Crosslingual Generalization through Multitask Finetuning

& xP3

PAPER

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, Colin Raffel

Motivation: Instruction Tuning & Multilingual Models

*Instruction-tuning teaches models to **perform natural language tasks** to make them useful.*

Sanh et al. (2021), Wei et al. (2021), Ouyang et al. (2022), inter alia

What is the capital of Japan?

Pre-trained Model: , Where is the capital of Japan? How many islands are there in Japan? These are all great questions that require a more complex answer than a name or a number and…

Instruction-tuned: Tokyo

---

*Multilingual models **can write in dozens of languages***

T5: 🇬🇧

mT5: 🇬🇧🇷🇺🇪🇸🇩🇪🇫🇷🇮🇹🇵🇹🇵🇱🇳🇱🇹🇷🇯🇵🇻🇳🇮🇩🇨🇿🇨🇳…

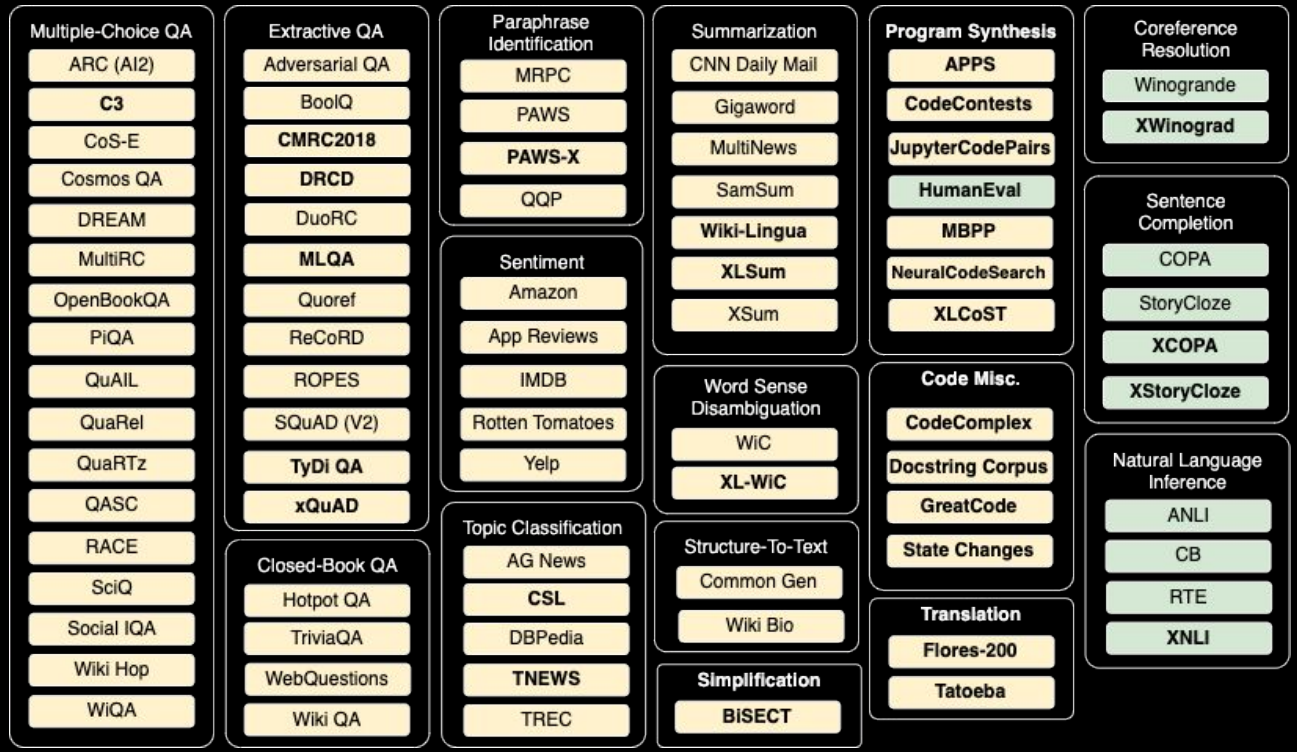
BLOOM: 🇬🇧🇨🇳🇫🇷🇪🇸💻🇵🇹…

Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., 2022)

BLOOMZ & mT0

xP3: Crosslingual Corpus of prompted tasks in up to 277 languages

xP3 Task Taxonomy | Yellow: Training | Green: Evaluation

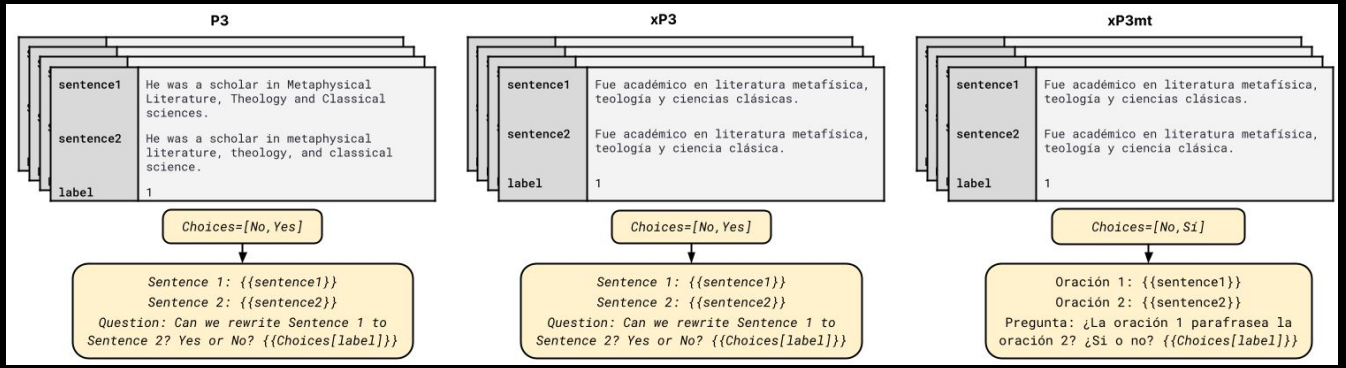


Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., 2022)

BLOOMZ & mT0

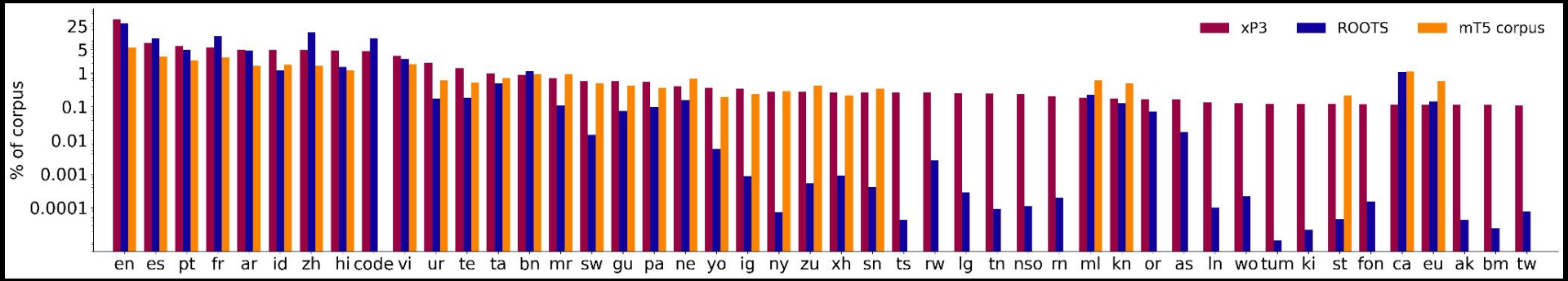
xP3: Crosslingual Corpus of prompted tasks in up to 277 languages

Dataset variations + *xP3x (2023)*



xP3 eXtension:

- 🌍 277 languages
- 🎡 700GB
- 🤯 500M samples



Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., 2022)

BLOOMZ & mT0

Multitask Fine-tuning / Instruction tuning

Sanh et al. (2021), Wei et al. (2021), Ouyang et al. (2022), inter alia

1. Pretrained Language Model (LM)

BLOOM: 560M - 176B parameter *decoder* LMs trained on *48 languages*

mT5: 300M - 13B parameter *encoder-decoder* LMs trained on *101 languages*

2. Finetune on many tasks with instructions

P3: English instructions; English data

xP3: English instructions; Multilingual data

xP3mt: Multilingual instructions; Multilingual data

3. Evaluate on tasks *NOT* trained on

Sentence Completion: XCOPA & XStoryCloze

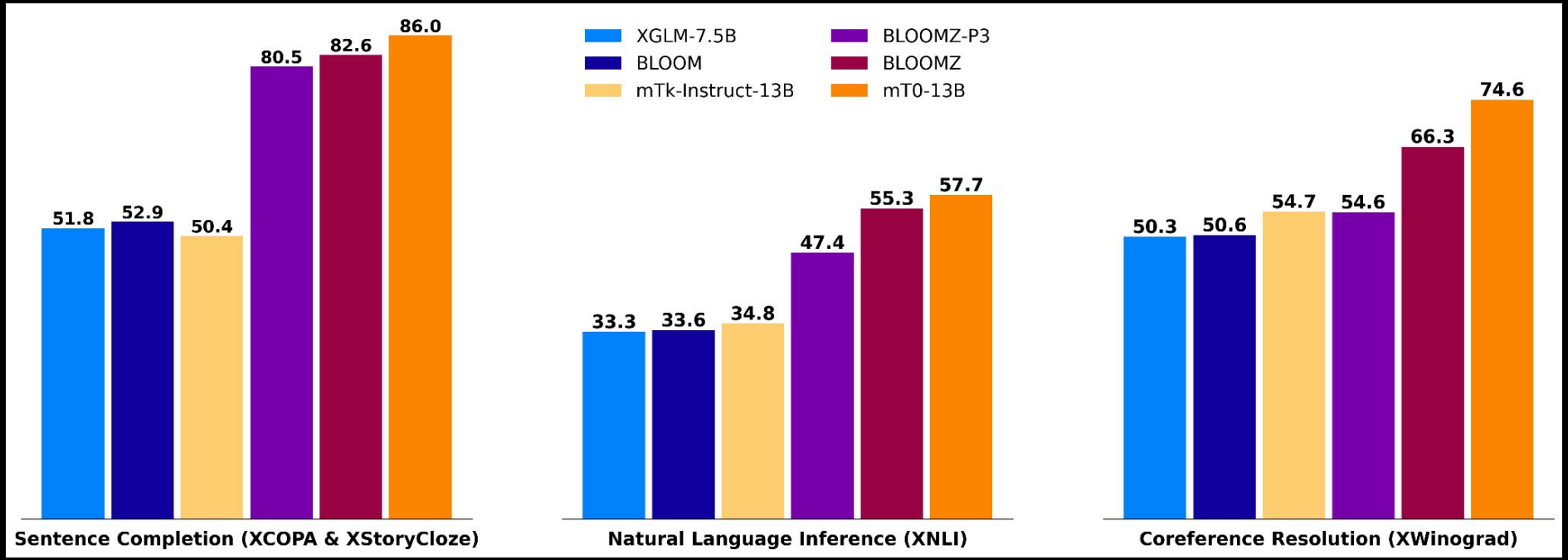
Natural Language Inference: XNLI

Coreference resolution: XWinograd

Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., 2022)

BLOOMZ & mT0

BLOOMZ & mT0: Generalization to *unseen* Tasks



Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., 2022)

BLOOMZ & mT0

BLOOMZ & mT0: Generalization to *unseen* Tasks

Why is the sky blue? Give me a detailed explanation in Latin American Spanish. El cielo azul se debe a la dispersión de la luz solar por las moléculas de gas en la atmósfera. La luz azul es dispersada más que otros colores, por lo que el cielo se ve azul.

Input Output

Computed in 4 seconds

BLOOMZ & mT0

మీరు మనసులో ఏం ఫీలవుతున్నారో అది మీ ముఖంలో కనిపిస్తుంది. అందుకే ఎప్పుడూ పాజిటివ్‌గా ఆలోచిస్తూ ఆనందంగా ఉండే ప్రయత్నం చేయాలి. Translate to Vietnamese. Cảm xúc của bạn sẽ được thể hiện rõ trên gương mặt, vì thế hãy luôn cố gắng duy trì thái độ tích cực và mỉm cười nhé!

Input Output

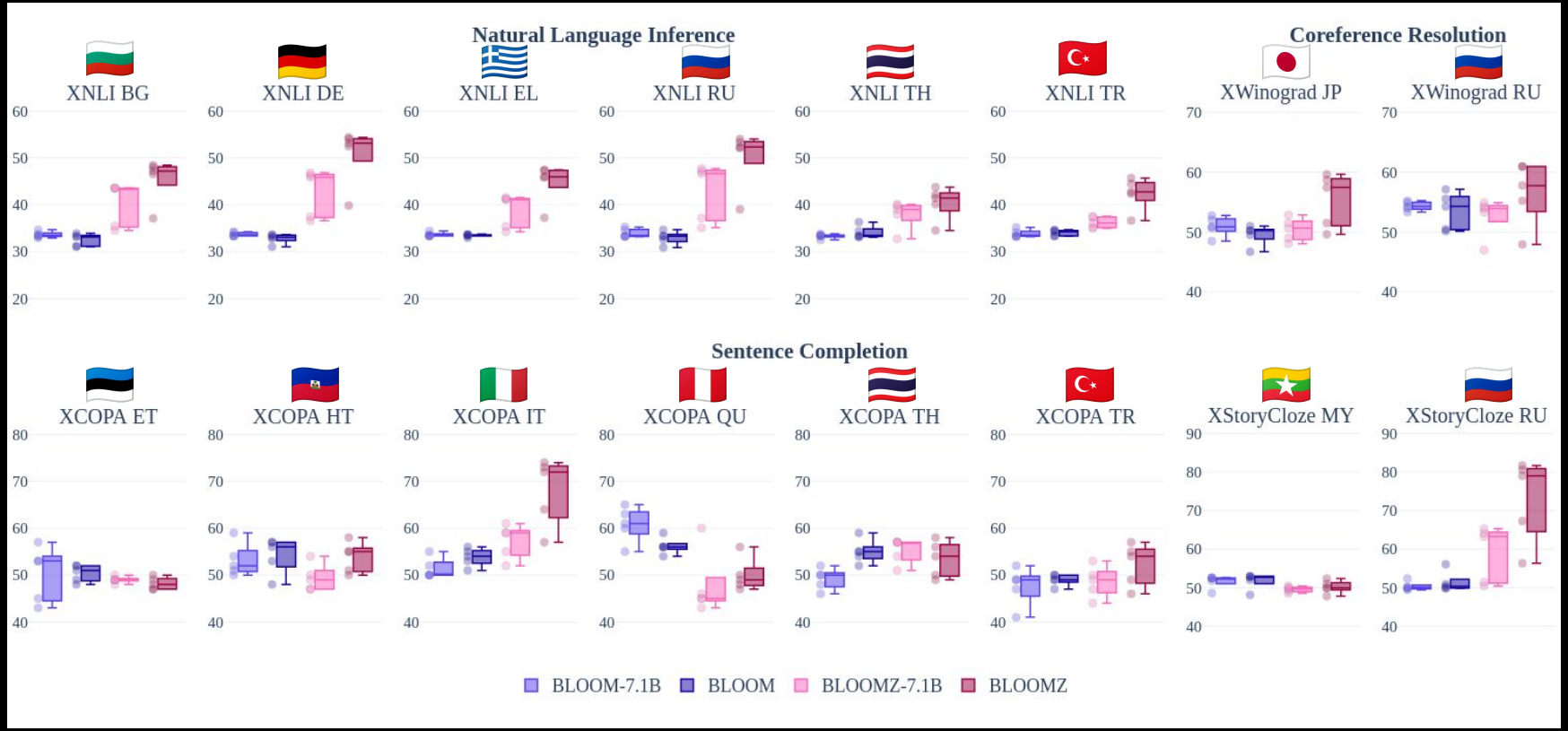
Computed in 3 seconds

BLOOMZ & mT0

Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., 2022)

BLOOMZ & mT0

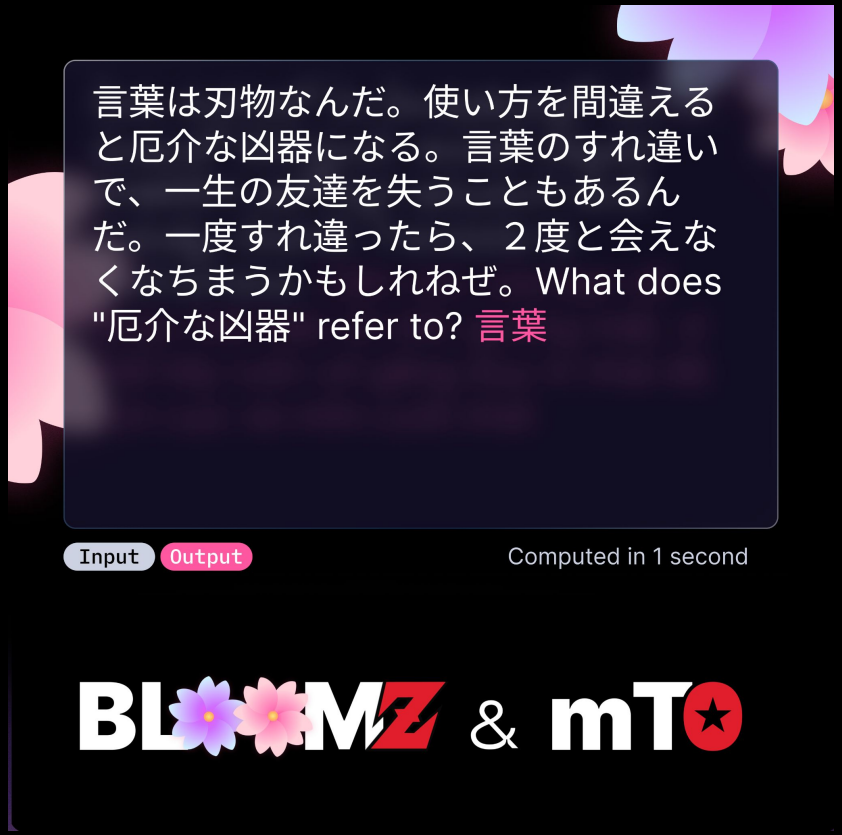
BLOOMZ & mT0: Generalization to *unseen* Tasks in *"unseen"* Languages



Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., 2022)

BLOOMZ & mT0

BLOOMZ & mT0: Generalization to *unseen* Tasks



Crosslingual Generalization through Multitask Finetuning (Muennighoff et al., 2022)

BLOOMZ & mT0

Thanks!

Check the paper for *more details* on:

- Impact of non-English instructions (xP3mt)
- Generative tasks such as code generation
- Impact of pre-training language distribution
- Language contamination

…

Paper:
arxiv.org/abs/2211.01786

Open-source **code/models/data:**
github.com/bigscience-workshop/xmtf

BLOOMZ & mT0

& xP3

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, Colin Raffel