Crosslingual Generalization through Multitask Finetuning

BLOOMZ & mT0

Paper: N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (Nov. 2022)

BigScience Workshop

Motivation

Large language models (LLMs) can solve tasks without explicit training GPT-3

Fine-tuning LLMs on groups of tasks boosts zero-shot task generalization FLAN T0 MetaICL

Prior work has focused on English LLMs/tasks

Multilingual LLMs show zero-shot abilities

XLM-R$_{XXL}$ XGLM mT5+SAP AlexaTM

But zero-shot trails task/lang. specific finetuning

Hard to address for low-resource languages and tasks

Goal: study multilingual multitask finetuning for zero-shot task generalization on non-English tasks

Key findings:

(1) English multitask finetuning helps non-English tasks

(2) Multingual finetuning data further helps

(3) Larger models benefit more from multitask finetuning

(4) Finetuning helps tasks on rarely seen languages

References:

(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)

(FLAN) J. Wei et al., "Finetuned Language Models are Zero-Shot Learners", ICLR (2022)

(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)

(MetaICL) S. Min et al., "MetaICL: Learning to learn in context", arXiv (2021)

(XLM-R$_{XXL}$) N. Goyal et al., "Larger-scale transformers for multilingual masked language modeling", arxiv (2021)

(XGLM) V. Lin et al., "Few-shot learning with multilingual language models", arxiv (2021)

(mT5+SAP) A. Patel et al., "Bidirectional Language Models Are Also Few-shot Learners", arxiv (2022)

(AlexaTM) S. Soltan et al., "AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model", arxiv (2022)

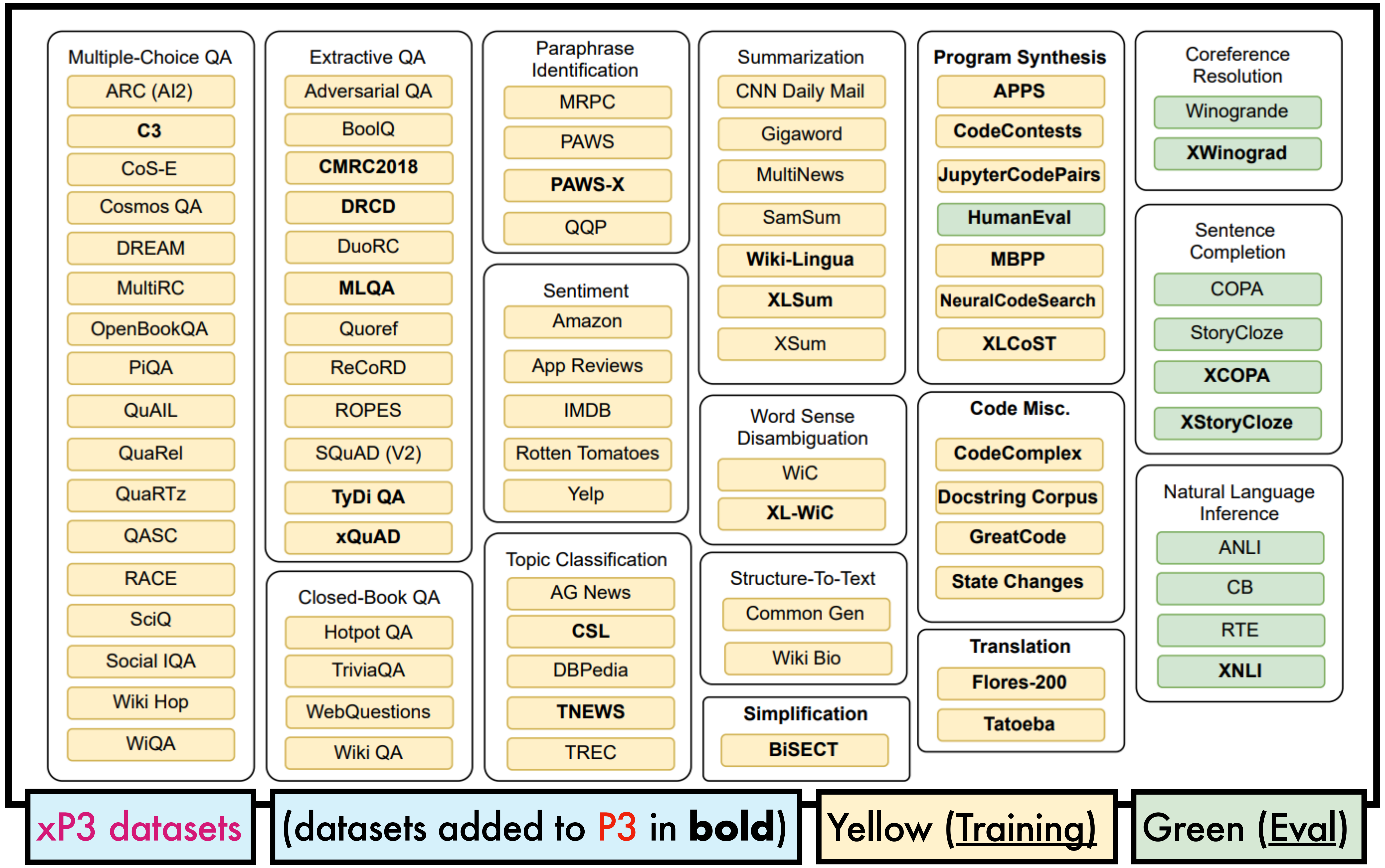
Finetuning data

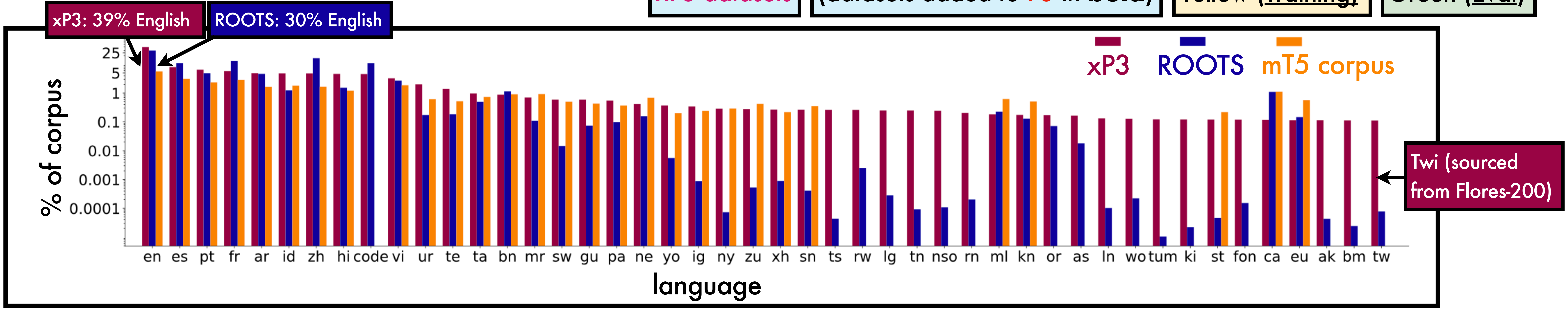
xP3: adds 28 multilingual datasets to Public Pool of Prompts (P3) used by T0

New task clusters: Translation Simplification Program Synthesis Code Miscellaneous

xP3 aims to replicate ROOTS distribution (same 46 languages and code are used)

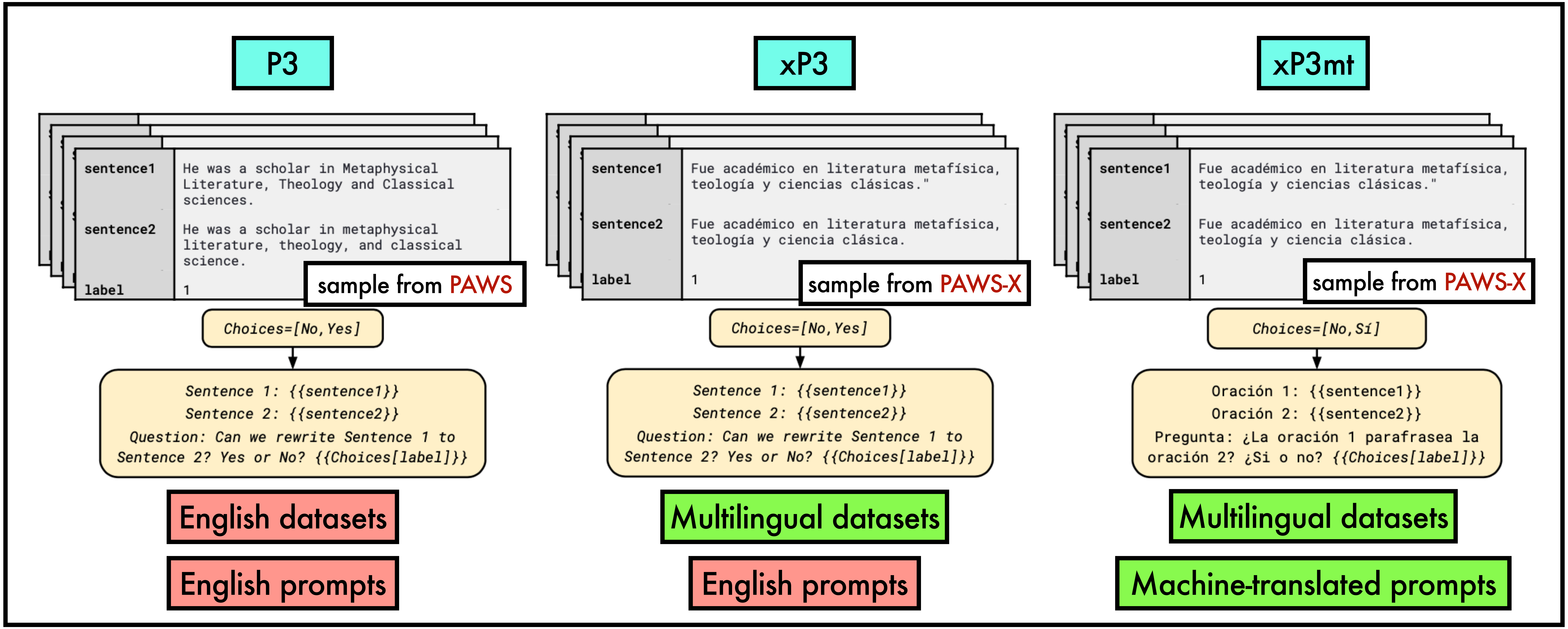
Non-English: xP3mt (machine translations)





References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
(ROOTS) H. Laurençon et al., "The BigScience ROOTS Corpus: A 1.6 TB Composite Multilingual Dataset", NeurIPS Datasets Track (2022)
(mT5) L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer", arxiv (2020)
(Flores-200) M. Costa-jussà et al., "No language left behind: Scaling human-centered machine translation", arxiv (2022)

Dataset variants



References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(PAWS) Y. Zhang et al., "PAWS: Paraphrase Adversaries from Word Scrambling", NAACL-HLT (2019)
(PAWS-X) Y. Yang et al., "PAWS-X: A cross-lingual adversarial dataset for paraphrase identification", EMNLP (2019)

Finetuned models

Two families of models are finetuned:

BLOOM

- Decoder-only (560M → 176B params)
- Pretrained on ROOTS for 350B tokens
- Finetuned for 13B tokens
- Loss is only enforced on target tokens

Input	Target
Translate to English: Je t'aime.	I love you.

Loss enforced here

mT5

- Encoder-decoder (300M → 13B params)
- Pretrained on mC4 for 1T tokens
- Similar finetuning to BLOOM
- Inputs are fed into the encoder (rather than space-separated)

Finetuned variants

Finetuned on **P3**		Finetuned on **xP3**		Finetuned on **xP3mt**	
BLOOMZ-P3	mT0-P3	BLOOMZ	mT0	BLOOMZ-MT	mT0-MT

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(BLOOM) Big Science Workshop, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model" (2022)
(ROOTS) H. Laurençon et al., "The BigScience ROOTS Corpus: A 1.6 TB Composite Multilingual Dataset", NeurIPS Datasets Track (2022)
(mT5) L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer", arxiv (2020)

Evaluation details

Evaluate on three held-out task clusters:

Coreference Resolution | Sentence Completion | Natural Language Inference

Additional evaluation for program synthesis (not held-out task cluster): HumanEval

Rank classification used for selection (score log-likelihood of completions) like GPT-3, T0, etc.

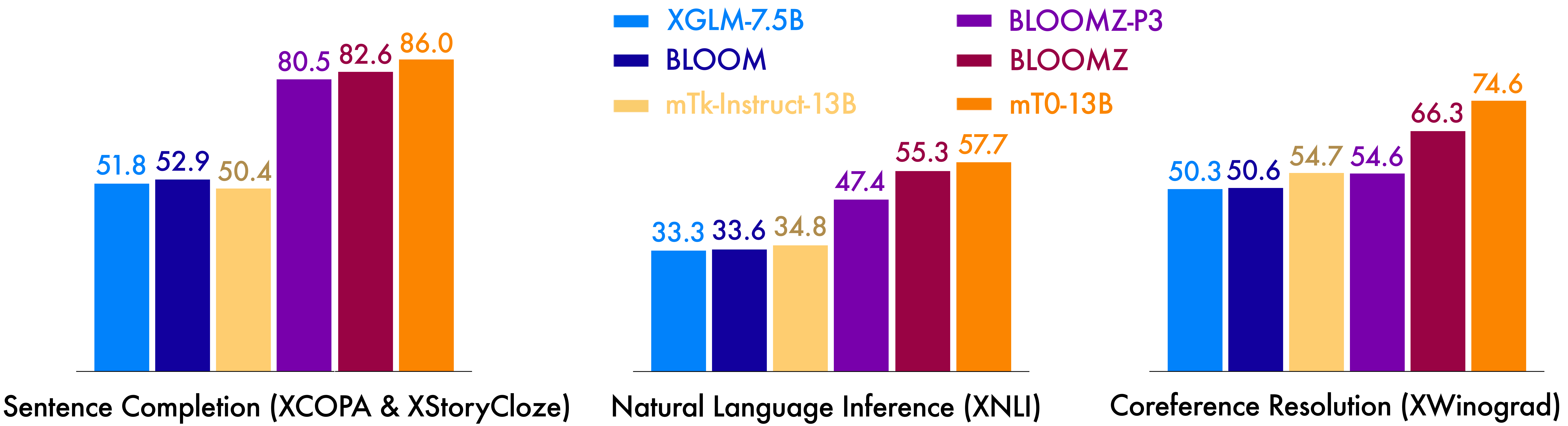
For each eval dataset: 5 prompts randomly chosen from PromptSource & used for all language splits

The median score of the 5 prompts per language split is reported

Generation evaluations are assessed with the LM-evaluation-harness implementation

References:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(HumanEval) M. Chen et al. "Evaluating large language models trained on code", arxiv (2021)
(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
(LM-Evaluatio-harness) L. Gao et al., "A framework for few-shot language model evaluation." Version v0.0.1 (2021)

Results: Zero-shot multilingual task generalization



Observations

mT0-13B outperforms BLOOMZ

mT0-13B outperforms mTk-Instruct-13B

xP3 outperforms P3 for finetuning

also holds on English tasks

Potential explanations

accords with T. Wang et al. (2022); Tay et al. (2022)

longer pretraining

encoder-decoder

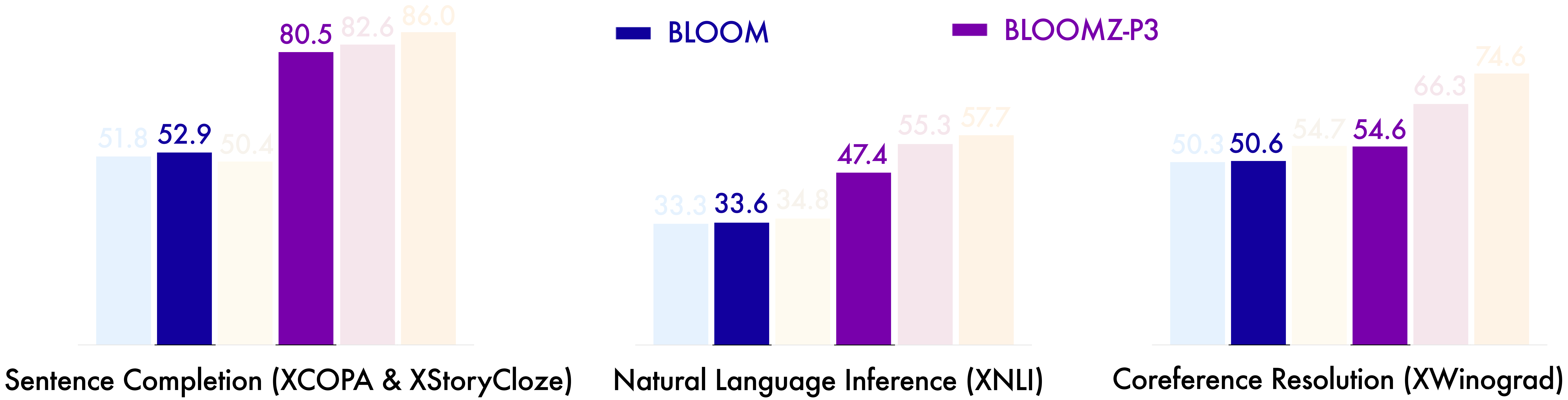
mask objective

Difference in prompting style (structured vs informal)

Extra tasks/prompts help generalization (Chung et al., 2022)

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(mTk-instruct) Y. Wang et al., "Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks", arxiv (2022)
T. Wang et al., "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?", arxiv (2022)
Y. Tay et al., "Unifying Language Learning Paradigms", arxiv (2022)
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (2022)

Results: Language generalization



Observation BLOOMZ-P3 outperforms BLOOM

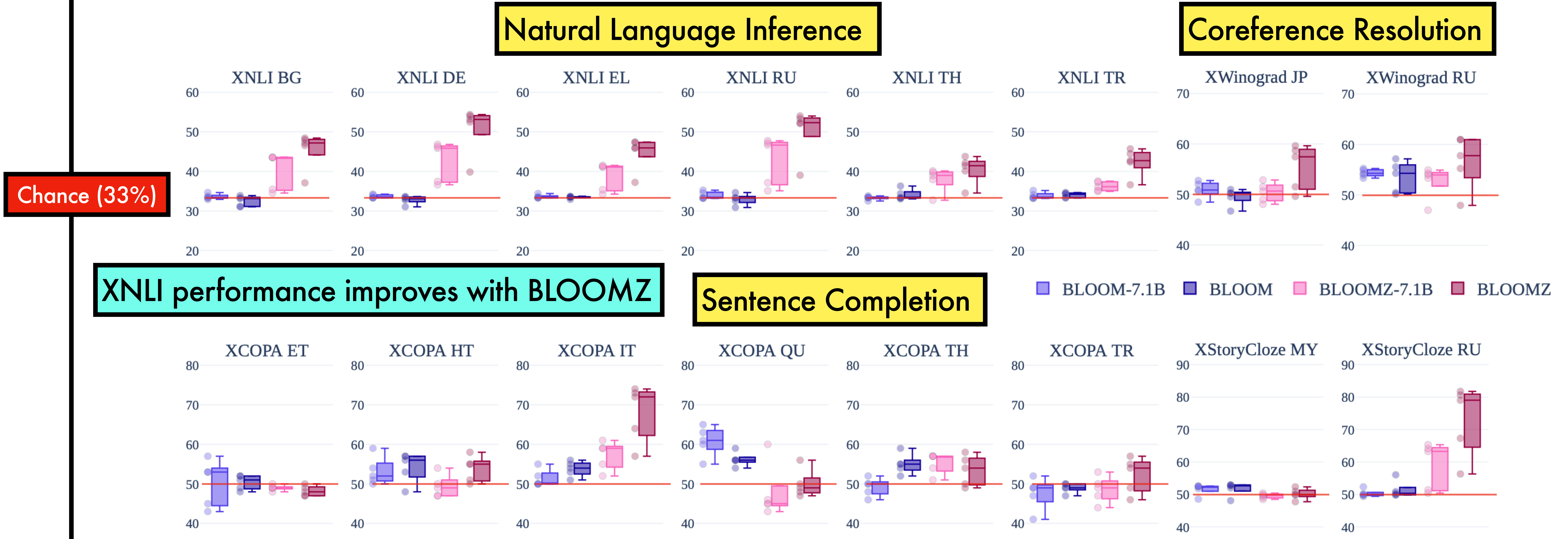
Takeaway Zero-shot multilingual task performance improves after finetuning on English-only prompts on languages were only seen during pretraining

Significant potential benefits (difficult to collect low-resource language data)

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)

Results: Language generalization (cont.)

Study performance on languages "never intentionally seen" (due to scale it's hard to be sure)



In practice, tiny fractions of these languages are found in pretraining (e.g. Thai constitutes 0.006% of data)

Note: XNLI labels correlate with Levenshtein edit distance (character inspection could outperform chance)

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)

Results: Multilingual prompting

To study performance on non-English prompts, xP3 prompts are translated (xP3mt)

Task	Prompt	Average accuracy			
		BLOOMZ	BLOOMZ-MT	mT0-13B	mT0-13B-MT
XNLI	EN	**53.58**	49.74	48.43	**51.52**
	MT	37.87	**42.03**	39.83	**42.64**
	HT	41.13	**44.55**	45.19	**47.03**
XCOPA	EN	75.5	**75.75**	**84.45**	81.6
	MT	71.95	**74.25**	**82.9**	81.1
XStoryCloze	EN	**84.42**	84.07	82.52	**82.58**
	MT	84.37	**85.31**	**84.01**	83.31
XWinograd	EN	**60.07**	59.15	70.49	**73.24**
	MT	58.48	**60.14**	66.89	**72.33**

human translations work better

BLOOMZ-MT outperforms BLOOMZ on non-English (translated) prompts, but worse on English

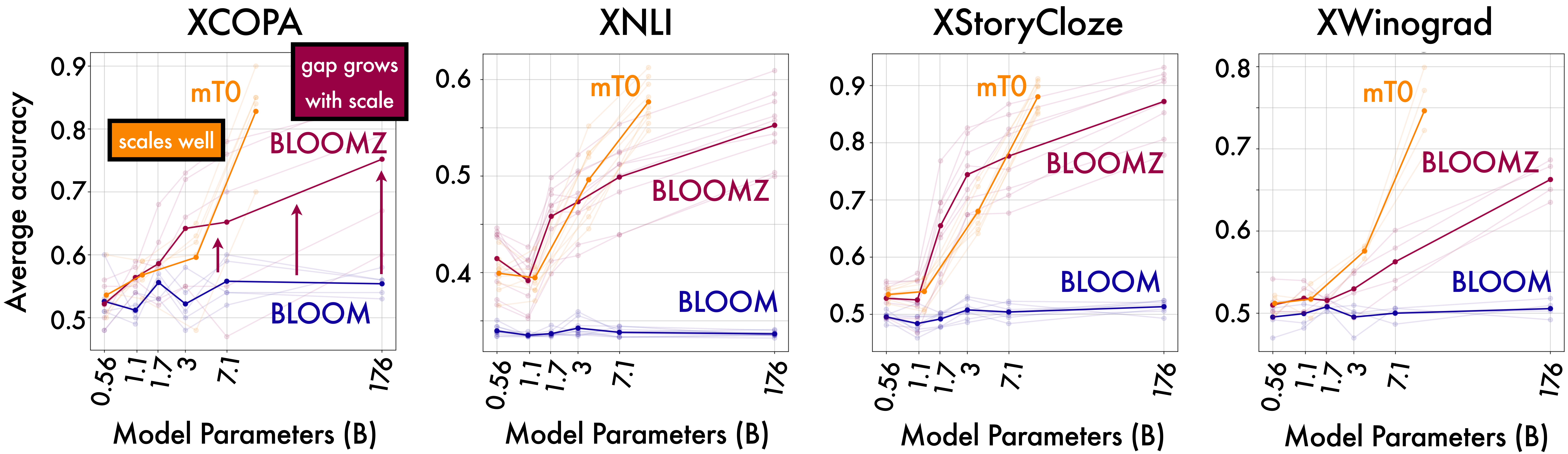
Results on mT0-13B are mixed

Human-translated prompts achieve better scores than machine-translated prompts

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)

Results: Scaling multitask finetuning

Study of the interaction of model scale and multitask finetuning



mT0 improves particularly rapidly with model size

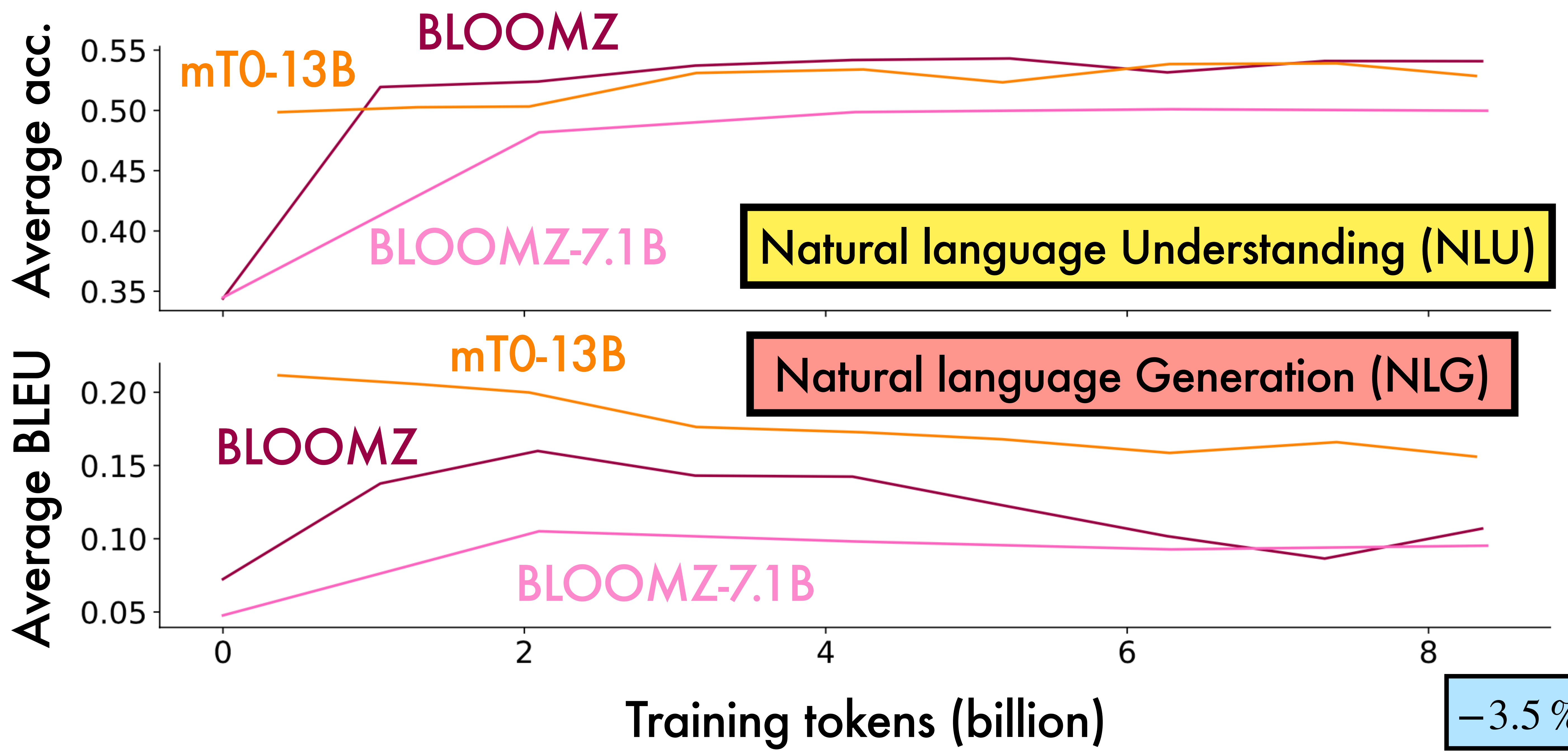
The gap between BLOOM and BLOOMZ grows with model size

Multitask finetuning benefits larger models more

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)

Influence of multitask finetuning on generative tasks

How does generation evolve with multitask finetuning?



NLU tasks improve under finetuning

NLG tasks worsen as finetuning continues longer

HumanEval (function completion)

higher is better

	Pass@k $k=1$	$k=10$	$k=100$
GPT-Neo 1.3B	4.79%	7.47%	16.30%
GPT-Neo 2.7B	6.41%	11.27%	21.37%
GPT-J 6B	11.62%	15.74%	27.74%
GPT-NeoX 20B	15.4%	25.6%	41.2%
Codex-300M	13.17%	20.37%	36.27%
Codex-679M	16.22%	25.7%	40.95%
Codex-2.5B	21.36%	35.42%	59.5%
Codex-12B	28.81%	46.81%	72.31%
BLOOM-560M	0.82%	3.02%	5.91%
BLOOM-1.1B	2.48%	5.93%	9.62%
BLOOM-1.7B	4.03%	7.45%	12.75%
BLOOM-3B	6.48%	11.35%	20.43%
BLOOM-7.1B	7.73%	17.38%	29.47%
BLOOM	15.52%	32.20%	55.45%
BLOOMZ-560M	2.18 %	4.11%	9.00%
BLOOMZ-1.1B	2.63%	6.22%	11.68%
BLOOMZ-1.7B	4.38%	8.73%	16.09%
BLOOMZ-3B	6.29%	11.94%	19.06%
BLOOMZ-7.1B	8.06%	15.03%	27.49%
BLOOMZ	12.06%	26.53%	48.44%
BLOOMZ-P3	6.13%	11.79%	18.73%

$-3.5\,\%$

(no code)

Performance drops for BLOOMZ

Possible cause for generative drop: finetuning encourages shorter completions

Partial mitigation: force longer generation (ignore <EOS>)

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(HumanEval) M. Chen et al., "Evaluating large language models trained on code", arxiv (2021)

Code generation

BLOOM generation

```python
from typing import List


def has_close_elements(numbers: List[float], threshold: float) -> bool:
    """ Check if in given list of numbers, are any two numbers closer to each other than
    given threshold.
    >>> has_close_elements([1.0, 2.0, 3.0], 0.5)
    False
    >>> has_close_elements([1.0, 2.8, 3.0, 4.0, 5.0, 2.0], 0.3)
    True
    """
    if len(numbers) <= 1:
        return False
    numbers = sorted(numbers)
    i = 0
    while i < len(numbers) - 1:
        # Get distance between current and next elements
        distance = numbers[i] - numbers[i + 1]

        # Check if distance is closer than threshold
        if distance < threshold:
            return True
        i += 1
    return False
```

BLOOMZ generation

```python
from typing import List


def has_close_elements(numbers: List[float], threshold: float) -> bool:
    """ Check if in given list of numbers, are any two numbers closer to each other than
    given threshold.
    >>> has_close_elements([1.0, 2.0, 3.0], 0.5)
    False
    >>> has_close_elements([1.0, 2.8, 3.0, 4.0, 5.0, 2.0], 0.3)
    True
    """
    numbers = sorted(numbers)
    return any(numbers[i + 1] - numbers[i] < threshold for i in range(len(numbers) - 1))
```

BLOOMZ is biased towards more concise code

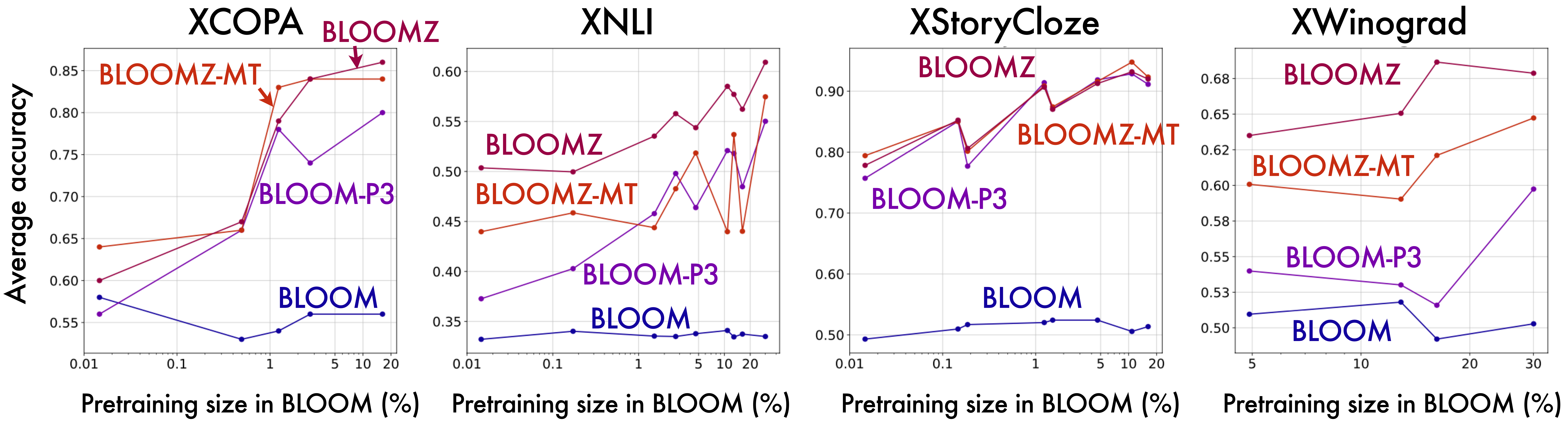
Data (→)	HumanEval generations BLOOM	HumanEval generations BLOOMZ	xP3 code datasets
Average characters	247	144	530
Average Python comments (#)	0.69	0.04	0.85

BLOOMZ produces 17x fewer comments

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(HumanEval) M. Chen et al., "Evaluating large language models trained on code", arxiv (2021)

Influence of language proportions

How do language proportions in pretraining affect the performance of multitask-finetuned models?



Finetuned models (broadly) do better on languages better represented in pretraining (ROOTS)

Note: xP3 aims to match the ROOTS language distribution

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(XCOPA) E. Ponti et al., "XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning", EMNLP, (2020)
(ROOTS) H. Laurençon et al., "The BigScience ROOTS Corpus: A 1.6 TB Composite Multilingual Dataset", NeurIPS Datasets Track (2022)
(XNLI) A. Conneau et al., "XNLI: Evaluating Cross-lingual Sentence Representations", EMNLP (2018)
(XStoryCloze) X. Lin et al., "Few-shot learning with multilingual language models", arxiv (2021)
(XWinograd) A. Tikhonov et al. "It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning", ACL/IJCNLP (2021)

Qualitative results

Forcing longer generations

Input:

Write a fairy tale about a troll saving a princess from a dangerous dragon. The fairy tale is a masterpiece that has achieved praise worldwide and its moral is "Heroes Come in All Shapes and Sizes". Story (in Spanish):

BLOOMZ: **Un reino en el bosque estaba bajo un terrible dragón. Todos los que intentaron derrotar al dragón murieron. El rey y la reina le pedían al pueblo que les diera un niño, que les gustaría salvar el reino de la maldición del dragón. Un día, la reina dio a luz a un niño con una cabeza enorme y ojos que brillaban como estrellas. El rey y la reina llamaron al niño Elliot. Elliot crecería y se convertiría en un joven gigante. Él salvaría al pueblo.**

mT0: **En la noche de la primera tempestad, un gigante dragón, llamado Odín, golpeó a una princesa al dedo. Así que un troll, llamado Odín, fue a salvar a la princesa. Luego, este fue querido por todos.**

Machine translations

A kingdom in the forest was under a terrible dragon. All those who tried to defeat the dragon died. The king and queen asked the people to give them a child, that they would like to save the kingdom from the curse of the dragon. One day, the queen gave birth to a boy with a huge head and eyes that shone like stars. The king and queen named the boy Elliot. Elliot would grow up to be a young giant. He would save the town.

On the night of the first storm, a giant dragon, named Odin, struck a princess to the finger. So a troll, named Odin, went to save the princess. Then this one was loved by everyone.

Input:

Why is the sky blue?

BLOOM:	**"Why is the sky blue?" "Why is the sky blue?" "Why is the sky blue?"...**
BLOOMZ:	**Rayleigh scattering.**
BLOOMZ with 10 tokens min length:	**Rayleigh scattering. The blue color of the sky is due to Rayleigh scattering of sunlight by atmospheric molecules.**
mT0-13B:	**because of the sun**
mT0-13B with 10 tokens min length:	**because of the sun's radiation sensitivity to light**

Answer to "Why is the sky blue?" https://xkcd.com/1818/

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(Google translate) https://translate.google.co.uk/

Related Work

Multitask prompted finetuning (or "Instruction finetuning")

Brings boost to zero-shot task generalisation T0 (2022) Flan (2022)

This work builds on T0 (2022) PromptSource (2022)

Increasing scale/diversity of tasks and datasets brings benefits

MetaICL (2021) BigBIO (2022) CT0 (2022) Flan-PaLM (2022) Natural Instructions (2022)

Most prior work has focused on English prompts and datasets

Concurrent work Super-NaturalInstructions (2022) train English & multilingual models on prompted datasets (Tk-Instruct/mTk-Instruct)

Comparisons are reported with BLOOMZ, mT0 and mTk-Instruct

"Never intentionally seen" generalisation

GPT-2 (2019) discovered that unintentional inclusion of non-English text enables some non-English generation

This work: finetuning supports languages "not intentionally seen"

Multilingual models

Aim to enable processing of multiple languages XLM (2019) XLM-R (2019) M2M-100 (2021)

Two particular families of architectures have emerged:

- Encoder-decoder mBART (2020) mT5 (2020)
- Decoder-only mGPT (2022) XGLM (2021) BLOOM (2022)

This work uses mT5 (2020) (mT0) BLOOM (2022) (BLOOMZ)

References:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
(FLAN) J. Wei et al., "Finetuned Language Models are Zero-Shot Learners", ICLR (2022)
(PromptSource) S. Bach et al., "PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts", ACL Demo (2022)
(MetaICL) S. Min et al., "MetaICL: Learning to learn in context", arXiv (2021)
(BigBIO) J. Fries et al. "BigBIO: A Framework for Data-Centric Biomedical Natural Language Processing", arxiv (2022)
(CT0) T. Scialom et al., "Continual-T0: Progressively Instructing 50+ Tasks to Language Models Without Forgetting", arxiv (2022)
(Flan-PaLM) H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (2022)
(Natural Instructions) S. Mishra et al., "Natural instructions: Benchmarking generalization to new tasks from natural language instructions", ACL (2022)
(Super-NaturalInstructions) Y. Wang et al., "Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks", arxiv (2022)
(GPT-2) A. Radford et al., "Language models are unsupervised multitask learners", Technical Report (2019)
(XLM) A. Conneau and G. Lample, "Cross-lingual language model pretraining", NeurIPS (2019)
(XLM-R) A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale", ACL (2020)
(M2M-100) A. Fan et al., "Beyond English-Centric Multilingual Machine Translation", JMLR (2021)
(mBART) Y. Liu et al., "Multilingual denoising pre-training for neural machine translation", ACL (2020)
(mT5) L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer", arxiv (2020)
(mGPT) O. Shliazhko et al., "mGPT: Few-Shot Learners Go Multilingual", arxiv (2022)
(XGLM) V. Lin et al., "Few-shot learning with multilingual language models", arxiv (2021)
(BLOOM) Big Science Workshop, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model" (2022)

Takeaways and released assets

Takeaways

Finetuning on English-only data (P3) helps a multilingual model generalize to tasks in other languages

Finetuning on multilingual data (xP3) brings even better performance

Finetuned models can generalize to languages "never intentionally seen" to some extent

Finetuning on machine-translated prompts can improve non-English prompt performance

Larger models benefit more from finetuning

Finetuning on short targets may bias models to produce short outputs (hurts generative performance)

Released assets

xP3 - a corpus of tasks in 46 languages (together with machine-translated prompts, xP3mt)

All trained models produced in this work

Released models include BLOOMZ & mT0

References/image credits:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)

Nuts and bolts of finetuning

BLOOM finetuning

The token loss is scaled to the length of the target it belongs to

E.g. multiple-choice QA targets are much shorter than translation targets

Samples longer than 2048 tokens are skipped

Packing is used to train efficiently on multiple samples concurrently

The final checkpoint is chosen based on validation performance

mT5 finetuning

mostly similar to BLOOM, but inputs are processed by the encoder

Implemented with the T5X framework on TPUs

References:
N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(Packing) M. Kosec et al., "Packing: Towards 2x nlp bert acceleration", arxiv (2021)
(T5X) A. Roberts et al., "Scaling Up Models and Data with t5x and seqio", arxiv (2022)