct contamination tests on test datasets for our target languages. We perform the following tests:

- Dataset Card filling: Generate dataset card (supported languages, dataset description, #instances in each split, etc.)
- Completion: Given a few words, complete the sentence and their labels, and
- Generation using first few instances: Given first K instances (K=5) in the dataset, generate next few instances following them.

We observe negligible contamination as depicted in table 8. The 40% accuracy for Quechua was a result of all the labels passed for the exemplars being entailment labels. As a result, the model repeated the same label for all the other examples, giving a 40% accuracy. *Following these results, to prevent any chance of contamination, we remove Quechua from our evaluation dataset.*

Stage 1:

Field	Sentence in target language (Gn)	En translation (GPT-4-turbo stage 1)	En translation (gold)
Premise	Péva ha'e, eikuaáma, emaña, neapañuâima.	That is, you know, look, you're not in trouble anymore. (Error)	This is it, you know, look, you're in trouble.
Hypothesis	Ikatu reñemosê ko tetâgui.	You may be expelled from the country.	You can be deported from this country.
Label	neutral	contradiction	neutral

Stage 2:

Premise: Ha upéichako, akârasy memete, ja'ekuaa ko árape arekopaite mba'érepa cheakârasy hağua, nde nereikuaái mba'éichapa ojejapo peteî mba'e ha he'i ndéve hikuái: péina, ejapo. , Hypothesis: Ko árape ojerure cheve ajapo hağua tembiapo che katupyrývape. , Answer: neutral

Premise: upéichaite, ha'eséko che ko'â léi pyahukuéra rehe hasy ko'ága. , Hypothesis: Upevarehete, umi temimoîmby pyahu reheve, ko'ága hasyve. , Answer: neutral

Premise: Péva ha'e, eikuaáma, emaña, neapañuâima. , Hypothesis: Ikatu reñemosê ko tetâgui. , Answer: **contradiction**

Premise: Néi, ñağuahêniko ko'âvape, peteî arapokôindýpe oîha mokôi térâ mbohapy aviô ha ndoikuaái moôpa ovevéta. , Hypothesis: Hetave aviô oîramo upéva apañuâima. , Answer: neutral

Premise: Ha aha hógape ha ahenói upe papapy oje'evakue ahenói hağua ağuahê vove upépe. , Hypothesis: Ahenói upe papapýpe ağuahêvo hógape. , Answer: neutral

Premise: Pe kuñataî ikatúva chepytyvô oî amo táva mboypýri. , Hypothesis: Upe mitâkuña chepytyvôtava oî águi 5km hápe. , Answer: neutral

Premise: Ha'e ou, oipe'a okê ha chemandu'a amaña che rapykuévo ha ahecha hova, ahechakuaa ndaha'éihague upe oha'arôva. , Hypothesis: Oñeha'â ani hağua roñeñandu vai katu roikuaa orekáusa iñapañuâiha. , Answer: neutral

Premise: Ajeíma upe oje'évagui, aipo peteî kuimba'e oikutihague hembirekópe ojuka peve ha'e oke rupi ambue kuimba'e ndive, hembirekokue ha'eséko, nde reikuaa mba'érepa añe'ê. , Hypothesis: Peteî kuimba'e ojuka hembirekópe oñeno rupi ambue kuimba'e ndive ramoite ojepoi rire chupe upe haguetére. , Answer: contradiction

---

Premise: Ha'ese che ha'ekuéra orekoha amo 5 ñemoñare rupinte. Peteîva omanova'ekue. , Hypothesis: Peteîva umi 5 apytépe omano. , Answer: **entailment**

Figure 7: Stage 1: Impact of translation error on translate-test performance in Gn language for NLI task. Stage 2: GPT-4-turbo correctly predicts the label for given NLI query in Gn language, even though the 3rd exemplar is incorrectly labeled. This depicts the SSP's robustness to stage 1 noise due to errors in translation (NLLB) model.

Premise: Ah, huk chaypi allinqa apakurqa allin qawasqayqa paniypa ñawpaq yuyariyninmi, chaypas hina hipa pampapim karqa.
Hypothesis: Yuyaruniqa hipa pampapi huk ima apakusqantam.
Answer: entailment

Premise: Yaykuykuptiykuqa punkukunaqa wichqasqam kachkarqa.
Hypothesis: Punku wichqasqa kachkaptinpas yaykurqanikum.
Answer: entailment

Premise: Yanapawaqniy atiq sispasmi hatun llaqtapa waklawninpiraq tiyan.
Hypothesis: Yanapawaqniy warmi warman 5 millas nisqan karupiraq tiyan.
Answer: **neutral**

Premise: Manam mayman risqanta yacharqanikuchu.
Hypothesis: Mayman risqantam yacharqaniku.
Answer: entailment

Premise: Chayna kaptinqa hamutachkanim huktapiwan Ramonawan rimariyta.
Hypothesis: Ramonawanmi huktapiwan rimarqani.
Answer: entailment

Premise: Ripukusqañam hinaspam amaña llakikunaypaq niwarqa.
Hypothesis: Ama llakikunaytam niwarqa.
Answer: entailment

Premise: Ichapasyá huk kaq mana yachasqaymanta hamun ichaqa
Hypothesis: Apurawtam hamun, ichaqa maymanta hamusqanta yachanim.
Answer: entailment

Premise: Locust Hill oh awriki, ari, kusa
Hypothesis: Locust Hill nisqaqa allinmi.
Answer: contradiction

Premise: Oh, payllam isqun iskay iskayraq regulador nisqapi inyecciónta qinaq karqa.
Hypothesis: Martes punchawtam inyector nisqata hinarqani.
Answer: neutral

Premise: Ah, huk chaypi allinqa apakurqa allin qawasqayqa paniypa ñawpaq yuyariyninmi, chaypas hina hipa pampapim karqa.
Hypothesis: Yuyaruniqa hipa pampapi huk ima apakusqantam.
Answer: entailment

Premise: Yaykuykuptiykuqa punkukunaqa wichqasqam kachkarqa.
Hypothesis: Punku wichqasqa kachkaptinpas yaykurqanikum.
Answer: entailment

Premise: Manam mayman risqanta yacharqanikuchu.
Hypothesis: Mayman risqantam yacharqaniku.
Answer: entailment

Premise: Chayna kaptinqa hamutachkanim huktapiwan Ramonawan rimariyta.
Hypothesis: Ramonawanmi huktapiwan rimarqani.
Answer: entailment

Premise: Manam pachay karqachu ima kaqpas ruranaypaq.
Hypothesis: Mana pacha llapan qinanaypaq haypawarqachu
Answer: entailment

Premise: Ripukusqañam hinaspam amaña llakikunaypaq niwarqa.
Hypothesis: Ama llakikunaytam niwarqa.
Answer: entailment

Premise: Ichapasyá huk kaq mana yachasqaymanta hamun ichaqa
Hypothesis: Apurawtam hamun, ichaqa maymanta hamusqanta yachanim.
Answer: entailment

Premise: Locust Hill oh awriki, ari, kusa
Hypothesis: Locust Hill nisqaqa allinmi.
Answer: contradiction

Premise: Oh, payllam isqun iskay iskayraq regulador nisqapi inyeccióntа qinaq karqa. Hypothesis: Martes punchawtam inyector nisqata hinarqani.
Answer: **contradiction**

Figure 8: Correct case of 'Neutral' detected by ILP (left), while 'w/o label' variant misses it (right). We note that exact one 'neutral' class has been sampled by ILP, while no 'neutral' is sampled in 'w/o label' version.

Premise: Ah, one there good thing took away is my best view is my sister's old memory, which was also on the same hip floor.
Hypothesis: I remember something carrying on the floor.
Answer: entailment

Premise: The doors were locked when we entered.
Hypothesis: We got in even though the door was locked.
Answer: entailment

Premise: The sister who can help me lives just on the other side of the big city.
Hypothesis: My assistant lives 5 miles away.
Answer: **neutral**

Premise: We didn't know where he was going.
Hypothesis: We knew where he was going.
Answer: entailment

Premise: In that case I'm coming up with another conversation with Ramona.
Hypothesis: I talked to Ramona again.
Answer: entailment

Premise: He had left and told me not to worry.
Hypothesis: He told me not to worry.
Answer: entailment

Premise: Maybe it comes from something I don't know though
Hypothesis: It comes quickly, but I know where it comes from.
Answer: entailment

Premise: Locust Hill oh yeah, yeah, great
Hypothesis: Locust Hill is good.
Answer: contradiction

Premise: Oh, he was the only one who still injected nine seconds into the regulator.
Hypothesis: I applied the injector on Tuesday.
Answer: neutral

Premise: Ah, one there good thing took away is my best view is my sister's old memory, which was also on the same hip floor.
Hypothesis: I remember something carrying on the floor.
Answer: entailment

Premise: The doors were locked when we entered.
Hypothesis: We got in even though the door was locked.
Answer: entailment

Premise: We didn't know where he was going.
Hypothesis: We knew where he was going.
Answer: entailment

Premise: In that case I'm coming up with another conversation with Ramona.
Hypothesis: I talked to Ramona again.
Answer: entailment

Premise: I didn't have time to do anything.
Hypothesis: I didn't have enough time to cover everything
Answer: entailment

Premise: He had left and told me not to worry.
Hypothesis: He told me not to worry.
Answer: entailment

Premise: Maybe it comes from something I don't know though
Hypothesis: It comes quickly, but I know where it comes from.
Answer: entailment

Premise: Locust Hill oh yeah, yeah, great
Hypothesis: Locust Hill is good.
Answer: contradiction

Premise: Oh, he was the only one who still injected nine seconds into the regulator. Hypothesis: I applied the injector on Tuesday.
Answer: **contradiction**

Figure 9: English translations of Exemplars shown in Fig. 8

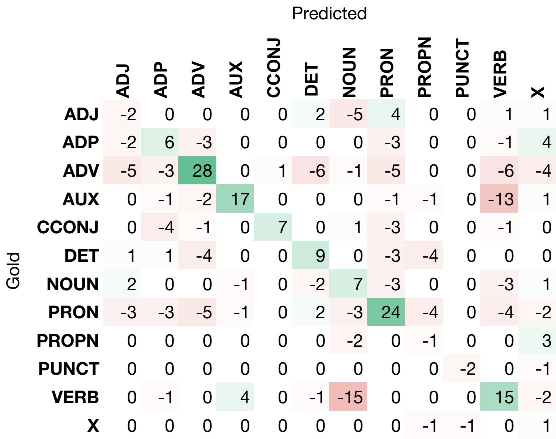

Figure 10: Difference in confusion matrices between similarity-based SSP Stage 1 and Stage 2 for the POS task, summed across all languages (tags with less than 100 instances have been omitted). The increase in correct tags is visible along the diagonal, and misclassifications between VERB and AUX tags / NOUN and VERB tags have also improved.

Task	Card Filling	Completion	Few-Shot Generation
NER	Didn't predict correct languages; no split sizes generated	No match found	NA
POS	predicted 33 languages, but doesn't contain any of our target languages	No match found	NA
NLI	predicts 3 languages, of which only one matches with our target language (Quechua); wrong test split size	Refuses to generate for 3 out of 4 target languages, except for Quechua - for which it predicts 100% of the tokens wrong and only 40% labels correctly (out of 10 instances)	Repeats the premise of last instance, copies the premise string to hypothesis as well (No match detected)

Table 9: Results of Contamination Study

Pool Size	Hau	Ibo	Kin	Lug	Luo	Avg.	Fo	Got	Gsw	Avg
8	68.8	80.2	67.6	75.6	53.8	69.2	80.6	54.8	80.9	72.1
32	70.1	80.8	71.3	74.8	53.9	70.2	81.8	62.8	82.2	75.6
64	70.5	79.7	**72**	**77**	**55.2**	70.9	82.1	63.4	**85.7**	77.1
100	**71.2**	**82.4**	71.4	75.4	55.1	**71.1**	**82.2**	**71.5**	85.6	**79.8**

Table 10: Language-wise F1 scores for African NER and Germanic POS as a function of candidate pool size in SSP