LLMs as Polyglots: On the Emergence of Multilingual Inference Abilities

A0220603Y, A0219702J, A0214563J, A0223837B, A0225744A
Group 19
Mentored by Yisong Miao
{e0555795, e0550360, e0588286, e0564878, e0527279}@u.nus.edu

Abstract

Cross-lingual transfer refers to the ability of language models to learn from a set of well-documented languages and perform tasks in unseen, possibly low-resource languages. The community is yet to study the efficacy of LMs on logical inference in a cross-lingual setting. In this work, we make two contributions: we release **Multilingual e-SNLI** (MLe-SNLI), a cross-lingual dataset built on top of e-SNLI (Camburu et al., 2018) comprising of samples translated to Spanish, Dutch, French, and German grounded by their similarity to English. We also propose a new prompting paradigm to study this emergence of LMs by finetuning Flan-T5-Large (Chung et al., 2022) on MLe-SNLI and empirically demonstrate the emergence of multilingual inference skills by comparing it to a zero-shot Flan-T5-Large solely pretrained on ∼1.8K English tasks. Our experiments demonstrate that a finetuned Flan-T5-Large achieves significantly higher classification and explanation accuracies compared to a zero-shot Flan-T5-Large on an unseen language. Specifically, our model achieved a classification accuracy of 75% and an explanation accuracy of 51%, whereas the zero-shot model scored 64% and 40% respectively.

1 Introduction

Cross-lingual transfer is the ability of models to learn tasks in different, possibly high-resource languages and apply the acquired knowledge in a zero-shot or few-shot manner in an unseen, possibly low-resource language. The community has shown active interest in this area given the emergence of Transformers and their emergent abilities exhibited through scaling. Such areas mainly involve question-answering (Lee and Lee, 2019; Zhou et al., 2021; Pouran Ben Veyseh, 2016), text classification (Kozhevnikov and Titov, 2013; Ahmad et al., 2021; Chalkidis et al., 2021), and reasoning (Pfeiffer et al., 2020; K et al., 2021; Tikhonov and Ryabinin, 2021).

This cross-lingual reasoning task has roots in language learning theory. Hyltenstam (2016) and EFEOTOR et al. (2022) show that polyglots have an easier experience learning additional languages because of existing syntactical and semantical similarities in the grammar. Furthermore, Hakuta and Diaz (2014) show that bilingual children have greater capacities of working memory and cognitive reasoning, allowing them to perform inference at a higher level than monolingual children. We aim to validate this theory for LMs owing to recent trend in the community to study emergent abilities (Wei et al., 2022b).

To our knowledge, we are the first to study cross-lingual transfer in the context of natural language inference tasks involving both classification label (entailment, contradiction, neutral) and explanation, whereas existing benchmarks like Conneau et al. (2018); Hu et al. (2020); Williams et al. (2018); Ponti et al. (2020) solely look at either the former or the latter separately. As such, we provide our own baselines specifically geared towards this multilingual inference task, along with a comprehensive ablation[1]. We hope to pave the way for showcasing the abilities of LMs to do inference for low-resource languages in the world, for which data is scarce and difficult to obtain.

To summarise, we introduce the following key contributions:

- We release Multilingual e-SNLI (MLe-SNLI), a multilingual inference corpus built on top of e-SNLI (Camburu et al., 2018), on Hugging Face's Hub, additionally available in Spanish, Dutch, French, and German[2]. This dataset is dedicated to finetuning large language mod-

[1]Given compute restrictions, we were unable to train very large models (# params > 3B).

[2]MLe-SNLI (with descriptions and usage instructions) can be found at https://huggingface.co/datasets/rish16/MLe-SNLI and our code at https://github.com/rish-16/cs4248-project.

els to study the emergence of cross-lingual transfer abilities.

- We empirically demonstrate the emergence of cross-lingual logical inference skills by finetuning Flan-T5-Large (Chung et al., 2022) on MLe-SNLI and outperforming an off-the-shelf (zero-shot) Flan-T5-Large **on an unseen but similar** language – something not observed in smaller-scale models.

2 Preliminaries and Related Work

Our work is primarily related to T5 and its fine-tuned versions, which are cutting-edge models used for explanation generation and multi-task training. Additionally, we examine the emergence of models' cross-lingual transfer capabilities under scaling laws.

2.1 Text-to-Text Transfer Transformer (T5)

Text-to-Text Transfer Transformer (T5) is a text-to-text model that has been pre-trained to follow the sequence-to-sequence framework(Raffel et al., 2020). This model is designed to take an input sequence and generate an output sequence by maximizing the probability of the output given the input and the previous output tokens, which is called an autoregressive modelling approach.

Because of its encoder-decoder architecture, T5 exhibits state-of-the-art performance in natural language inference (NLI) tasks and is suited for multi-task framework (MT) (Hase et al., 2020). Specifically, T5 is widely used as a baseline model for reasoning tasks (Li et al., 2022) and multi-tasks such as MT-Re (Hase et al., 2020) and MT-Ra (Camburu et al., 2018) that require both labels and explanation as output due to its excellent balance between performance and computational requirements (Narang et al., 2020). In our case, we fine-tune our model based on the powerful T5 model as both labels and quality explanations are required by the e-SNLI dataset.

2.2 Finetuning Language Models (Flan)

Although LLMs such as T5 have demonstrated impressive performance in few-shot learning, they have not been as successful in zero-shot learning for tasks like natural language inference and reading comprehension (Brown et al., 2020). To address this limitation, a new approach called Flan (**F**inetuning **lan**guage models) was developed by finetuning original LLMs using datasets expressed through natural language instructions (Wei et al., 2022a). Unlike traditional LLMs such as GPT-3 and T5, which format prompts to resemble data that the model was pre-trained on, Flan formats prompts as natural language instructions to improve models' zero-shot performance on unseen tasks (Wei et al., 2022a). Flan-T5-Large, in particular, was observed to outperform T5-Large in a variety of tasks, including MMLU, BBH, TyDiQA, and MGSM, achieving an 18.8-point higher normal accuracy score (Chung et al., 2022).

However, the Flan-T5 family cannot transfer its multilingual ability to reason about or handle an unseen language. As a result, we fine-tune Flan-T5 on MLe-SNLI to unleash its ability to transfer learning to a new language.

2.2.1 Chain of Thought Prompting

The Chain of Thought (CoT) Prompting technique is used to enhance language models by generating intermediate reasoning steps leading to the final answer of a multi-step problem (Wei et al., 2023). However, experiments have shown that finetuning only on non-CoT examples can significantly reduce performance on CoT tasks (Chung et al., 2022). To address this issue, Flan involves joint finetuning on both non-CoT and CoT data. The used CoT data includes tasks such as multi-hop reasoning (Geva et al., 2021) and natural language inference (Camburu et al., 2020). This joint finetuning results in better CoT performance while maintaining performance on non-CoT tasks. However, we show the contrary in this report in Section 6.1.

2.3 Cross-lingual Transfer

Given the relative nascency of cross-lingual transfer, there are very few benchmark tasks that help study this area. Notable among them is XNLI (Conneau et al., 2018) which is a multilingual dataset comprising of premise-hypothesis pairs that must be classified into three categories (entailment, contradiction, neutral). Similarly, MultiNLI (Williams et al., 2018) extends SNLI (Bowman et al., 2015) into multiple genres spanning entertainment, public reports, telephone records, and more. The task is also to classify a premise and hypothesis as entailing or contradicting each other. Under the cross-lingual umbrella, XCOPA (Ponti et al., 2020) is a multilingual commonsense reasoning dataset involving a natural language premise and two alternatives where the task is to choose which alternative aligns with the premise the best. MLe-SNLI dif-

Language	Premise/Hypothesis	Label
English (original)	Children smiling and waving at camera. They are smiling at their parents.	neutral
Spanish	Niños sonriendo y saludando a la cámara. Están sonriendo a sus padres.	neutral
French	Enfants souriant et agitant à la caméra. Ils sourient à leurs parents.	neutral
Dutch	Kinderen lachen en zwaaien naar de camera. Ze lachen naar hun ouders.	neutral
German	Kinder lächeln und winken vor der Kamera. Sie lächeln ihre Eltern an.	neutral

Table 1: A randomly chosen premise-hypothesis pair from e-SNLI (Camburu et al., 2018) translated into Spanish, French, Dutch, and German. Samples like these altogether form MLe-SNLI.

fers from these in that, aside from the multilingual premise-hypothesis pairs, we provide explanations in the associated languages as well in order to study cross-lingual emergent abilities of LMs to justify inferences made in different languages.

2.4 Scaling Laws and Emergent Abilities

Scaling language models along the size of the model, training data, and computational resources has been observed to improve their performance (Kaplan et al., 2020; Hoffmann et al., 2022). *Emergence* is defined as special abilities or skills that large models possess due to large compute and data size, which smaller models (with lesser compute or data) cannot possess (Wei et al., 2022b). Scaling Laws (Kaplan et al., 2020) are the best way to get a model to exhibit emergent abilities, with such emergence not easily ascertained via extrapolating performance on smaller-scale models, requiring large-scale training runs. For instance, Rae et al. (2022) showed that small Gopher models performed no better than random on the TruthfulQA benchmark until being scaled up to $\sim$ 280B parameters. A similar emergence of inference ability in German for large models is also observed when finetuning on MLe-SNLI, as reported in Sections 4 and 6.

3 MLe-SNLI: A Multilingual NLI Dataset

To study the effect of language models on logical inference in a cross-lingual setting, we introduce MLe-SNLI, a multilingual version of e-SNLI, as there are no similar datasets around.

3.1 Data Generation

The English Corpus. The English portion of MLe-SNLI is from e-SNLI (Camburu et al., 2018). The original e-SNLI dataset contains around 570K premise-hypothesis pairs, among which $\sim$ 9.8K are test pairs.

Translating the corpus. Previous research has demonstrated that the transfer of learning between language models is most effective when the languages in question are topologically similar (Pires et al., 2019). As a result, in our current study, we have identified and selected four additional languages that exhibit the greatest structural similarities to English (**en**), namely Spanish (**es**), Dutch (**nl**), French (**fr**), and German (**de**) (Williams and Elizabeth, 2022). We utilize Hugging Face's MarianMT (Junczys-Dowmunt et al., 2018), a potent neural machine translation (NMT) model to translate English into these four languages. MarianMT models have been proven to give quality translation for a wide range of languages (Liu et al., 2021). Due to limited compute, we select 100K data points from the training dataset and all $\sim$ 9.8K from the test data points for translation. For all languages, on average, the premises are twice as long as the hypotheses. Additionally, the explanation for neutral labels is longer than that of contradiction and entailment labels(See Table 2).

To ensure the preservation of the original context, we translate the premises and hypotheses separately. And the label is simply copied from the English source text. Table 1 provides a random developmental example of this approach.

Subsequently, we have compared the performance of the finetuned Flan-T5-Large model with

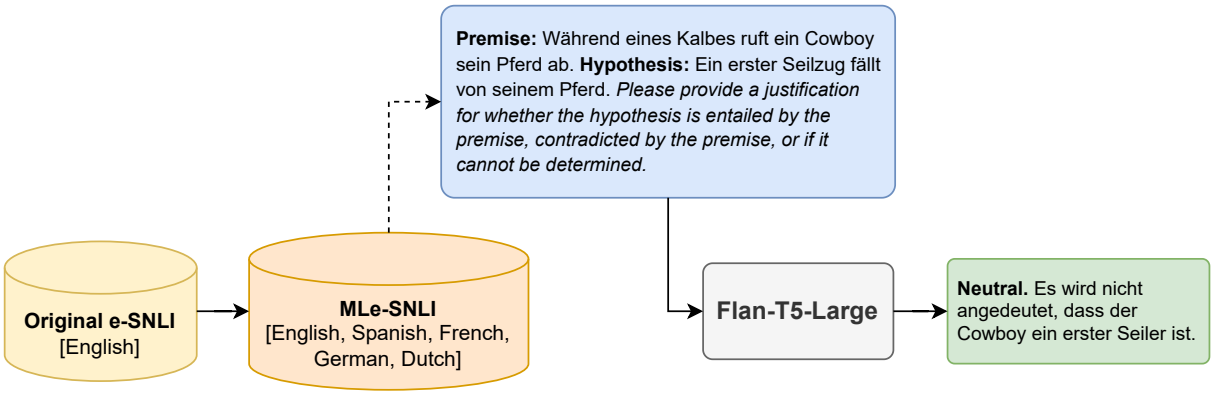

Figure 1: An overview of finetuning Flan-T5 on MLe-SNLI: a prompt including the premise and hypothesis is crafted and passed into the model. The model is instruction-tuned to specifically output the classification first followed by the explanation, which is separately extracted for evaluation. In this work, the German split is the hold-out task while all other languages are used for finetuning the Flan-T5.

	en	nl	fr	es	de
Premise	12.8	12.3	13.4	13.7	12.0
Hypothesis	7.4	7.1	7.6	7.8	6.8
Neutral	14.0	13.7	15.0	14.6	13.5
Contradiction	12.6	11.8	13.8	13.0	11.7
Entailment	11.5	11.2	12.2	12.1	11.1

Table 2: Average number of tokens (words) for explanations in the MLe-SNLI corpus for each language (top) as well as a breakdown for each category (bottom).

the off-the-shelf Flan-T5-Large on the German split to demonstrate that training in multiple languages can enhance the model's reasoning ability when working with an unseen but related language; see Figure 1 for the whole process.

4 Methodology

Task Statement. Given a premise `prem` and a hypothesis `hypo`, both in natural language, an instruction-tuned model should output a classification label `l` $\in$ {`entailment`, `contradiction`, `neutral`} and a natural language explanation `expl` justifying this label `l`.

4.1 Bi-level Evaluation

As the model outputs a classification label `l` as well as an explanation `expl`, we evaluate the model separately on both fronts. We make use of instruction finetuning (Wei et al., 2022a) to force the model to generate outputs in the format of classification followed by the explanation, which are extracted for downstream evaluation, as shown in Figure 1; this is done by modelling the finetuning prompt as follows:

"Premise: ⟨prem.⟩
Hypothesis: ⟨hypo.⟩

Please provide a justification for whether the hypothesis is entailed by the premise, contradicted by the premise, or if it cannot be determined."

The model was instruction-tuned to specifically generate outputs in the following format:

"⟨l⟩. ⟨expl.⟩"

As we are using autoregressive transformers that perform next-token prediction via masking, the explanations are inherently label-conditioned. Unless stated otherwise, we report a pair of accuracies for all models: classification label accuracy, Cls %, and explanation accuracy, Expl %. Classification label accuracy is trivially computed by extracting the generated label at the start of the output. Explanation accuracy is computed by manual inspection of 50 randomly-selected outputs from the test set and check if the generated explanation matches the ground truth or has any semblance of meaning in the context of the premise and hypothesis[3]. Furthermore, as the samples are evenly distributed amongst entailment, contradiction, and neutral, we use simple accuracy as a representative measure of model performance, as is done in the literature.

[3]We handed over these 50 samples to three unbiased peers (with no conflict of interest) who collectively voted on whether an explanation contextually makes sense. They had access to Google Translate to verify and validate the outputs.

4.2 Training Setup

Given computational constraints, we employ Mixed Precision training (Micikevicius et al., 2018). The pretrained models, specifically the Flan-T5 family, were trained using FP32 while our finetuning process made use of a mixture of FP32 and FP16. In situations where FP16 was not possible, we default to BF16 instead. If not stated otherwise, the Flan-T5 models were finetuned for 3 epochs on MLe-SNLI with 10K samples per language. We use AdamW (Loshchilov and Hutter, 2019) as the optimiser with $\alpha = 5 \times 10^{-5}$. Finetuning was mostly done on an A100 GPU using DeepSpeed (Aminabadi et al., 2022) as the accelerator. We use the SentencePiece tokenizer (Kudo and Richardson, 2018), specifically the version used to pretrain Flan-T5, given its versatility in multilingual settings. We found that padding the sentences was useful in adapting to the different context lengths used by different models.

5 Results and Discussion

We study the zero-shot and few-shot capabilities of logical inference in an unseen language. We pick German because the Flan-T5 family of models were not trained on many German tasks beyond translation and gender classification from the MMLU benchmark (Hendrycks et al., 2021). Regardless, we believe our finetuning method can be extended to most languages that may or may not be low-resource. We empirically show the emergence of cross-lingual inference abilities in models finetuned on different but similar languages (across all model sizes), as opposed to off-the-shelf models solely relying on pretraining; see Table 3 for Cls % and Table 4 for Expl %.

Due to the computational constraints, we were unable to further investigate the impact of using more data on only English and Dutch.

Flan-T5	Zero-shot	Finetuned
Small (80M)	**0.44**	0.42
Base (250M)	0.55	**0.66**
Large (750M)	0.64	**0.75**

Table 3: Larger finetuned models exhibit better cross-lingual transfer of inference skills than zero-shot models and smaller finetuned models. Classification accuracy (Cls %) reported.

Specifically, the models shown in Table 3 and Table 4 were fine-tuned with only English and Dutch, using 10K samples per language. English and Dutch were selected as it was shown in the Ablations (Section 6) later on, to have the best improvement in Expl %. As seen in Table 4, the fine-tuned Flan-T5-Large model outperforms the zero-shot model, while the fine-tuned Flan-T5-Base model did not perform better and the fine-tuned Flan-T5-Small model performed worse. From this observation, we conclude that there is an emergence of cross-lingual inference ability that only appears in the fine-tuned Large model while not in smaller models. Some examples of such emergence are provided in Appendix A.3.

Flan-T5	Zero-shot	Finetuned
Small (80M)	**0.08**	0.06
Base (250M)	0.26	**0.26**
Large (750M)	0.45	**0.51**

Table 4: Larger finetuned models exhibit better cross-lingual transfer of inference skills than zero-shot models and smaller finetuned models. Explanation accuracy (Expl %) reported.

On fluency and inferential skills. The zero-shot Flan-T5, across model sizes, already possesses a working knowledge of the languages chosen in this study but may not necessarily be able to use this language to perform logical inference in said languages, especially when transferring that skill to unseen languages. However, we only report these results without *proving* if the gain in classification and explanation accuracies is due to the additional finetuning on similar languages.

It is challenging to decouple knowledge of a language with the ability to infer in this language to demonstrate that the emergent abilities were born of the finetuning. As such, we simply show and ablate this emergence of cross-lingual transfer for NLI and hope the community can shed light on how to better perform this decoupling of abilities, which we leave for future work.

6 Ablations

Our ablation study aims to evaluate the impact of several variables on classification and explanation accuracy. Specifically, we investigate the effect of Chain of Thought Prompting, data size, finetuning language type, and the number of finetuning languages. By measuring the performance of our model under different conditions, we hope to gain insight into the most effective strategies for improv-

ing accuracy in this context.

6.1 Chain of Thought (CoT) Prompting

While it is natural to believe CoT Prompting would improve model performance, we show empirically, across all model sizes, that **CoT Prompting (without exemplars) does not help**; see Figure 2. We trained all the models below using 5K samples per language, using all 4 languages (i.e. Dutch, English, French and Spanish). We posit that the inference problem we are solving requires a different form of explicit reasoning that CoT does not elicit.

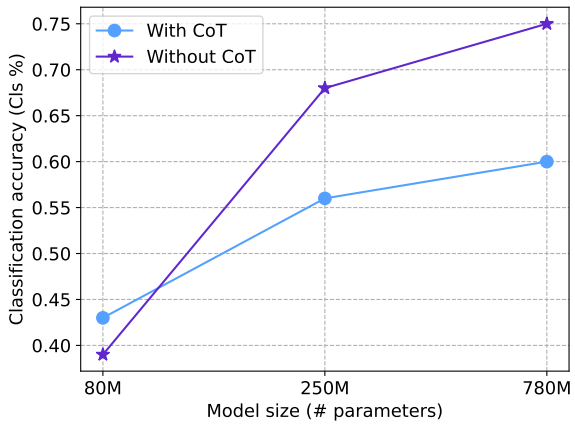

Figure 2: Chain of Thought Prompting (without exemplars) does not improve cross-lingual transfer to unseen German split across model sizes.

Suppose a sample has a premise, `prem`, and hypothesis, `hypo`. The new instruction-tuning prompt has an additional message to "think step-by-step" prepended (in green) as follows:

> "Answer whether the premise entails or contradicts the hypothesis or if it's impossible to tell. Let us think step-by-step.
>
> Premise: ⟨prem⟩.
> Hypothesis: ⟨hypo⟩."

The tasks for which CoT was used in Wei et al. (2023) and Chung et al. (2022) (arithmetic and logic puzzles, specifically) required multiple steps of reasoning for the tasks considered compared to NLI task in this work that often requires a single hop in reasoning. In conjunction, the explanations given in e-SNLI, and hence MLe-SNLI, are simplistic and do not feature iterative leaps in logic to derive the final answer, which may be a probable cause for poorer performance compared to non-CoT counterparts[4].

[4]We also tried prompts with exemplars, as done in Wei et al. (2023). We do not report the associated performance on the German split but the prompt can be found in Appendix A.2.

6.2 Scaling Data Size

In this section, we focused on investigating how the emergence of cross-lingual inference ability is affected by scaling the sample count per language when the model is trained using all four languages. We observed that **there is a positive correlation between the emergence and the size of the finetuning set** when it is increased; see Table 5. While the classification accuracies did not increase with data size, our model did show improvement in terms of its ability to provide accurate explanations for its predictions. This finding suggests that the increase in data size may have helped the model better understand the semantic relationships between words, leading to emergent properties in its performance. This observation supports the finding of Kaplan et al. (2020) who demonstrated that model size and dataset size should be scaled proportionally to achieve good performance. However, due to limitations in computational resources, we were unable to investigate this further and we leave it for future research.

Per-language Samples	Cls %	Expl %
5K	0.75	0.20
10K	0.75	0.31
30K	0.75	0.39

Table 5: Effect of number of per-language samples on performance on unseen German split. An increase in data size leads to an increase in explanation accuracy.

6.3 Type of Finetuning Languages Chosen

Further investigation was conducted to examine how the selection of language impacts the model's ability to make inferences. The model was finetuned on each language with 20K samples. **We observe that the classification accuracies of the models finetuned with English and Dutch are higher**, while the models finetuned with French and Spanish have lower accuracies; see Table 6. These results align with the fact that French and Spanish are Romance languages, whereas English, Dutch, and German are Germanic languages with high mutual intelligibility (Gooskens et al., 2018).

6.4 Number of Finetuning Languages

Here, we vary the number of different languages part of our finetuning set and continue to evaluate

Language Type	Expl %
English	0.74
Spanish	0.72
French	0.72
Dutch	0.74

Table 6: Effect of number of type of language on performance on unseen German split.

on the unseen German samples. **We observe constant Cls % but a drop in Expl % when more dissimilar languages (to German) are used**. Based on the results obtained in Section 6.3, we incrementally add Dutch, English, French and Spanish to the mixture as shown in Table 7. The order was chosen such that languages closer to German are added first. We also add the performance of the zero-shot model as a baseline.

# languages	Expl %	Expl %
0 (zero-shot)	0.64	0.45
1	0.74	0.35
2	0.75	0.51
3	0.75	0.37
4	0.75	0.31

Table 7: Effect of finetuning on more languages during inference time on the unseen German split.

We speculate that the drop in Expl % when fine-tuning on more than 2 languages is due to the inclusion of dissimilar languages in fine-tuning, resulting in worse performance when generating an explanation. There is also a noticeable drop for the Expl % from the zero-shot model to the model fine-tuned with 1 language, however, we attribute this to the model learning German through an intermediary language (Dutch) that is highly similar. This created a shift in language from English to Dutch when it comes to the generation of an explanation, which results in the initial decrease in performance. This is supported by the explanations generated by the languages, where all fine-tuned models shifted to a mix of Dutch and German as a medium to generate explanations:

"Premise: Ein Mann in einem Boomlift-Eimer schweißt.
Hypothesis: Ein Mann arbeitet.

Please provide a justification for whether the hypothesis is entailed by the premise, contradicted by the premise, or if it cannot be determined."

Zero-Shot: "A man in a boomlift-eimer is a man who works. The answer: yes."
NL: "Als ein mann in einem boomlift-eimer schweißt, arbeitet er."
NL+EN: "Als een mann in een boomlift-eimer schweißt, betekent dat hij werkt"
NL+EN+FR: "Als ein mann in ein boomlift-eimer schweißt, so arbeitet er."
NL+EN+FR+ES: "Als ein mann in ein boomlift-eimer schweißt, arbeitet er."

7 Limitations and Future Work

Mixed outputs. When training on a multilingual corpus, we observe the tendency of the model to speak in a mix of languages. For example, given a German premise and hypothesis, the generated output from the T5 is in *Denglisch* (shown in green below)[5]. While the explanation may be correct, it would be desirable if the explanations are either fully in English or German (or the unseen, low-resource language in question). We leave the evaluation of prompt phrasing for future work. There are words invariant to language changes (like "Cowboy" in Figure 1) which have to be dealt with care.

[5] *Denglisch* is a hybrid of English and German in the same vein as *Singlish*.

Example of *Denglisch* explanations:

"Premise: Der Hund ist im Schnee vor einigen Bäumen.
Hypothesis: Eine Katze spielt im Schnee.

Please provide a justification for whether the hypothesis is entailed by the premise, contradicted by the premise, or if it cannot be determined."

– – – –

"Contradiction. Der hund isn't necessarily eine katze."

Contextual Embedding Alignment. Huang et al. (2021) and Cao et al. (2020) show that the embeddings for words from different languages in a multilingual corpus can be aligned; in fact, Cao et al. (2020) demonstrate improved zero-shot performance on XNLI (Conneau et al., 2018) using this technique. For instance, the embedding for the word **cat** in the English phrase "*the cat sits*" and its German equivalent **katze** in "*Die katze sitzt*" would have similar representations. This word embedding alignment technique could be used to help Flan-T5 generate coherent explanations. However, this technique may exacerbate the aforementioned issue on language inter-mixing in the generated output as a side-effect given that the model learns a language-agnostic representation of concepts, allowing it to *context-switch* between languages in the same explanation. We leave this investigation for future work.

e-SNLI Explanation Quality. Manual inspection of e-SNLI reveals occasional poor-quality explanations for text. Translating these faulty explanations to other languages, further exacerbated by the uncontrollable translation errors, result in poor-quality finetuning data for the models. Additionally, our translation models may not pick up on certain nuances in the explanations which result in poor downstream explanations in other languages. We leave the construction of a refined MLe-SNLI for future work.

Additional Language Support. As done in Conneau et al. (2018) and Hu et al. (2020), we hope to expand MLe-SNLI to other languages like Russion, Bahasa Indonesia, Vietnamese, Greek, Japanese, Chinese, as well as popular South Asian languages like Hindi, Tamil, Bengali, Urdu chief among others. Some of these languages have dialects, pidgins, and creoles that are low-resource and may benefit from having a scalable model that exhibits the emergence of multilingual inference, as shown in this report for the subset of languages we have chosen.

8 Conclusion

In this work, we demonstrate the emergence of cross-lingual transfer abilities in LMs through the use of instruction-tuning. Across all scales, we empirically show that models finetuned on globally similar languages outperform zero-shot models on unseen, often-low-resource languages (German, here) in the context of natural-language-based logical inference. We also release Multilingual e-SNLI, a cross-lingual transfer corpus used in this study which further improves on existing benchmarks in the logical inference space. We believe our discoveries are important in view of rare, low-resource languages out there. We hope that the community continues to find similar languages (in the same vein of high-resource Mandarin being similar to low-resource Hokkien for speech-to-text) to alleviate difficulties of training models solely on low-resource languages which is an arduous ordeal.

Acknowledgements

We would like to thank the CS4248 teaching team and our project mentor, Yisong Miao, for allowing us to embark on this exciting project. We also express our sincere gratitude to Delip Rao (UPenn SEAS), Yi Tay (Reka, Google Brain), Sayak Paul (HuggingFace), and Sam Witteveen (Red Dragon AI, GDE), for the advice, pointers, and informal conversations that led to the completion of this project.

References

Wasi Uddin Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual bert for cross-lingual transfer.

Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. 2022. Deepspeed inference: Enabling efficient inference of transformer models at unprecedented scale.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations.

VOKE EFEOTOR et al. 2022. *What do polyglots know about learning languages? Assessing the beliefs and perceptions of polyglots vis-à-vis language learning*. Ph.D. thesis, Durham University.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Charlotte Gooskens, Vincent J van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in europe. *International Journal of Multilingualism*, 15(2):169–193.

Kenji Hakuta and Rafael M Diaz. 2014. The relationship between degree of bilingualism and cognitive ability: A critical discussion and some new longitudinal data. In *Children's language*, pages 337–362. Psychology Press.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language?

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training.

Kenneth Hyltenstam. 2016. The exceptional ability of polyglots to achieve high-level proficiency in numerous languages. *Advanced proficiency and exceptional ability in second languages*, pages 241–272.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++.

Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. 2021. Analyzing the effects of reasoning types on cross-lingual transfer performance.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Chia-Hsuan Lee and Hung-Yi Lee. 2019. Cross-lingual transfer learning for question answering.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. 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 5834–5846, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning.

Amir Pouran Ben Veyseh. 2016. Cross-lingual question answering using common semantic space. In *Proceedings of TextGraphs-10: the Workshop on Graph-based Methods for Natural Language Processing*, pages 15–19, San Diego, CA, USA. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis insights from training gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference.

Elizabeth Williams and Elizabeth WilliamsI'm Elizabeth. 2022. Languages similar to english- top 10: Higherlanguage.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. 2021. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834, Online. Association for Computational Linguistics.

Statement of Independent Work

1A. Declaration of Original Work. By entering our Student IDs below, we certify that we completed our assignment independently of all others (except where sanctioned during in-class sessions), obeying the class policy outlined in the introductory lecture. In particular, we are allowed to discuss the problems and solutions in this assignment, but have waited at least 30 minutes by doing other activities unrelated to class before attempting to complete or modify our answers as per the class policy.

We have documented our use of AI tools (if applicable) in a following table, as suggested in the NUS AI Tools policy. This particular document did not use any AI Tools to proofcheck and was constructed and edited purely by manual work.

1B. Exception to the Class Policy. We did not follow the CS4248 Class Policy in doing this assignment. This text explains why and how we believe we should be assessed for this assignment given the circumstances explained.

Signed, A0220603Y, A0219702J, A0214563J, A0223837B, A0225744A

A Appendix

A.1 Alternative Standard Prompts

Here is an unordered list of other standard prompts prepended before the sample's premise and hypothesis that we experimented with during training and testing but do not report in this work:

- "*Does the premise entail or contradict the hypothesis? Give the rationale before answering.*"
- "*Determine if the premise entails or contradicts the hypothesis, or is not possible to tell. Provide the answer and the explanation.*"
- "*Is there an entailment or contradiction between the premise and the hypothesis, or is it unclear at this point? Please give a rationale for your answer.*"
- "*Can it be inferred from the available evidence whether the premise contradicts or entails the hypothesis, or is the relationship uncertain? Please provide a justification for your answer.*"
- "*Determine which of the following is true and give the rationale. A: The premise entails the hypothesis. B: The premise contradicts the hypothesis. C: It is not possible to tell.*"

A.2 Alternative CoT Prompt

We tried an alternative CoT prompt aside from the one mentioned in Section 6.1 but could not successfully run it on limited hardware (OOM) due to its large context length. Given a premise, `prem`, and hypothesis, `hypo`, from the Spanish corpus, our instruction-tuning prompt with exemplars (in red) is as follows:

"Premise: Un hombre de camisa azul parado frente a una estructura de garaje pintada con diseños geométricos.
Hypothesis: Un hombre lleva una camisa azul.
Explanation: "in a blue shirt" is inferred as "wearing a blue shirt". The answer is Entailment.

Premise: En un evento al aire libre en una zona temática asiática, una multitud se congrega como una persona en un traje amarillo de dragón chino se enfrenta a la cámara.
Hypothesis: Un solo hombre está al lado de una cámara.
Explanation: The crowd can't be single. The answer is Contradiction.

Premise: Un perro blanco con el pelo largo salta para coger un juguete rojo y verde.
Hypothesis: Una mezcla de chihuahua blanco con el pelo largo rizado está saltando para atrapar un hueso de juguete rojo y verde.
Explanation: The jumping dog isn't necessarily a chihuahua with curly hair, and it's not necessarily trying to catch a red and green toy bone. The answer is Neutral.

Premise: ⟨prem⟩.
Hypothesis: ⟨hypo⟩.
Explanation:"

We created similar unseen exemplars for the other languages in MLe-SNLI.

A.3 Comparison of models with different sizes

We tried training models using different sizes as stated in Section 5. The example below suggests that the model has better understanding of the language, and therefore supports our hypothesis that larger models exhibits emergence of cross-lingual inference abilities. We present this in the boxes below (they are not prompts).

Premise: A group of young people pose for a picture in a field.
Hypothesis: The young people are outdoors.
Label: Entailment
Explanation 1: pose for a picture in a field means outdoors
Explanation 2: a group of people in a field implies that the people are outdoors
Explanation 3: young people are young people and a field is outdoors

Small: younger people cannot be younger people and younger people at the same time
Base: a gruppe (German of group) of younger people is a group of younger people
Large: a group of younger people posing for a picture on a field means they are outside

Below is another example:

"Premise: A girl in a black sweater and jeans pours water into an empty soda bottle.
Hypothesis: A girl pours water into the ocean.

Label: Contradiction
Explanation 1: A soda bottle is not the ocean.
Explanation 2: The ocean is a different place than an empty soda bottle.
Explanation 3: The girl is either pouring water into an empty soda bottle or into the ocean.

Small: the girl gieÃŸt wasser in a leere sodaflasche is in a leere sodaflasche (German: the girl pours water into an empty soda bottle is in an empty soda bottle)
Base: wasser is geen sodaflasche, geen ozean is geen sodaflasche, geen sodaflasche i (Incomplete Dutch: washer is not a soda bottle, no ocean is not a soda bottle, not a soda bottle i)
Large: een sodaflasche is geen ozean (Dutch: a soda bottle is not an ocean)