Treating Models Better for Language Agnostic Understanding

Brian Yu, Hansen Lillemark, Kurt Keutzer

BAIR

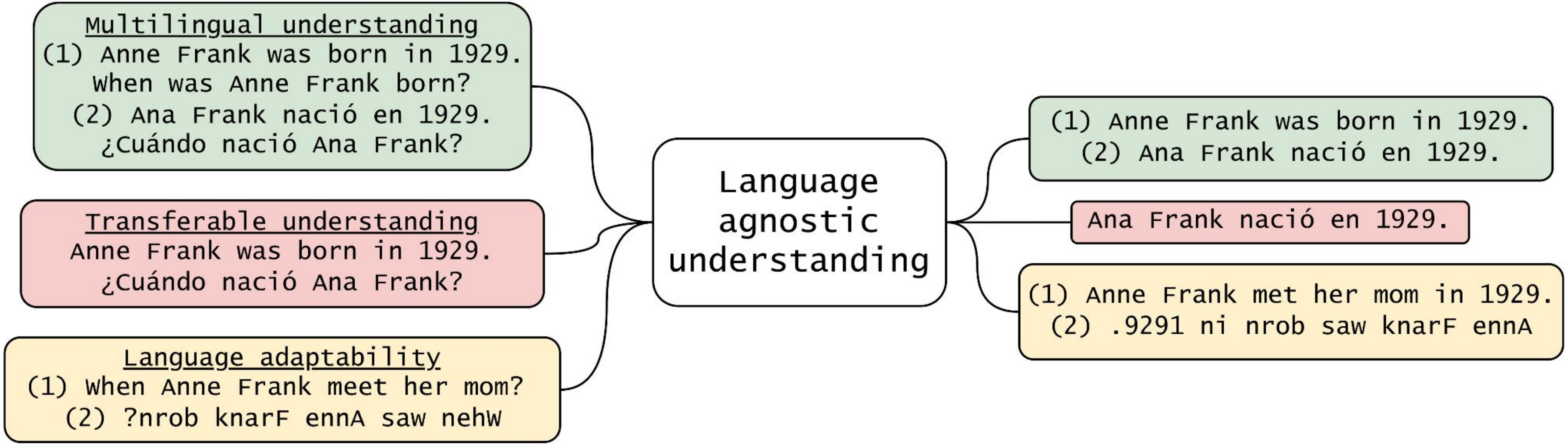
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Foundation models are the future of NLP

- Foundation models are **SOTA** on all downstream language and vision tasks
- Models today **do not focus on multilingual performance**, they just happen to pretrain on a multilingual corpus
- Models are **incredibly strong at pattern matching** within an input context and **completing an input**, enabling them to be prompted for downstream tasks
- These strengths (and other strengths) have **not been applied to finetuning**

Language agnostic understanding

Multilingual NLP lacks a clear goal



Multilingual understanding
(1) Anne Frank was born in 1929.
When was Anne Frank born?
(2) Ana Frank nació en 1929.
¿Cuándo nació Ana Frank?

Transferable understanding
Anne Frank was born in 1929.
¿Cuándo nació Ana Frank?

Language adaptability
(1) When Anne Frank meet her mom?
(2) ?nrob knarF ennA saw nehW

Language agnostic understanding

(1) Anne Frank was born in 1929.
(2) Ana Frank nació en 1929.

Ana Frank nació en 1929.

(1) Anne Frank met her mom in 1929.
(2) .9291 ni nrob saw knarF ennA

Multilingual models and where they fall short

- Most popular approach for multilingual understanding is mT5-like: apply monolingual pretraining on a multilingual dataset.
  - However, mT5 has **poor transferable understanding**
- Best translation-only model: NLLB, trained explicitly on translation
  - NLLB is unfit for multilingual understanding because it has not been trained to respond to inputs. For example, asking it a question in English and asking for a response in English yields the original input.
- Arguably models today are **incapable of language adaptability** since they require enormous amounts of data to train

mT5: A massively multilingual pre-trained text-to-text transformer. Xue et al; Google (Oct 2020). https://arxiv.org/abs/2010.11934
No Language Left Behind: Scaling Human-Centered Machine Translation. Fan et al; FAIR (Sep 2022). https://arxiv.org/abs/2207.04672

mT5 has poor transferable understanding

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Experimental setup: mT5 monolingual pretrained on a multilingual corpus, finetuned in English, and tested in different languages.

Tell the model a novel fact in English and ask about that fact in Arabic

Hypothesis: If mT5 has perfect transferable understanding, model performance on the same task in different languages should match English performance.

Results: Non-English performance lags significantly behind.

Model can't answer the question correctly in Arabic.

Observation: mT5's performance in different languages correlates with the amount of pretraining data seen in that language

Conclusion: mT5's performance can be explained by pattern-matching on the finetuning task and leveraging strong monolingual capabilities and **not transferable understanding**

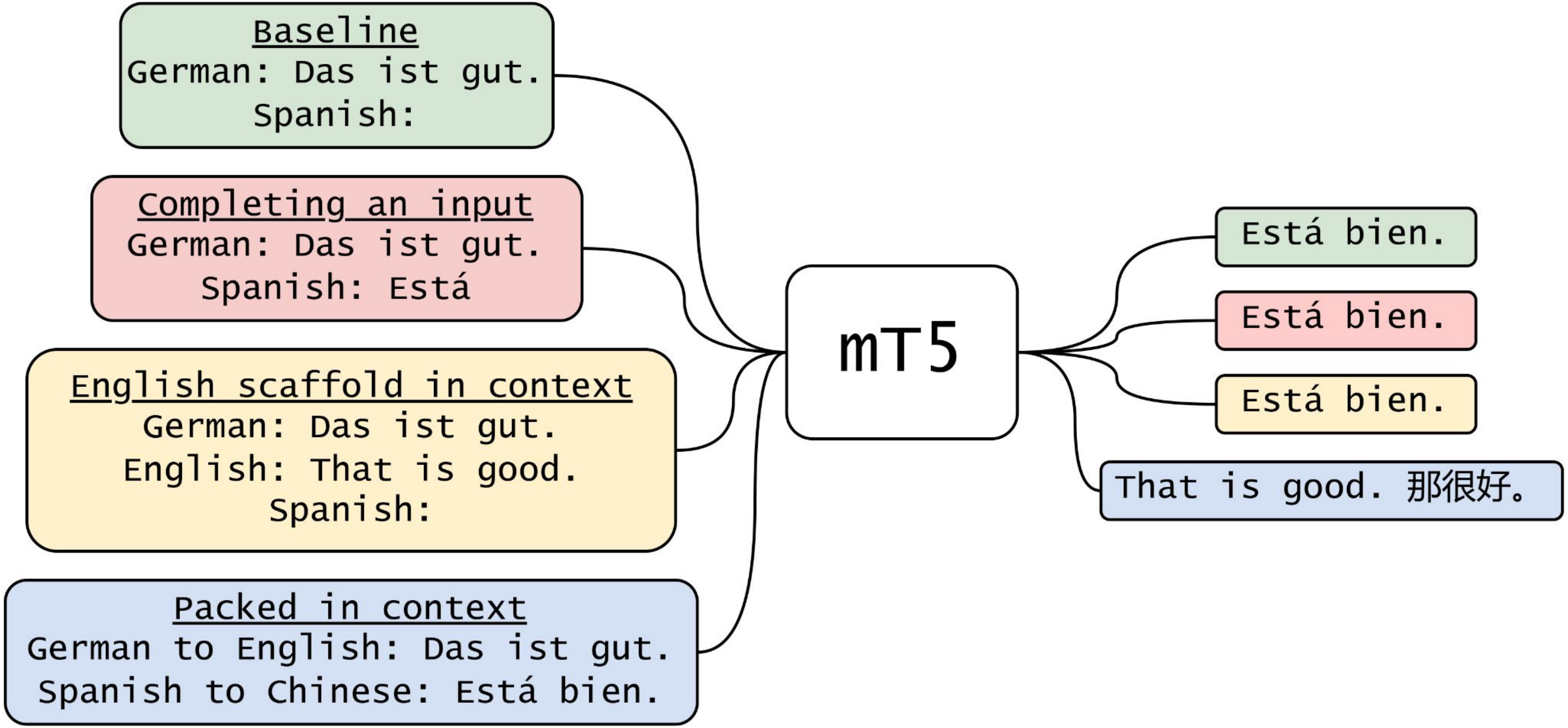
Proposition: Leverage foundation models **strengths** to improve their **language agnostic understanding**

Punchline: By including an **input context** during finetuning, we directly improve **transferable understanding**

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Fine tuning task reformulations



Baseline
German: Das ist gut.
Spanish:

Completing an input
German: Das ist gut.
Spanish: Está

English scaffold in context
German: Das ist gut.
English: That is good.
Spanish:

Packed in context
German to English: Das ist gut.
Spanish to Chinese: Está bien.

mT5

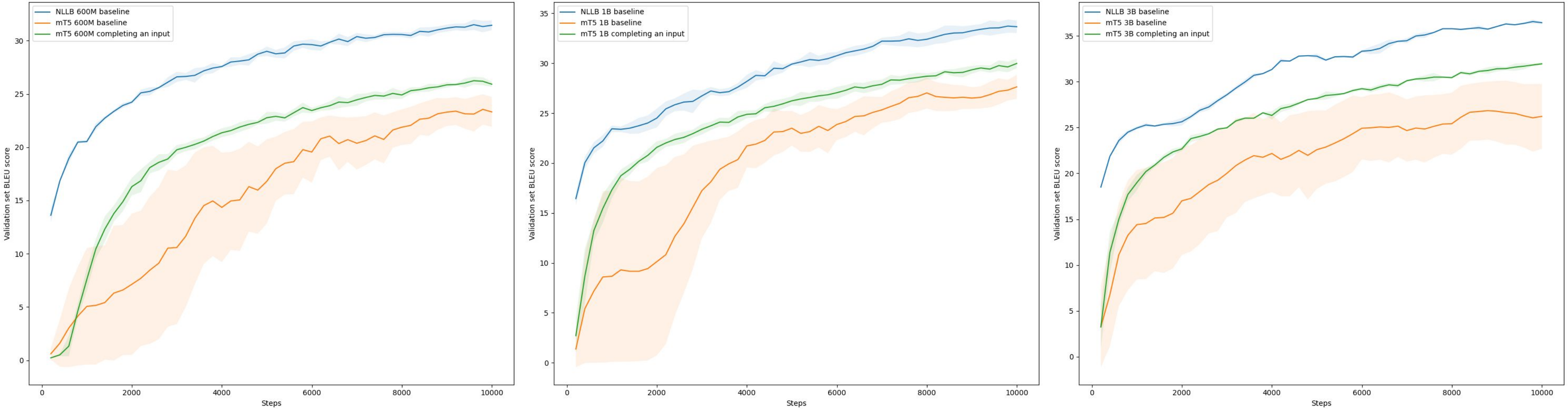
Está bien.

Está bien.

Está bien.

That is good. 那很好。

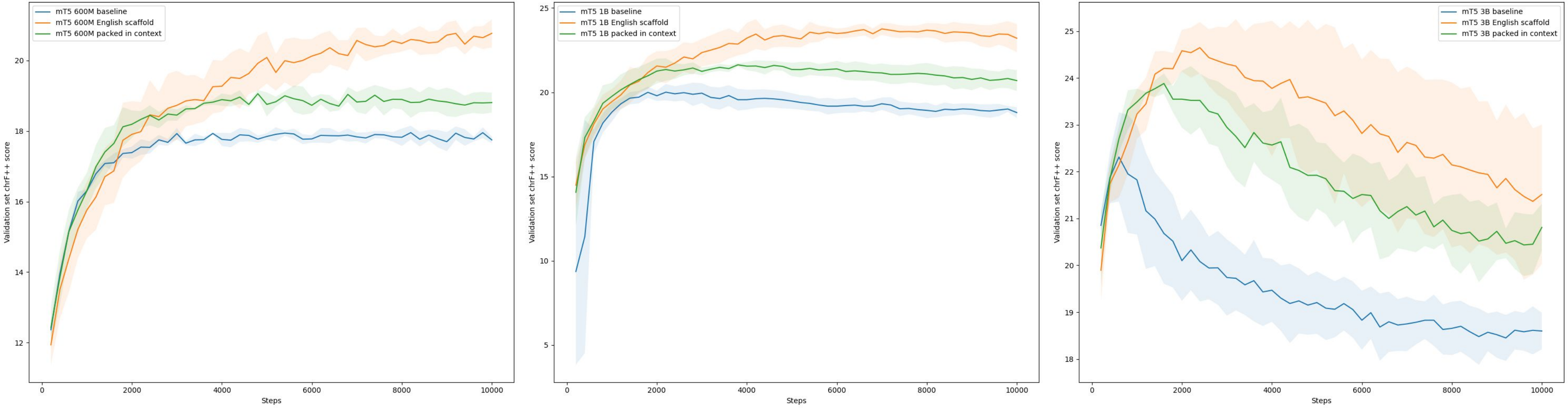
Completing an input reformulation



Classical Tibetan to English translation performance.
Left: 600M params. Middle: 1B params. Right: 3B params.
Blue: NLLB gold standard. **Green**: mT5 with reformulation. **Orange**: mT5 baseline.

mT5 performance improved up to 10.3% / 2.8 BLEU

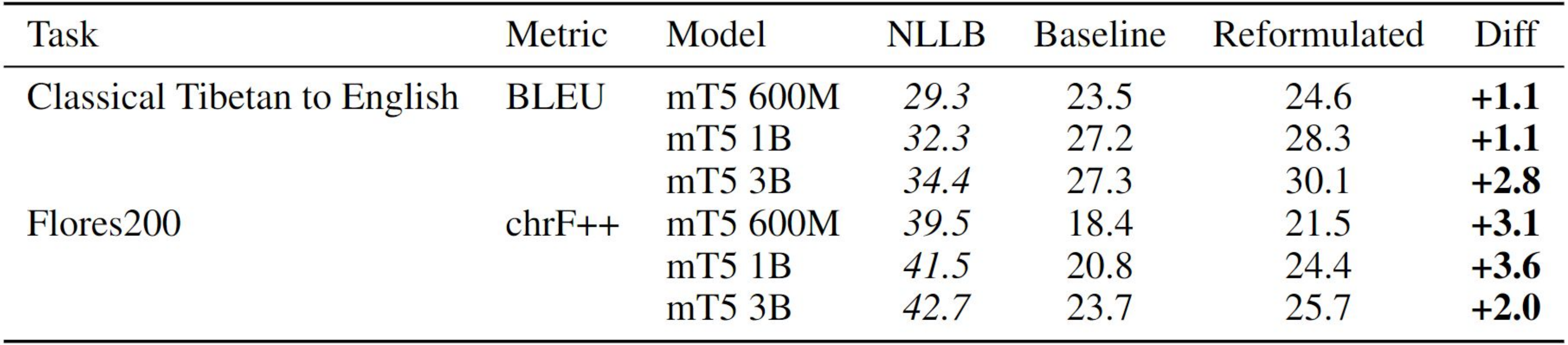
Scaffold and packed reformulations



mT5 Flores200 benchmark translation performance.
Left: 600M params. Middle: 1B params. Right: 3B params.
Orange: English scaffolding. Green: Packed in context. Blue: Baseline.

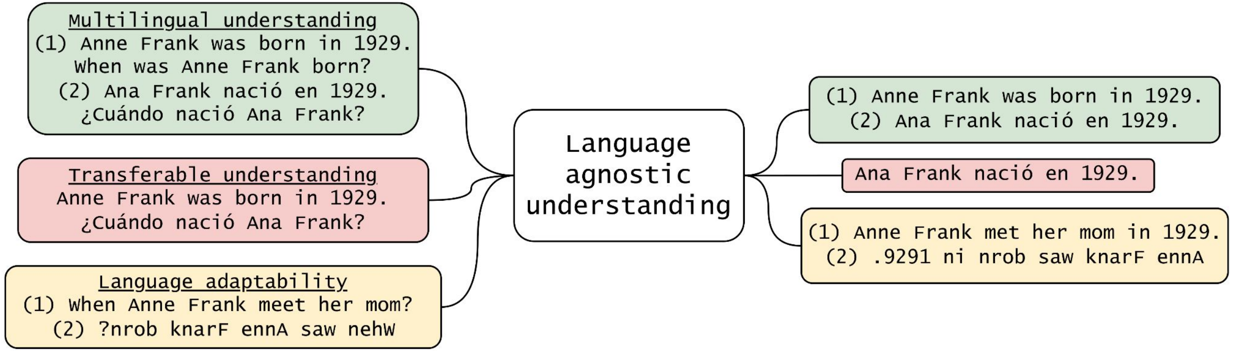
mT5 performance improved up to 17.3% / 3.6 chrF++

Results



Task	Metric	Model	NLLB	Baseline	Reformulated	Diff
Classical Tibetan to English	BLEU	mT5 600M	*29.3*	23.5	24.6	**+1.1**
		mT5 1B	*32.3*	27.2	28.3	**+1.1**
		mT5 3B	*34.4*	27.3	30.1	**+2.8**
Flores200	chrF++	mT5 600M	*39.5*	18.4	21.5	**+3.1**
		mT5 1B	*41.5*	20.8	24.4	**+3.6**
		mT5 3B	*42.7*	23.7	25.7	**+2.0**

Summary



Multilingual understanding
(1) Anne Frank was born in 1929.
When was Anne Frank born?
(2) Ana Frank nació en 1929.
¿Cuándo nació Ana Frank?

Transferable understanding
Anne Frank was born in 1929.
¿Cuándo nació Ana Frank?

Language adaptability
(1) When Anne Frank meet her mom?
(2) ?nrob knarF ennA saw nehW

Language agnostic understanding

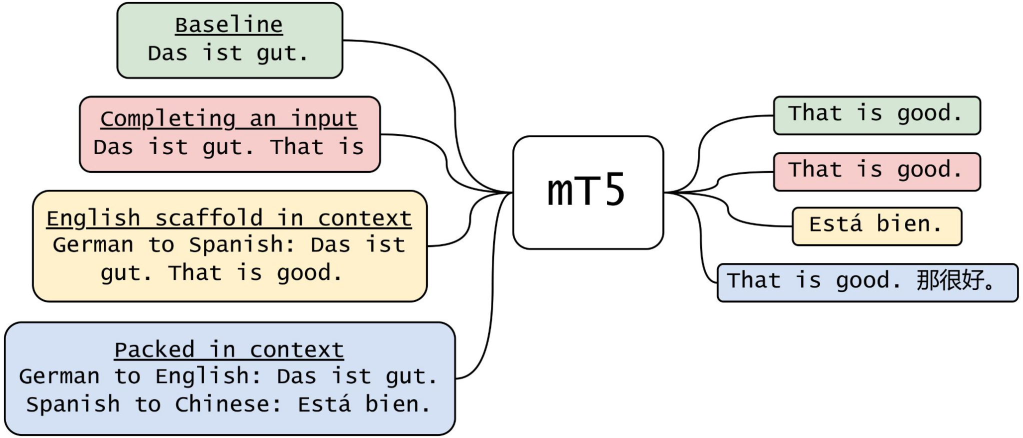
(1) Anne Frank was born in 1929.
(2) Ana Frank nació en 1929.

Ana Frank nació en 1929.

(1) Anne Frank met her mom in 1929.
(2) .9291 ni nrob saw knarF ennA

Our proposal for the goal of multilingual NLP

Reformulate inputs that leverage model strengths. The particular strength shown here is pattern matching in-context.



Baseline
Das ist gut.

Completing an input
Das ist gut. That is

English scaffold in context
German to Spanish: Das ist gut. That is good.

Packed in context
German to English: Das ist gut.
Spanish to Chinese: Está bien.

mT5

That is good.

That is good.

Está bien.

That is good. 那很好。

Conclusion and future work

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

- Analysis on mT5 Flores200 performance e.g. broken down by in-pretrain vs out-pretrain
- Transferable understanding is poor in current models but **may be less challenging than initially thought**. Translation alignment in our reformulated inputs is very cheap (~20M examples)
  - Bring back Translation Language Modeling (TLM) during pretrain with packed examples.
- **Simple and effective data efficiency technique** for finetuning on seq2seq tasks — only change data preprocessing
- We hope our research inspires further work that leverages foundation model strengths and further work on language agnostic understanding!

Thanks!

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Extra slides

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

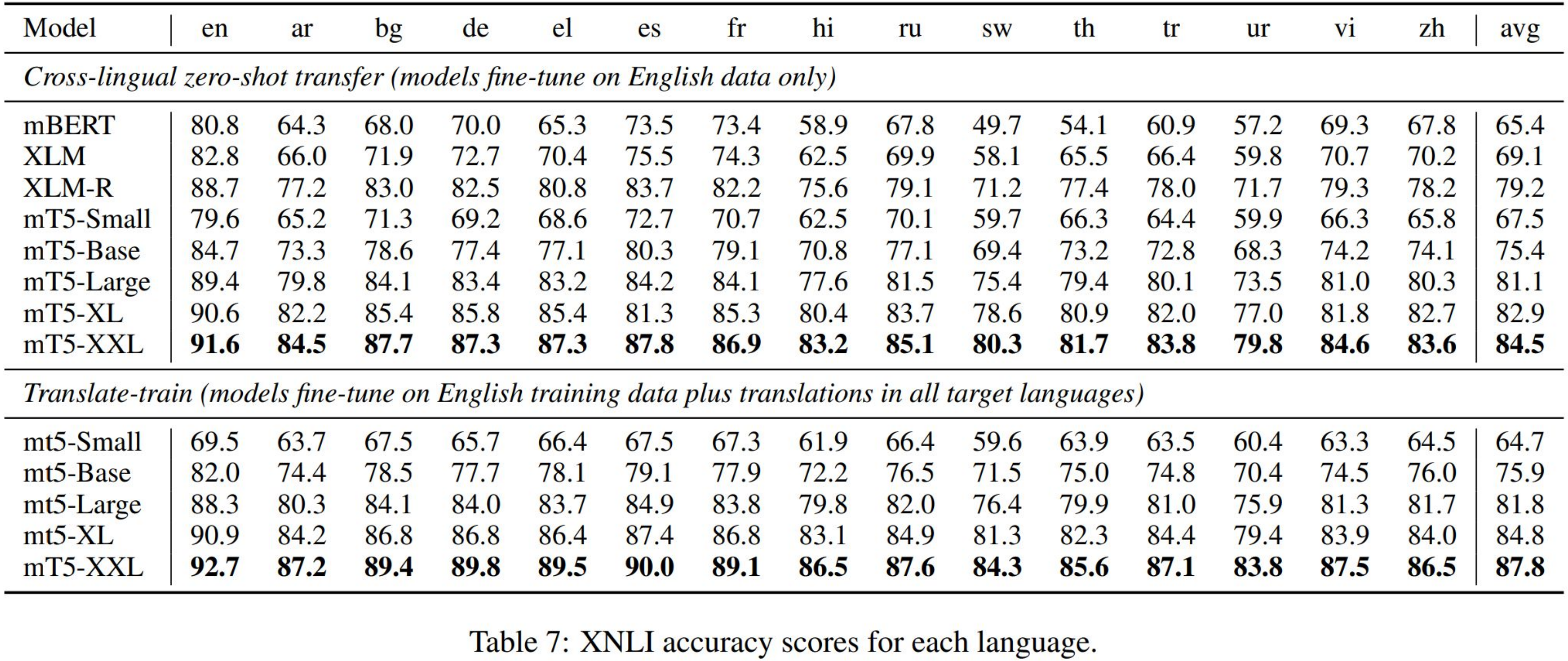
Foundation model history

Date	Model	Authors	Classification tasks	Context sensitive	Pre-trained	Transformer	All tasks	1B params	10B params	Optimized	100B params	Data/compute optimal	Open source	Efficient and accessible
Oct 2014	GloVe	Pennington et al	GloVe											
Aug 2017	CoVe	McCann et al	🟩	CoVe										
Feb 2018	ELMo	Peters et al	🟩	🟩	ELMo									
Oct 2018	BERT	Devlin et al	🟩	🟩	🟩	BERT								
Dec 2018	GPT	Radford et al	🟩	🟩	🟩	🟩	GPT							
Feb 2019	GPT-2	Radford et al	🟩	🟩	🟩	🟩	🟩	GPT-2						
Sep 2019	Megatron-LM	Shoeybi et al	🟩	🟩	🟩	🟩	🟩	🟩	Megatron-LM					
Oct 2019	T5	Raffel et al	🟩	🟩	🟩	🟩	🟩	🟩	🟩	T5				
May 2020	GPT-3	Brown et al	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	GPT-3			
Mar 2022	Chinchilla	Hoffman et al	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	Chinchilla		
May 2022	OPT	Zhang et al	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	OPT	
Feb 2023	LLaMA	Touvron et al	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	LLaMA

Multilingual approaches

Date	Model	Author	Type	Contribution	Comparison
Jan 2019	N/A	Lample and Conneau	Different pre-training	Translation language modeling for pretraining	N/A
Nov 2019	XLM-R	Conneau et al	Vanilla pre-training	Roberta architecture instead of BERT	Better than mBERT
Jan 2020	mBART	Liu et al	Translation	Sentence shuffling pre-training objective	Better than XLM-R
Oct 2020	M2M100	Fan et al	Translation	New translation dataset	Better than mBART
Oct 2020	mT5	Xue et al	Vanilla pre-training	T5 architecture	Better than XLM-R
Dec 2020	Ernie-M	Ouyang et al	Different pre-training	Cross attention and back translation MLM	Better than XLM-R, worse than mT5
May 2021	XLM-R XL	Goyal et al	Vanilla pre-training	Larger XLM-R models	Worse than mT5
May 2021	ByT5	Xue et al	Tokenization	Use bytes directly rather than tokenizing	Worse than mT5
Oct 2021	mLUKE	Ri et al	Different pre-training	mLUKE multilingual entity-based alignment	Better than XLM-R, worse than mT5
Jan 2023	XLM-V	Liang et al	Tokenization	New tokenization procedure	Worse than XLM-R

Poor transferable understanding in XNLI

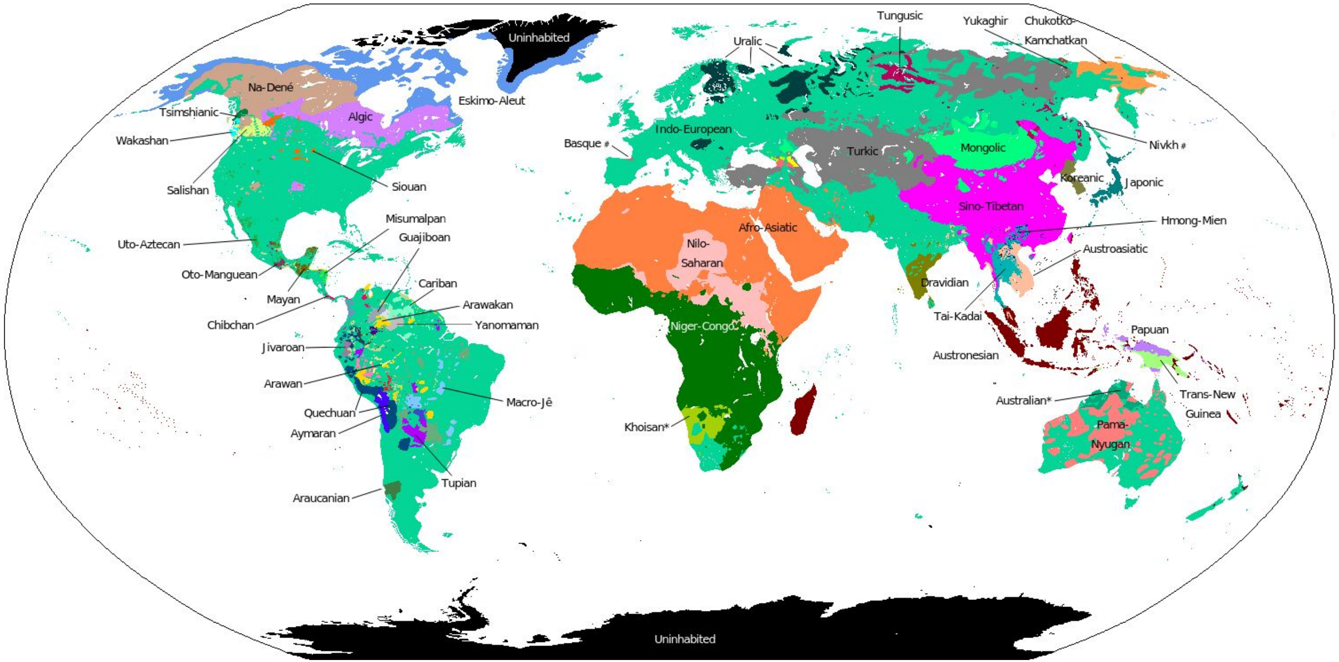


Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
Cross-lingual zero-shot transfer (models fine-tune on English data only)																
mBERT	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
XLM	82.8	66.0	71.9	72.7	70.4	75.5	74.3	62.5	69.9	58.1	65.5	66.4	59.8	70.7	70.2	69.1
XLM-R	88.7	77.2	83.0	82.5	80.8	83.7	82.2	75.6	79.1	71.2	77.4	78.0	71.7	79.3	78.2	79.2
mT5-Small	79.6	65.2	71.3	69.2	68.6	72.7	70.7	62.5	70.1	59.7	66.3	64.4	59.9	66.3	65.8	67.5
mT5-Base	84.7	73.3	78.6	77.4	77.1	80.3	79.1	70.8	77.1	69.4	73.2	72.8	68.3	74.2	74.1	75.4
mT5-Large	89.4	79.8	84.1	83.4	83.2	84.2	84.1	77.6	81.5	75.4	79.4	80.1	73.5	81.0	80.3	81.1
mT5-XL	90.6	82.2	85.4	85.8	85.4	81.3	85.3	80.4	83.7	78.6	80.9	82.0	77.0	81.8	82.7	82.9
mT5-XXL	**91.6**	**84.5**	**87.7**	**87.3**	**87.3**	**87.8**	**86.9**	**83.2**	**85.1**	**80.3**	**81.7**	**83.8**	**79.8**	**84.6**	**83.6**	**84.5**
Translate-train (models fine-tune on English training data plus translations in all target languages)																
mt5-Small	69.5	63.7	67.5	65.7	66.4	67.5	67.3	61.9	66.4	59.6	63.9	63.5	60.4	63.3	64.5	64.7
mt5-Base	82.0	74.4	78.5	77.7	78.1	79.1	77.9	72.2	76.5	71.5	75.0	74.8	70.4	74.5	76.0	75.9
mt5-Large	88.3	80.3	84.1	84.0	83.7	84.9	83.8	79.8	82.0	76.4	79.9	81.0	75.9	81.3	81.7	81.8
mt5-XL	90.9	84.2	86.8	86.8	86.4	87.4	86.8	83.1	84.9	81.3	82.3	84.4	79.4	83.9	84.0	84.8
mT5-XXL	**92.7**	**87.2**	**89.4**	**89.8**	**89.5**	**90.0**	**89.1**	**86.5**	**87.6**	**84.3**	**85.6**	**87.1**	**83.8**	**87.5**	**86.5**	**87.8**

Table 7: XNLI accuracy scores for each language.

XNLI: Evaluating Cross-lingual Sentence Representations. Conneau et al; FAIR (Sep 2018). https://arxiv.org/abs/1809.05053
mT5: A massively multilingual pre-trained text-to-text transformer. Xue et al; Google (Oct 2020). https://arxiv.org/abs/2010.11934

Multilingual benchmarks typically lack diversity



XNLI has en, ar, bg, de, el, es, fr, hi, ru, sw, th, tr, ur, vi, zh

World map from https://en.wikipedia.org/wiki/Language_family#/media/File:Primary_Human_Languages_Improved_Version.png
XNLI: Evaluating Cross-lingual Sentence Representations. Conneau et al; FAIR (Sep 2018). https://arxiv.org/abs/1809.05053

Datasets and experimental setup

Experiment parameter	**Tibetan to English**	**Flores200**
Num train steps	10000	10000
Max seq len	256	256
Batch size	512	2048
Total datapoints / tokens seen	5.1M / 350M	20.5M / 1B
Learning rates (mT5)	1e-3, 2e-3, 3e-3	1e-4, 2e-4, 3e-4
Dataset statistic	**Tibetan to English**	**Flores200**
Total datapoints / epochs	450k / 11	40M / 0.5
Length of tokens per input (mean): NLLB / mT5	26 / 72	41 / 52

Tib to Eng Dataset Example

- mT5 is very good at completing the sentence
- Reframe translation mapping task to completing the sentence

Baseline input

Tibetan input:

ཟངས་མདོག་དཔལ་གྱི་རི་བོ་ར་སྐྱེ་བར་ཤོག །

English output:

May we be born on the
Copper-Coloured Mountain of Glory.

Completing the input

Tibetan input:

ཟངས་མདོག་དཔལ་གྱི་རི་བོ་ར་སྐྱེ་བར་ཤོག །May we be born on

English output:

May we be born on the
Copper-Coloured Mountain of Glory.

Flores200 Dataset Example

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

- Flores200 is a parallel dataset of 204 languages of 3 sets of 1000 sentences
- Each sentence has been professionally translated to all 204 languages

Baseline input (Spanish to German)

Input: Spanish: El lunes, los científicos de la facultad de medicina de la Universidad de Stanford anunciaron el invento de una nueva herramienta de diagnóstico que puede catalogar las células según su tipo: un pequeñísimo chip que se puede imprimir y fabricar con impresoras de inyección de uso corriente, por un posible costo de, aproximadamente, un centavo de dólar por cada uno.

German:

Output: Am Montag haben die Wisenschaftler der Stanford University School of Medicine die Erfindung eines neuen Diagnosetools bekanntgegeben, mit dem Zellen nach ihrem Typ sortiert werden können: ein winziger, ausdruckbarer Chip, der für jeweils etwa einen US-Cent mit Standard-Tintenstrahldruckern hergestellt werden kann.

Scaffold input (Spanish + English to German)

Input: Spanish: El lunes, los científicos de la facultad de medicina de la Universidad de Stanford anunciaron el invento de una nueva herramienta de diagnóstico que puede catalogar las células según su tipo: un pequeñísimo chip que se puede imprimir y fabricar con impresoras de inyección de uso corriente, por un posible costo de, aproximadamente, un centavo de dólar por cada uno.

English: On Monday, scientists from the Stanford University School of Medicine announced the invention of a new diagnostic tool that can sort cells by type: a tiny printable chip that can be manufactured using standard inkjet printers for possibly about one U.S. cent each.

German:

Output: Am Montag haben die Wisenschaftler der Stanford University School of Medicine die Erfindung eines neuen Diagnosetools bekanntgegeben, mit dem Zellen nach ihrem Typ sortiert werden können: ein winziger, ausdruckbarer Chip, der für jeweils etwa einen US-Cent mit Standard-Tintenstrahldruckern hergestellt werden kann.

No Language Left Behind: Scaling Human-Centered Machine Translation. Fan et al; FAIR (Sep 2022). https://arxiv.org/abs/2207.04672

Data efficient methods and where they fall short

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

- Our work can be viewed as a data efficient method for translation
- Past works have explored data augmentation, sample re-weighting, or curriculum learning
- These approaches vary in effectiveness, are not generalizable, and introduce unnecessary complexity

Date	Authors	Type	Technique/contribution
Aug 2015	Sennrich et al	Data augmentation	Back translation
May 2017	Fadaee et al	Data augmentation	Replace rare words with their synonyms
Jul 2017	Kocmi and Bojar	Curriculum learning	Minibatches of similar sentences and sentence types difficulty scaling
Mar 2018	Ren et al	Sample re-weighting	Meta-learning for sample weighting
Aug 2018	Gu et al	Sample re-weighting	Meta-learning specifically for low-resource translation
Nov 2018	Zhang et al	Curriculum learning	Dataset with difficulty scores, scaling over training
Feb 2019	Shu et al	Sample re-weighting	Loop between learning a sample weighting and training the model
Mar 2019	Platanios et al	Curriculum learning	Data point difficulty matched to model competence
May 2019	Zhang et al	Curriculum learning	Difficulty of datapoints by similarity to a domain
Jul 2022	Fan et al	Curriculum learning	High resource to low resource scaling

Tibetan to English curriculum learning ablations

Incomplete curriculum configuration	**Test BLEU**
First 2000 steps: 50% no addition, 50% uniformly distributed addition	23.9
Uniformly distributed addition	21.1
First 2000 steps 80% addition, 2000-6000 steps linearly scaling chance for addition, 6000-10000 no addition	24.7
First 2000 steps uniformly distributed addition (best)	**24.6**
Linear addition chance from 100% to 0%, First 2000 steps linear addition from 100% to 0%, first 2000 steps uniformly distributed prefix and suffix addition, first 2000 steps uniformly distributed suffix addition, …	Worse than best