

Aligning LLMs to Low-Resource Languages

AAAI Tutorial 2024

Today.

Why should we work on multilingual LLM models?

Why LLMs struggle with non-English, specially low-resource languages?


What is instruction finetuning?

How can we make models multilingual?

Languages are not treated equally by researchers. Some languages have received disproportionate attention and focus in NLP.

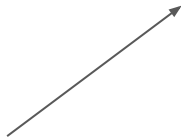
Language	# of papers per million speakers	# of speakers (in millions)
Irish	5235	0.2
Basque	2430	0.5
German	179	83
English	63	550
Chinese	11	1,000
Hausa	1.5	70
Nigerian Pidgin	0.4	30

Number of papers in top NLP venues referencing language per 1 million speakers. [[Van Etch et al. 2022](#)]



This uneven coverage also means that many languages have been left out of the technological progress.

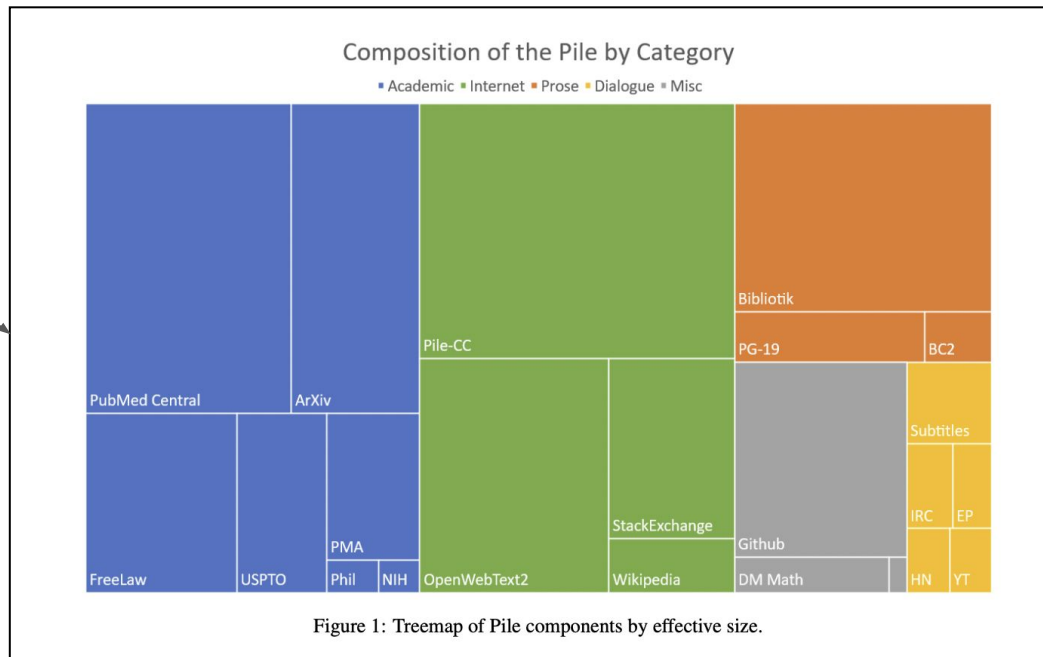
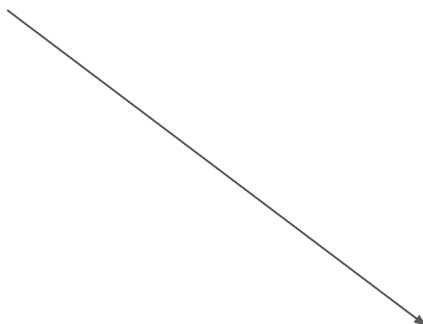
Multilingual Model Name	Number of Languages Trained On (pre-training)
BLOOM	46
mT5	101
XGLM	30



Open source multilingual state of art Large Language Models (LLM) are pre-trained a smaller subset of available languages.

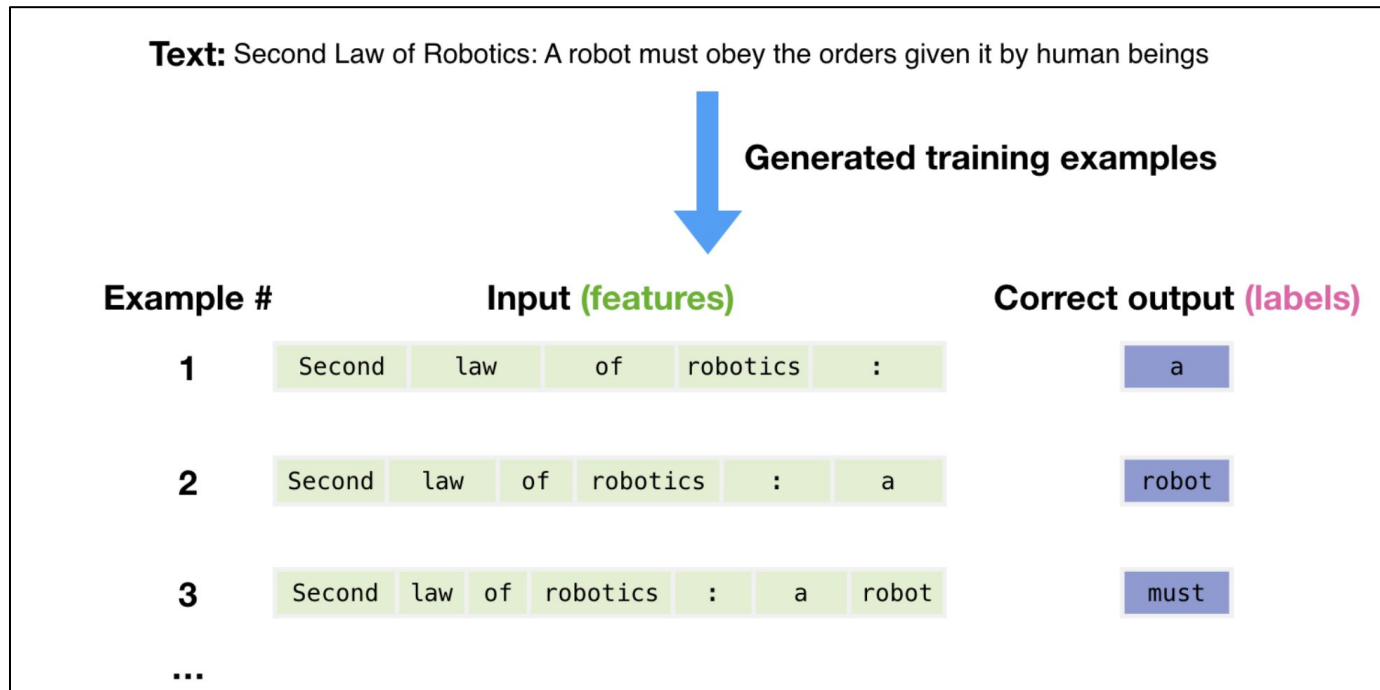
Why have some languages been left behind in technological progress?

Much of our data in large language model training comes from the internet.



Pretraining on larger and larger datasets in an unsupervised fashion.

Step 1:
Unsupervised
pre-training of
a transformer
model on a
massive web
crawled dataset
(i.e. train on
the internet).



<https://jalammar.github.io/how-gpt3-works-visualizations-animations/>

Changed to multi-task fine-tuning. Moving to a single global model – train on multiple tasks at once.

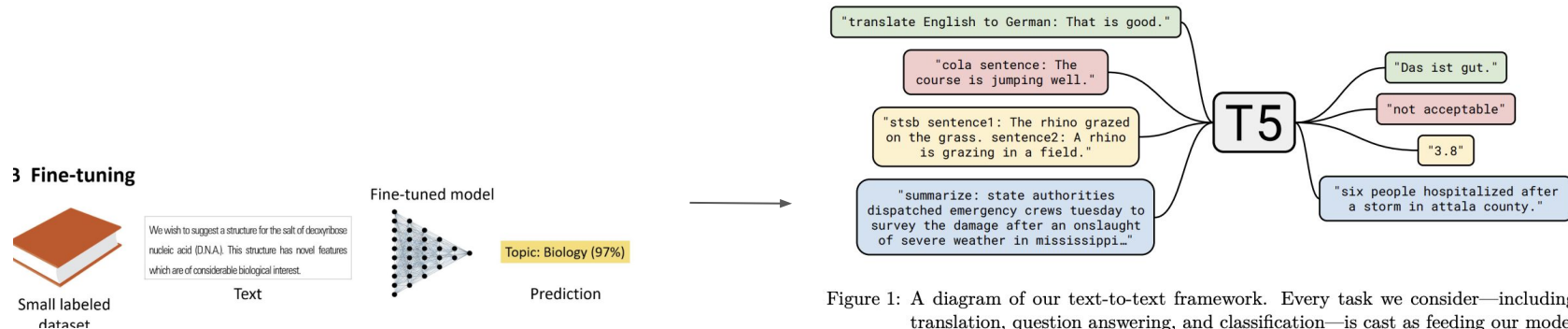


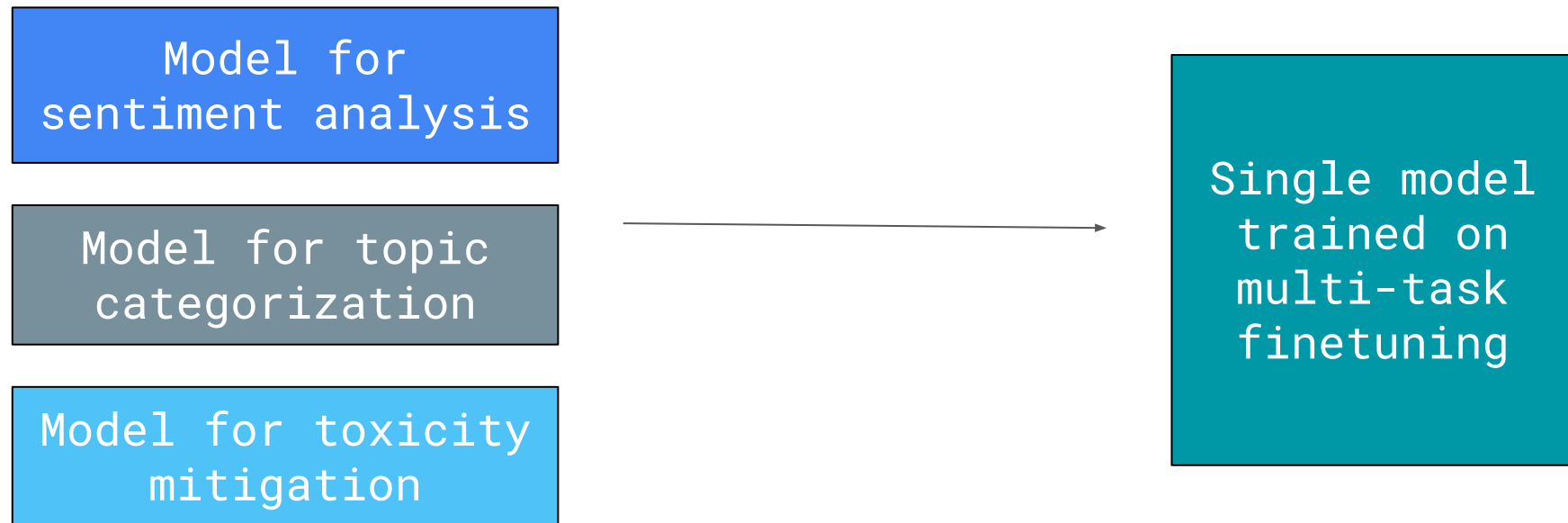
Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. “T5” refers to our model, which we dub the “Text-to-Text Transfer Transformer”.

Finetuning on a single task



Finetuning on many different tasks

Why is this a big deal – it transitions from having custom models for each task to having a single task-general model that can perform a lot of tasks, which only require zero or few examples



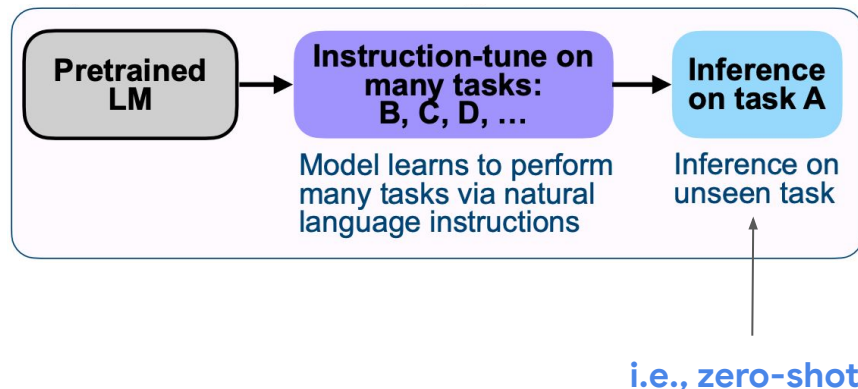
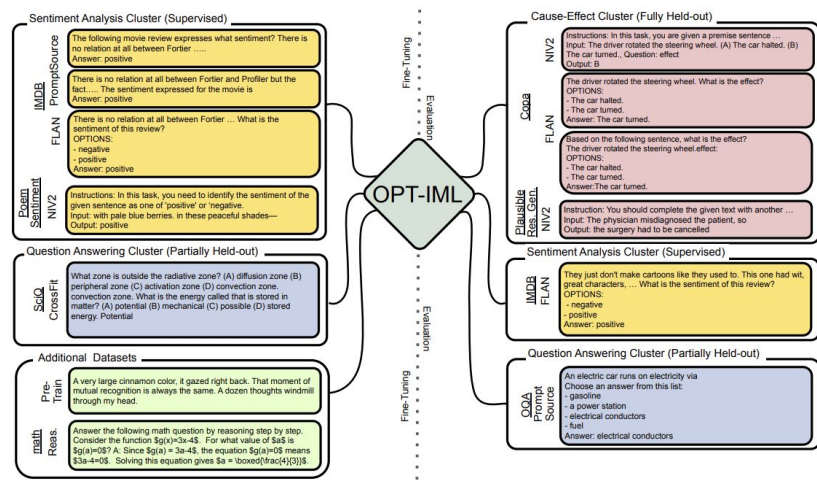
It turns out two ingredients have been particularly important at leading to breakthroughs in performance on zero and few shot tasks:

**1. Instruction tuning –
Structuring multi-task
fine-tuning data as
questions and
answers**

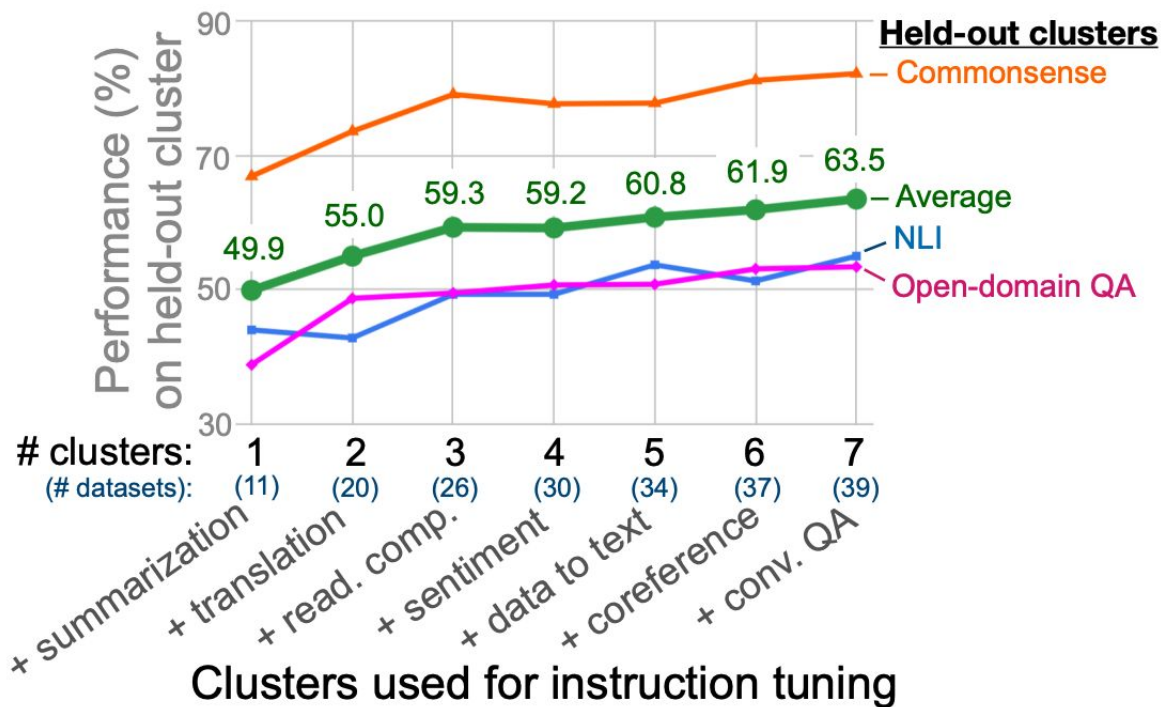
**2. Integrating human
feedback about
preferences**

Instruction Tuning – Finetuning a LLM on a collection of tasks described by instructions to improve performance on unseen tasks.

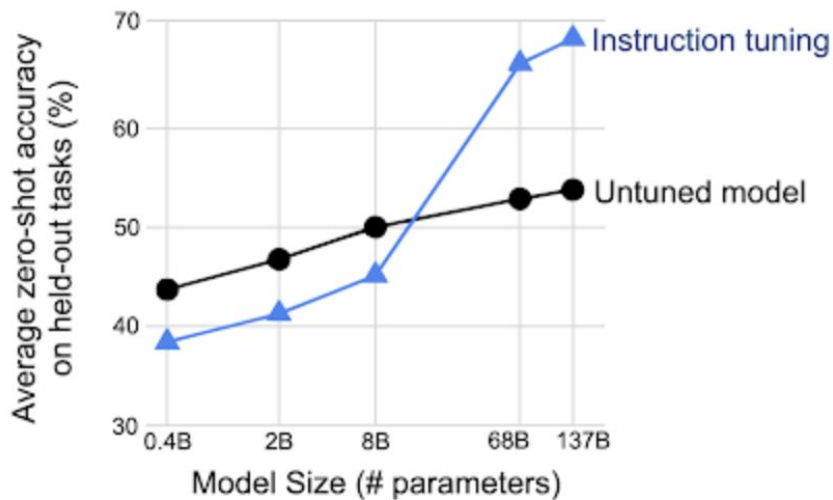
Leverage supervision to teach the model to perform many NLP tasks.



This combination – of multitask training and instruction style improves zero shot performance.



It also requires larger and larger models to take advantage of instruction tuning (partly explaining our race to ever larger models).



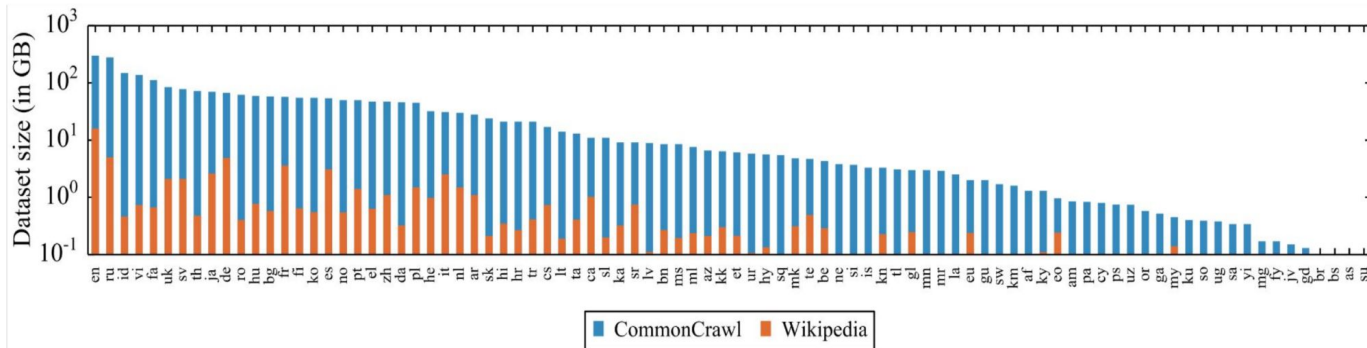
Instruction tuning only improves performance on unseen tasks for models of certain size.

Zero shot performance is particularly helpful for data limited regimes.

- Data limited regimes struggle to realize gains of fine-tuning.
- Fine-tuning large language models can be expensive (which typically impacts low resource languages more [Oreva et al. 2021](#)) – would be great if a model generalized to a task out of the box.

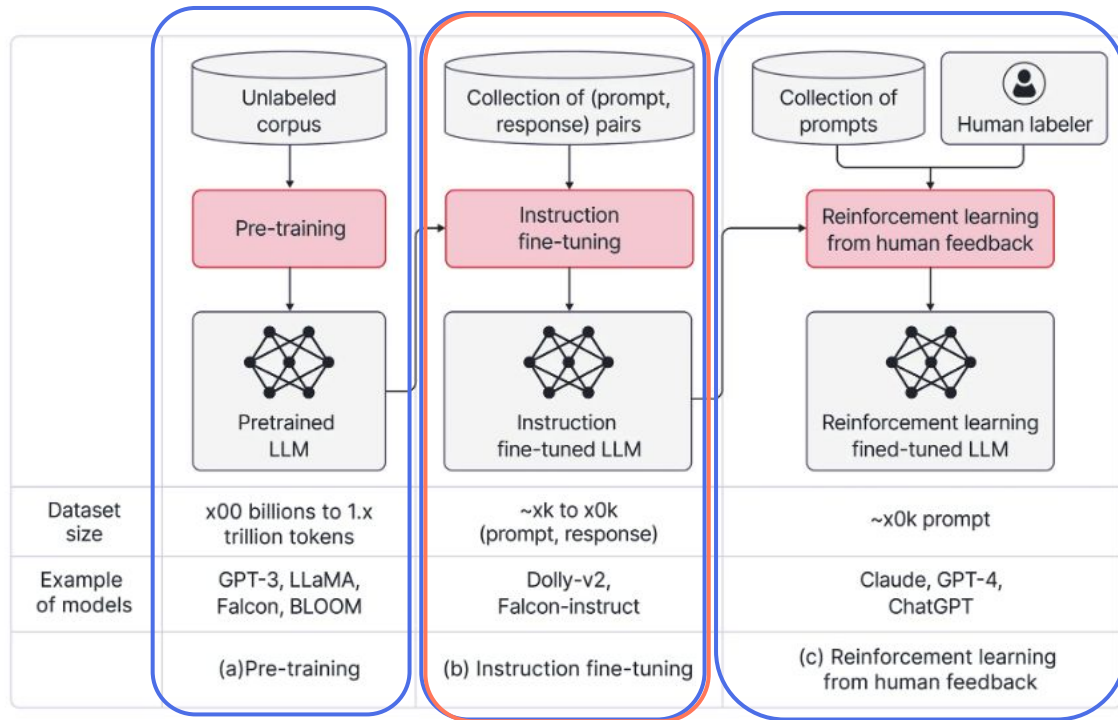
ACL [Keynote](#), [Conneau et al.](#)

This makes instruction finetuning a particularly promising research direction for multilingual, where there is a pronounced skew in the “haves” and “have nots.”

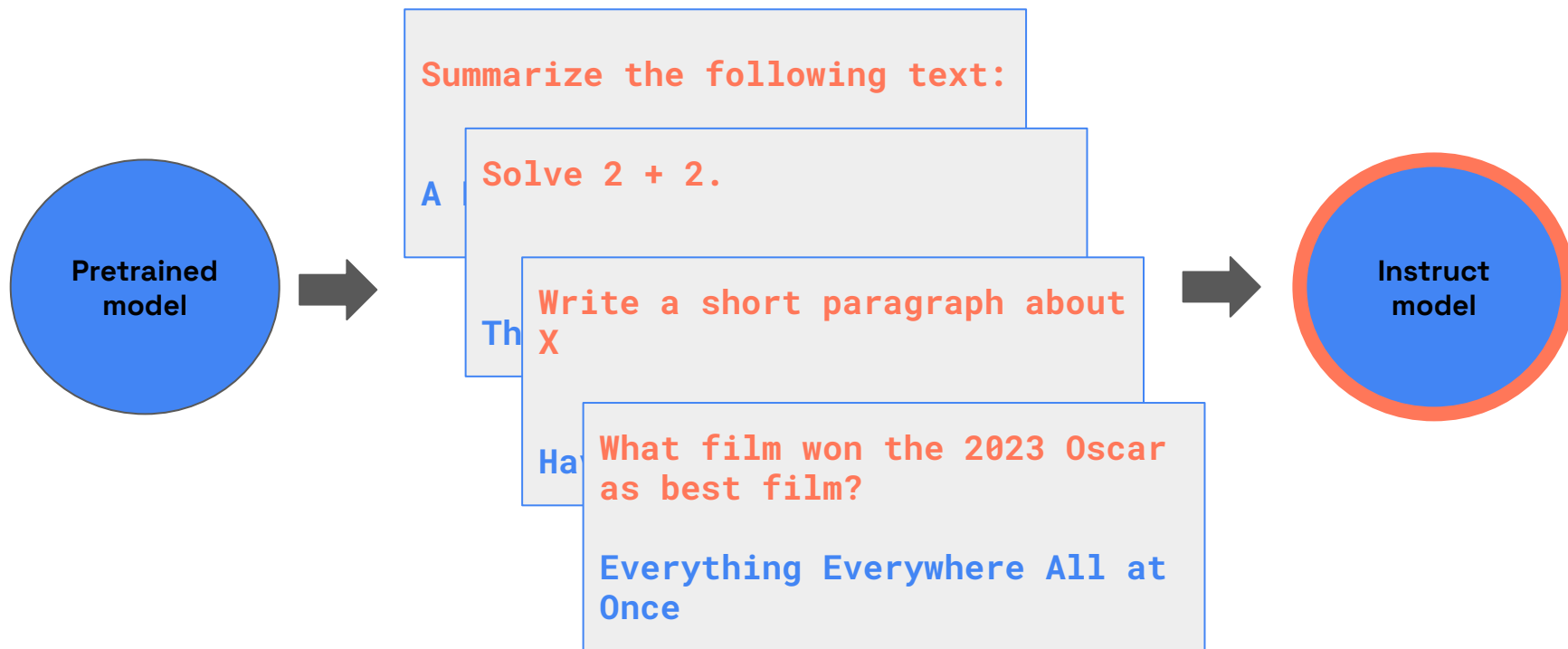


The long-tail of multilinguality, few high resource languages and many sparsely populated languages.

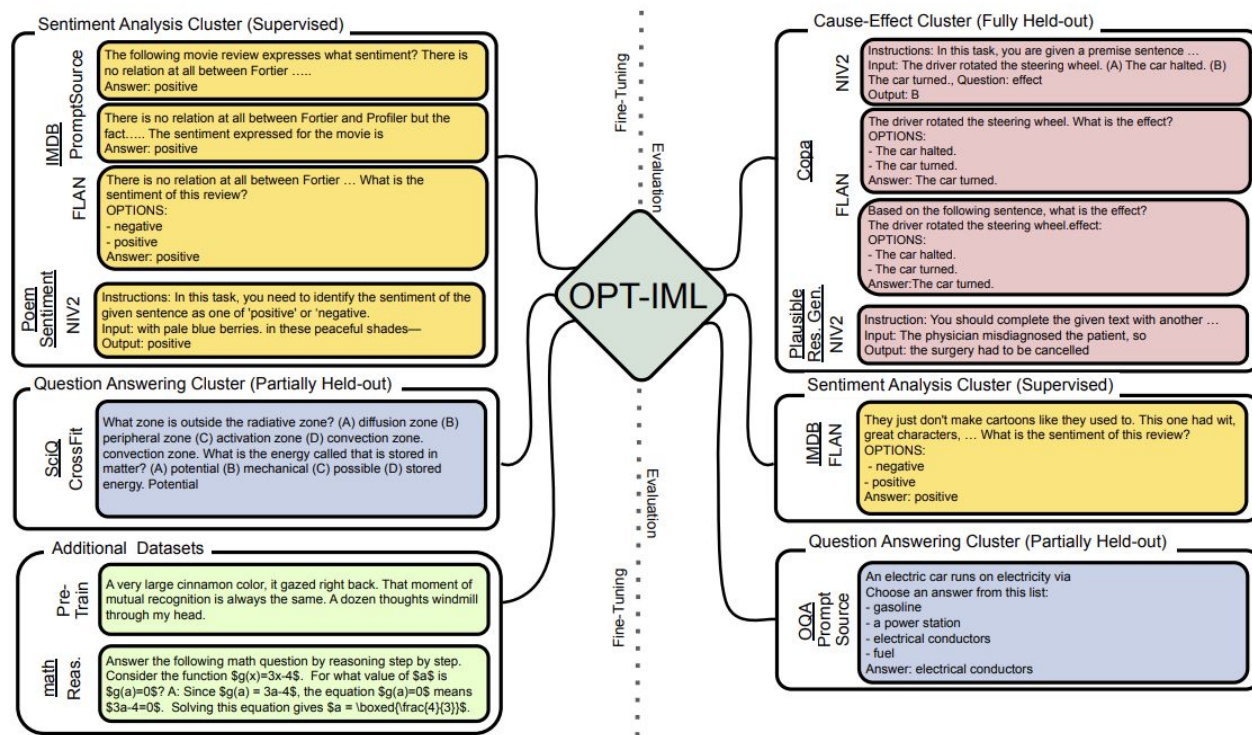
LLM recipe



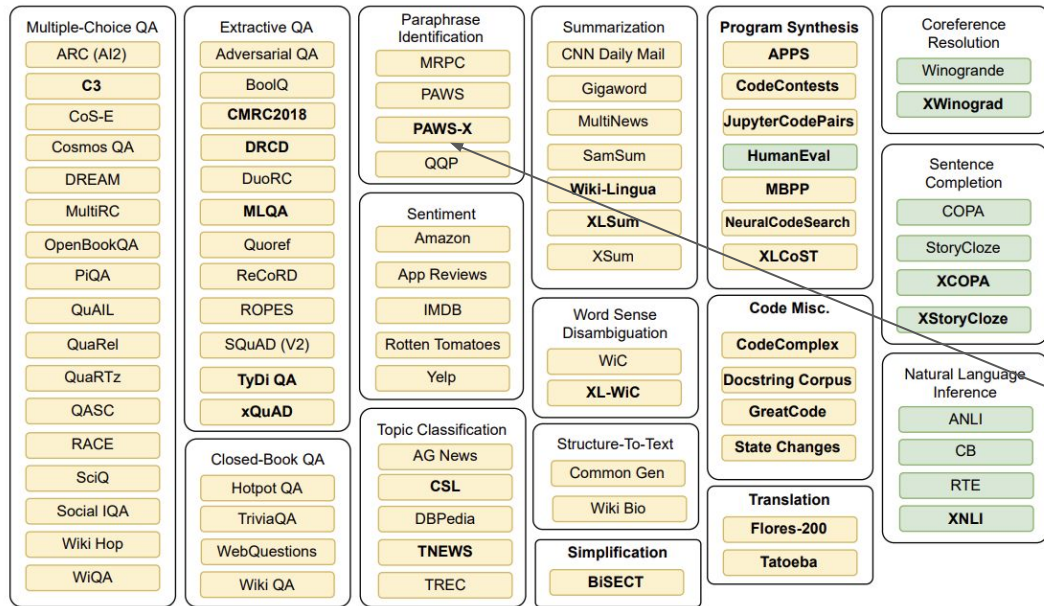
What does the data actually look like?



While considerable work has focused on English language instruct style datasets, far less has explored the benefits for a multilingual setting.



Most relevant is work released last year by [Muennighoff et al.](#)



Added multilingual and program synthesis datasets to P3.

Figure 1: An overview of datasets in xP3. Datasets added to P3 in this work are marked **bold**. Yellow datasets are trained on. Green datasets are held out for evaluation.

Observed boosts in performance over base multilingual models.

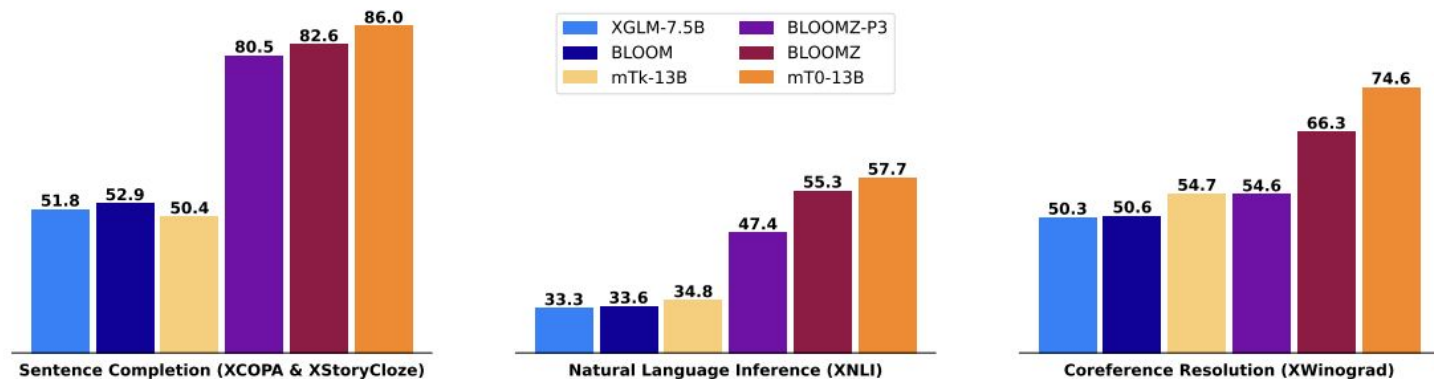
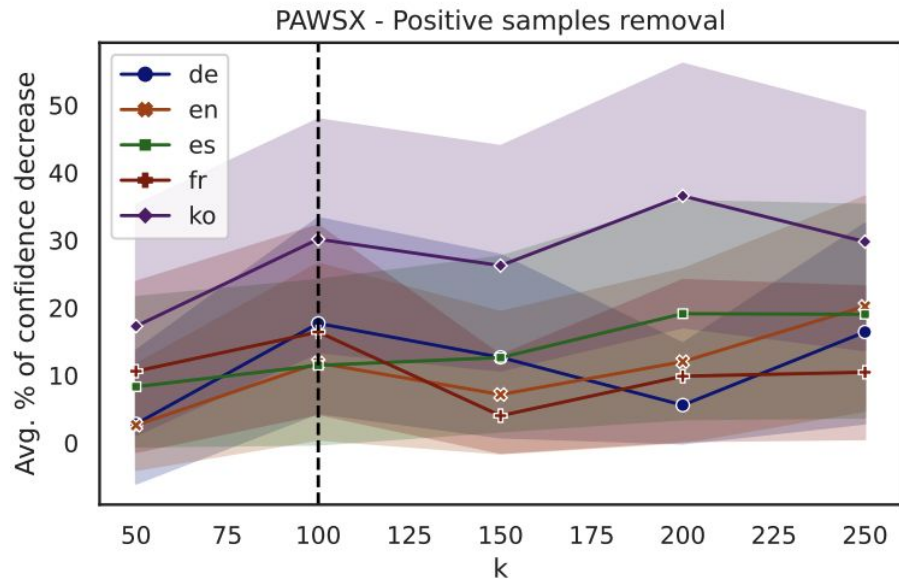


Figure 4: Zero-shot multilingual task generalization with English prompts. BLOOM models have 176 billion parameters. Scores are the language average for each task. Appendix §B breaks down performance by language.

Choenni et al. also observe that multi-task finetuning benefits multilingual tasks in-distribution performance.

Cross-lingual sharing increases as finetuning progresses.

Languages can support one another by playing both reinforcing as well as complementary roles.



This is despite the datasets added only covering 46 languages, and having no human feedback optimization.

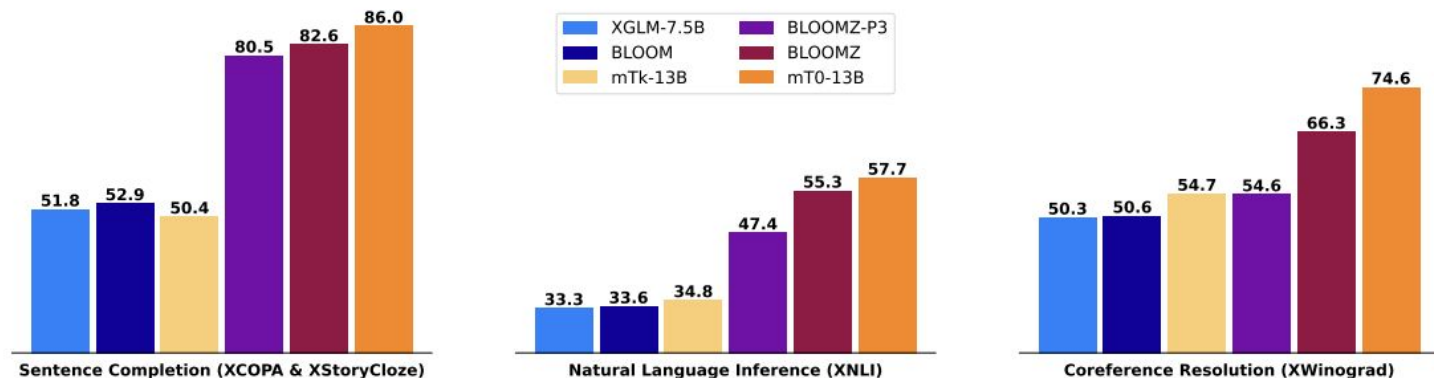


Figure 4: Zero-shot multilingual task generalization with English prompts. BLOOM models have 176 billion parameters. Scores are the language average for each task. Appendix §B breaks down performance by language.

