LANGBRIDGE: Multilingual Reasoning Without Multilingual Supervision

Anonymous ACL submission

Abstract

We introduce LANGBRIDGE, a *zero-shot* approach to adapt language models for multilingual reasoning tasks without multilingual supervision. LANGBRIDGE operates by "bridging" two models, each specialized in different aspects: (1) one specialized in understanding multiple languages (e.g., mT5 encoder) and (2) one specialized in reasoning (e.g., Orca 2). LANGBRIDGE connects the two models by introducing minimal trainable parameters between them. Despite utilizing only English data for training, LANGBRIDGE considerably enhances the performance of language models on low-resource languages across mathematical reasoning, coding, and logical reasoning. Our analysis suggests that the efficacy of LANGBRIDGE stems from the language-agnostic characteristics of multilingual representations. We publicly release our code and models.[1]

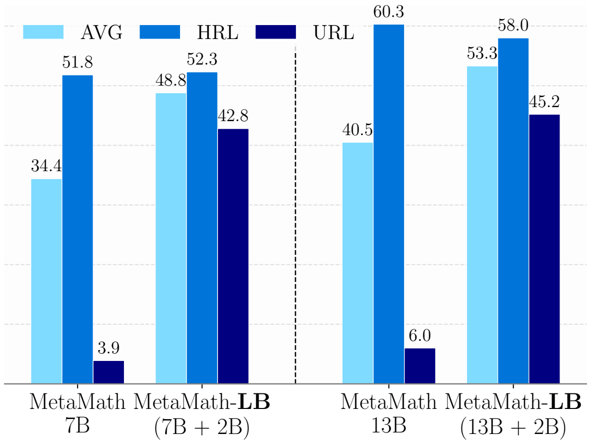

Figure 1: MGSM accuracy (%) of MetaMath models and models aligned with mT5-XL encoder (2B) via LANGBRIDGE (**LB**). In addition to the average (AVG) accuracy, we also report the average accuracy of high-resource languages (HRL) and underrepresented languages (URL) classified by Shi et al. (2023).

1 Introduction

Language models (LMs) are known to exhibit inferior performance in solving reasoning tasks such as math or coding in low-resource languages (Shi et al., 2022; Qin et al., 2023). This tendency primarily stems from the fact that LMs are predominantly trained on corpora comprised of a few high-resource languages (Touvron et al., 2023a,b). This results in low-resource languages being represented as long-tail knowledge (Lazaridou et al., 2021; Kandpal et al., 2023).

Prior works have mainly approached this problem by adapting English-centric LMs to other languages through continual training on the target language (Marchisio et al., 2023; Oba et al., 2023; Zhu et al., 2023; Kew et al., 2023). However, scaling this approach to a large number of languages is challenging, as it requires targeted training corpora for each language. This issue is particularly pronounced for LMs such as MetaMath (Yu et al., 2023) and Orca 2 (Mitra et al., 2023), which have undergone continuous domain-specific adaptation from Llama 2 (Touvron et al., 2023b). These specialized, domain-specific datasets are typically in English, complicating multilingual support for the underlying LM.

In this paper, we introduce LANGBRIDGE, a novel approach that adapts LMs to solve multilingual reasoning tasks without *explicitly* training on multilingual data. Inspired from the multimodal literature that integrates two independently pretrained modalities (Alayrac et al., 2022; Li et al., 2023a; Merullo et al., 2023; Liu et al., 2023b; Bavishi et al., 2023), we leverage the encoder from mT5 (Xue et al., 2021) and introduce a small number of trainable parameters between the encoder and the target LM. Most importantly, our approach does not require multilingual supervision and solely relies on English data while generalizing to multiple languages during test time, resembling zero-shot cross-lingual transfer (Pires et al., 2019; Conneau et al., 2020; Xue et al., 2021; Li and Murray, 2023).

[1] anonymous.4open.science/r/LangBridge-5334

We demonstrate the effectiveness of LangBridge by applying our method to LMs specialized in mathematical reasoning, coding, and logical reasoning. Our empirical results show LangBridge substantially enhances the multilingual reasoning performance of LMs. Notably, LangBridge applied to MetaMath-13B leveraging mT5-XL encoder (2.2B) boosts the average accuracy on MGSM (Shi et al., 2023) from 40.5% to 53.5%, matching the performance of PaLM-540B (Chowdhery et al., 2023), which stands at 51.3%.

We hypothesize that the effectiveness of LangBridge is anchored in the language-agnostic characteristics of multilingual representations (Pires et al., 2019; Libovický et al., 2020). By mapping these representations to the LMs' input space, we conjecture that the LM is able to grasp the semantics of these representations. As the representations are language-neutral, the comprehension makes the specific language of the input inconsequential. Our empirical analysis of LangBridge, using principal component analysis (PCA) and qualitative methods, supports this hypothesis.

2 Related Work

2.1 English-centric Language Models

Previous works have enhanced the reasoning capabilities of LMs in mathematical reasoning, code generation, and logical reasoning (Mukherjee et al., 2023; Azerbayev et al., 2023; Yu et al., 2023; Rozière et al., 2023; Mitra et al., 2023). However, the majority of these LMs are derived from English-centric LMs (Touvron et al., 2023a,b) and adapted with English domain-specific datasets. As a result, these LMs inherit limited proficiency in low-resource languages, which results in a significant performance discrepancy between high-resource and low-resource languages. The motivation of our proposed method, LangBridge, is to narrow this gap.

2.2 Zero-shot Cross-lingual Transfer

Multilingual models demonstrate remarkable zero-shot cross-lingual transfer capability, enabling them to perform tasks in languages different from the ones used during their finetuning stage (Conneau et al., 2020; Xue et al., 2021; Li and Murray, 2023). This feature significantly aids the low-resource language community, allowing multilingual models to handle tasks across a wide range of languages after being finetuned on high-resource languages. Our method, which leverages multilingual models, exhibits a similar ability, as it can enhance the reasoning capabilities of LMs across multiple languages while solely relying on English data during adaptation.

2.3 Aligning Pretrained Representations

The concept of combining independently pretrained representations has been widely explored in the realm of cross-modal alignment (Alayrac et al., 2022; Li et al., 2023a; Merullo et al., 2023; Liu et al., 2023b; Bavishi et al., 2023). These works focus on aligning visual encoder models with LMs to facilitate the visual comprehension abilities of LMs. In a concurrent work, Bansal et al. (2024) align two large language models to augment each other's capabilities. One of their experiments showcases that aligning a multilingual LM with another LM can lead to improved performance in multilingual tasks. However, in contrast to our method, their approach employs multilingual supervision for aligning.

3 LangBridge

3.1 Hypothesis

Previous works show that representations of multilingual encoder models are moderately language-agnostic (or language-neutral) out-of-the-box, facilitating zero-shot cross-lingual transfer (Pires et al., 2019; Libovický et al., 2020). Based on this premise, we hypothesize that by aligning a multilingual encoder model to the space of an LM, the LM will be able to understand the semantics of the languages supported by the multilingual encoder *without* training on a large set of languages.

To test this hypothesis, we align multilingual encoder models with LMs using only English corpora (**Left** of Figure 2). Then, we evaluate the aligned models using non-English inputs to determine if the LMs exhibit enhanced capabilities in multilingual reasoning tasks (**Right** of Figure 2).

3.2 Model Architecture

Building on the findings of previous works, where effective cross-modal understanding has been achieved by mapping representations from non-linguistic modalities to the soft prompts (Lester et al., 2021) of LMs (Merullo et al., 2023; Liu et al., 2023b; Bavishi et al., 2023), LangBridge maps the final hidden states of multilingual encoders to the soft prompts of LMs to attain multilingual un-

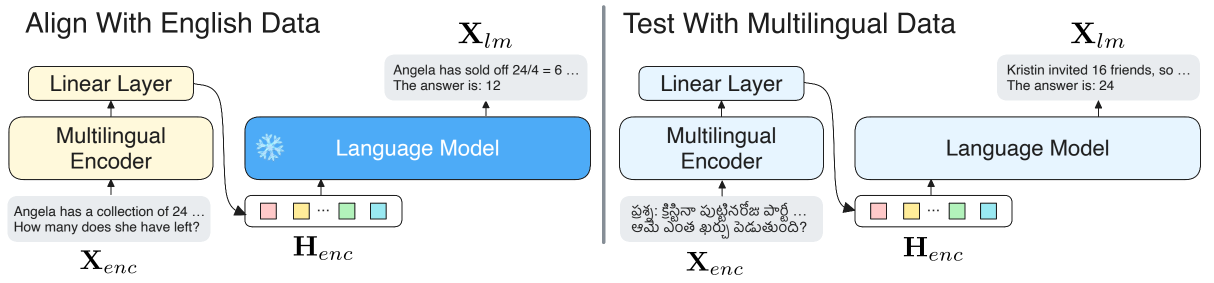

Figure 2: Overview of LANGBRIDGE. **Left**: A multilingual encoder with an added linear layer is aligned with the target language model using English data. We keep the language model frozen, whereas the linear layer is trainable. The multilingual encoder is trainable when adapting pretrained LMs and frozen when adapting finetuned LMs. **Right**: In test time, a LANGBRIDGE model can effectively solve multilingual reasoning tasks.

derstanding of the target LM. Following the studies above, we adopt the setting of using a single linear layer as the mapping function and append one trainable token to the end of the soft prompt as an `[EOS]` (end of sequence) token. Therefore, given the input token (padded if necessary) of the encoder $\mathbf{X}_{enc}$, the derived soft prompt $\mathbf{H}_{enc}$ is equivalent in sequence length to $\mathbf{X}_{enc} + 1$, and has the same dimensionality as the hidden state of the language model. Any tokens in $\mathbf{H}_{enc}$ that originate from padding tokens of $\mathbf{X}_{enc}$ are masked for the language model. We ablate the effect of using more complex architectures in Appendix D.

Language modeling objective of LANGBRIDGE resembles that of the "prefix LM" explored by Raffel et al. (2020), as it is conditioned on the prefix text given to the encoder. Formally, the language modeling likelihood of target token $\mathbf{X}_{lm}$ is denoted as:

$$p(\mathbf{X}_{lm}|\mathbf{X}_{enc}) = \prod_{i}^{L} p(x_i|\mathbf{X}_{enc}, x_{<i}) \qquad (1)$$

where L is the sequence length of $\mathbf{X}_{lm}$, and x_i is the ith token of $\mathbf{X}_{lm}$.

4 Main Experiments

4.1 Overview

We select three task categories for our main experiments: (1) **mathematical reasoning**, (2) **code completion**, and (3) **logical reasoning**. We additionally evaluate commonsense reasoning in Appendix A. For each task category, we apply LANGBRIDGE to LMs specialized in reasoning derived from Llama 2 (Touvron et al., 2023b), such as MetaMath or Orca 2.[2] We evaluate the models on existing multilingual benchmarks (e.g., MGSM) or translated English benchmarks (e.g., Bengali translation of Big Bench Hard (Suzgun et al., 2023)). As the evaluation tasks necessitate both multilingual understanding abilities and advanced reasoning capabilities, this complexity poses a significant challenge for general multilingual LMs and English-centric LMs specialized in reasoning. On the contrary, models aligned with LANGBRIDGE could take advantage of both.

Since Touvron et al. (2023b) discloses the language distribution of the pretraining data of Llama 2, this enables us to identify which languages are underrepresented in LMs initialized from Llama 2 weights. Throughout the paper, we classify a language as underrepresented if it comprises less than 0.1% of Llama 2's pretraining data.

In all of our experiments, we use the encoders of mT5 (Xue et al., 2021) as the multilingual encoders due to their availability across a wide range of parameters,[3] and their support for longer input sequences compared to other multilingual encoder models.[4] Specifically, we use the "LM adaptated" checkpoints from Vu et al. (2022). We align target LMs of 7B parameters with mT5-XL encoder to adapt 9B-sized models and 13B-sized LMs with mT5-XL encoder and mT5-XXL encoder to obtain 15B and 20B models, respectively. We ablate the effect of the encoder parameter size and encoder model type in Appendix D.

We use the original continual training data of the LM for LANGBRIDGE when accessible (e.g., MetaMathQA for MetaMath). If unavailable, we opt for the closest publicly accessible dataset (e.g., OpenOrca (Lian et al., 2023) for Orca 2). In all

[2] We test LANGBRIDGE on general-domain LMs (Llama 2 and Mistral 7B (Jiang et al., 2023)) in Appendix C.

[3] 270M (Small), 470M (Base), 820M (Large), 2.2B (XL) and 6.7B (XXL).

[4] mT5 was trained on input size of 1024 tokens, but can take longer sequences due to its use of relative position embeddings (Shaw et al., 2018).

our experiments, we restrict the size of the training dataset to 200,000 instances. However, our ablation on the effect of the training dataset size in Appendix D suggests LangBridge in practice may require much less data. We maintain the language model frozen throughout the alignment process to ensure efficiency. We also freeze the encoder (except the embedding layer) for aligning finetuned LMs, whereas for pretrained LMs, we keep the encoder trainable.[5] We provide further explanation and ablate the effect of freezing in Appendix D. We align the models by training on the prefix LM objective described in Section 3.2. In our preliminary experiments, we find that training on various lengths of $\mathbf{X}_{enc}$ is necessary to ensure robustness on inference time, as the language model is exposed to diverse lengths of $\mathbf{H}_{enc}$.

4.2 Mathematical Reasoning

4.2.1 Experimental Setup

Evaluation Datasets **MGSM** (Shi et al., 2023) comprises grade school math word problems in 11 typologically diverse languages, human translated from a sample of GSM8K (Cobbe et al., 2021). For evaluating pretrained LMs, we adopt the cross-lingual transfer chain-of-thought (CoT) reasoning (Wei et al., 2022) setting (NATIVE-EXEMPLARS + EN-COT) from Shi et al. (2023), where the few-shot exemplars are given in the target language, but the CoT rationales to solve the exemplars are provided in English. For finetuned LMs, we evaluate in zero-shot[6] setting. Additional evaluation on MSVAMP (Chen et al., 2023) is available in Appendix B.

Language Models **Llemma** (Azerbayev et al., 2023) is an LM for mathematics, continually pretrained from Code Llama (Rozière et al., 2023) on Proof-Pile-2, a mixture of scientific papers, web data containing mathematics, and mathematical code. **MetaMath** (Yu et al., 2023) was finetuned from Llama 2 (Touvron et al., 2023b) on MetaMathQA, a mathematical dataset based on GSM8K and MATH (Hendrycks et al., 2021b). As both Proof-Pile-2 and MetaMathQA are publicly available, we apply LangBridge using samples of their respective training datasets.

Baselines **Llama 2** (Touvron et al., 2023b) is an English-centric LM in which 89.7% of the pretraining data consists of English but has shown considerable performance on non-English languages (Lai et al., 2023). **mT5**[7] (Xue et al., 2021), **XGLM** (Lin et al., 2022), and **BLOOM** (Scao et al., 2022) are multilingual LMs. **MathOctopus** (Chen et al., 2023) is an LM for multilingual mathematical reasoning. It was initialized from Llama 2 and finetuned on translations of the GSM8K dataset across ten languages.[8] We use their best-performing checkpoints, xRFT-MathOctopus^{P}, which were further enhanced by data augmentation through rejection sampling (Yuan et al., 2023). We also report the performance of BLOOM models further trained on the training sets of LangBridge models, **BLOOM-Proof-Pile-2 (PP2)** and **BLOOM-MetaMath (MM)**. This is done to confirm that the capabilities of LangBridge models are derived from the LMs' inherent strength rather than solely from the training set utilized. We additionally report the performance of **PaLM** (Chowdhery et al., 2023) measured by Shi et al. (2023). Similar to Llama 2, PaLM was pretrained on English-heavy corpora.

4.2.2 Results

Table 1 shows the evaluation results of baselines and LangBridge models on MGSM. We highlight five main observations. (1) Llama 2, Llemma, and MetaMath exhibit critical performance degradation across languages that are underrepresented in the training data of Llama 2. (2) Despite this, multilingual LMs underperform Llama 2, even in the context of underrepresented languages.[9] This disparity underscores the robust mathematical reasoning capabilities inherent in Llama 2 and absent in multilingual LMs. (3) LangBridge enhances the multilingual performance of Llemma and MetaMath, especially in underrepresented languages. Most notably, LangBridge is able to bring Llemma and MetaMath performance on Telugu (TE) from zero or near zero to a range comparable to other languages. (4) Mathematical reasoning capabilities of LangBridge models come from their original LMs, not their training data. This is evident from BLOOM-7.1B-PP2 and BLOOM-7.1B-MM underperforming Llemma-**LB**-9B and MetaMath-**LB**-9B, respectively, by a large margin.

[5] We define *finetuned* LMs as LMs trained on labeled corpora and *pretrained* LMs as LMs trained on unlabeled corpora.

[6] Here, the term *zero-shot* refers to the lack of few-shot examples.

[7] We use the language model checkpoint from Vu et al. (2022).

[8] The ten languages overlap with the 11 languages included in MGSM, except Telugu.

[9] Note BLOOM models were not trained in German, Russian, Japanese, and Thai.

	AVG	HRL	URL	EN	DE	FR	ES	RU	ZH	JA	TH	SW	BN	TE
Lang. Freq. (Llama 2, %)	-	-	-	89.7	0.17	0.16	0.13	0.13	0.13	0.10	Less Than 0.005			
Few-Shot Cross-lingual CoT														
Llama 2-7B	9.1	12.1	3.9	15.2	11.6	13.2	11.2	11.6	11.2	10.8	7.2	5.2	3.2	0.0
XGLM-7.5B	1.5	1.6	1.2	0.4	1.6	1.2	1.6	2.0	2.8	1.6	2.0	0.4	1.2	1.2
mT5-XXL (13B)	2.9	3.5	2.0	3.6	2.4	4.0	3.6	2.8	3.6	4.4	2.8	1.2	3.2	0.8
BLOOM-7.1B	2.4	2.6	2.0	3.6	1.2	3.6	2.4	2.0	3.2	2.0	0.0	2.4	2.8	2.8
BLOOM-7.1B-PP2	2.3	2.5	1.9	4.8	1.2	2.0	2.0	1.6	4.0	1.6	0.8	2.8	2.0	2.0
PaLM-540B	51.3	52.3	46.8	62.4	53.6	51.2	58.0	55.6	46.0	49.6	49.6	44.4	46.4	46.8
Llemma-7B	**21.6**	**29.9**	7.2	**44.8**	**27.2**	**33.2**	**29.2**	**26.0**	**26.4**	**22.4**	14.0	8.4	6.4	0.0
Llemma-**LB**-9B	20.4	22.5	**16.7**	34.8	23.6	26.8	22.4	18.8	16.0	15.2	**20.8**	**17.6**	**12.4**	**16.0**
Llemma-34B	35.6	46.3	16.7	58.0	48.0	46.8	48.0	47.2	36.8	39.6	28.4	27.2	11.2	0.0
Zero-shot CoT														
MathOctopus-7B	37.1	42.7	27.2	51.6	40.0	38.4	47.2	42.4	44.0	35.6	39.2	31.6	37.2	0.8
MathOctopus-13B	42.9	48.6	32.9	50.8	49.2	50.4	52.8	47.2	52.4	37.2	44.4	40.4	46.4	0.4
BLOOM-7.1B-MM	16.7	21.7	7.8	41.2	19.6	24.4	26.8	9.6	21.2	9.2	0.8	15.6	6.8	8.0
MetaMath-7B	34.4	51.8	3.9	**64.8**	**57.6**	**55.6**	56.4	50.4	42.4	35.6	4.0	6.4	4.4	0.8
MetaMath-**LB**-9B	**48.8**	**52.3**	**42.8**	63.2	50.8	52.4	**58.0**	**56.4**	**45.2**	**40.0**	**50.4**	**43.2**	**42.8**	**34.8**
MetaMath-13B	40.5	**60.3**	6.0	**70.4**	**64.4**	**65.2**	**63.6**	**60.0**	50.8	**47.6**	4.8	11.6	6.8	0.8
MetaMath-**LB**-15B	53.5	58.0	45.2	67.6	63.6	61.6	63.2	**60.0**	48.0	42.0	52.8	41.6	50.0	36.4
MetaMath-**LB**-20B	**55.8**	58.7	**50.7**	66.4	64.0	64.0	60.4	58.8	**52.4**	45.2	**53.6**	**49.2**	**52.8**	**47.2**

Table 1: Accuracy (%) on MGSM. Alongside average (AVG) accuracy, we also report average accuracy of high-resource languages (HRL) and underrepresented languages (URL) classified by Shi et al. (2023). We include the language distribution of Llama 2 for reference. For pretrained models (**Top**), we prompt with 8-shot cross-lingual chain-of-thought (CoT) reasoning exemplars, except for PaLM-540B, for which we reference the 6-shot cross-lingual CoT performance reported by Shi et al. (2023). For finetuned models (**Bottom**), we evaluate zero-shot. The PP2 and MM suffixes denote models trained on Proof-Pile-2 and MetaMath, respectively. We compare LangBridge models (**LB**) to their original checkpoints and highlight the best-performing numbers in **bold**.

(5) Surprisingly, despite only being trained on English reasoning data, our MetaMath-**LB** models are competitive against MathOctopus models, which were finetuned on translations of GSM8K on ten out of 11 languages supported by MGSM. The performance gap between the two models on Telugu (TE), an unseen language by MathOctopus, spotlights the benefit of *zero-shot* multilingual adaptation of LangBridge. Continual training with curated target language data cannot generalize to unseen languages, whereas LangBridge models, by nature, has the potential to generalize to the large scale of languages included in the pretraining of mT5.

Overall, LangBridge models demonstrate competitive performance against baselines. LangBridge models vastly outperform similar-sized multilingual models, establishing LangBridge as a viable approach for developing mathematical reasoning models for low-resource languages. We provide an example of CoT rationale generated by MetaMath-**LB** in Appendix H.

4.3 Code Completion

4.3.1 Experimental Setup

Evaluation Datasets Leveraging the competitive translation performance of GPT-4 (OpenAI, 2023), as demonstrated in the study by Jiao et al. (2023), we extend **HumanEval** (Chen et al., 2021), a set of hand-written programming problems, into five underrepresented languages: Swahili, Bengali, Punjabi, Telugu, and Urdu. We name the resulting dataset **HumanEval-MT**. We select the five languages among those with reported MMLU (Hendrycks et al., 2021a) performance in the technical report of GPT-4 while also being included in the pretraining corpora of BLOOM. This choice is made to acquire high-quality translations from GPT-4, and give BLOOM-based baselines an even chance. To guide GPT-4 to only translate the embedded natural language instruction of the docstring while not modifying the code segments, we prompt GPT-4 with two human-annotated examples. Subsequently, the generated translations are executed in a Python interpreter environment to assert the absence of syntax errors. We provide

	AVG	EN	SW	BN	PA	TE	UR
Llama 2-7B	4.6	9.8	3.7	3.0	3.0	3.0	4.9
BLOOM-7.1B	4.9	7.3	2.4	5.5	4.3	6.1	3.7
BLOOM-7.1B-StarCoder	4.5	6.7	3.0	4.3	4.3	3.7	4.9
Code Llama-7B	16.9	**36.0**	14.0	14.6	10.4	9.8	16.5
Code Llama-**LB**-9B	**19.4**	31.7	**17.1**	**15.2**	**18.3**	**15.2**	**18.9**
Code Llama-13B	19.6	40.2	15.2	17.7	12.2	12.8	19.5
Code Llama-**LB**-15B	**23.6**	**41.5**	20.1	**20.1**	19.5	**19.5**	**20.7**
Code Llama-**LB**-20B	23.5	36.0	**24.4**	**20.1**	**20.7**	18.9	**20.7**
Code Llama-34B	22.7	43.9	17.1	23.8	15.9	12.8	22.6

Table 2: Pass@1 scores on HumanEval and HumanEval-MT. The models were evaluated on zero-shot code completion using greedy decoding. We compare LANGBRIDGE (**LB**) models to their original checkpoints and highlight the best-performing numbers in **bold**.

translation quality estimation of HumanEval-MT in Appendix E.

The examples within the HumanEval dataset often feature self-explanatory function names, which raises the possibility that LMs could potentially complete the associated code segments successfully without fully comprehending the natural language comments embedded within them. Consequently, we evaluate the models on *anonymized* versions of each language set, wherein the target function names of the code segments are uniformly altered to "func". An example of an anyonymized prompt and evaluation result on the non-anonymized version is available in Appendix F.

Language Models **Code Llama** (Rozière et al., 2023) is a family of models initialized from Llama 2 model weights and pretrained on a code-heavy dataset. In our experiments, we use Code Llama-Python models, which were further pretrained on a Python-heavy dataset. Since the datasets used to pretrain the Code Llama models are not publicly available, we sample from the Python subset of StarCoder data (Li et al., 2023b) as the training dataset for LANGBRIDGE.

Baselines We use **Llama 2** and **BLOOM** models as the baselines. Unlike other open multilingual models, the pretraining corpora of BLOOM contain code data with a 10.8% ratio (Laurençon et al., 2022). Additionally, we report the performance of **BLOOM-StarCoder**, a BLOOM model continually pretrained on the sample of StarCoder data used to train LANGBRIDGE models.

4.3.2 Results

Table 2 presents the Pass@1 scores on HumanEval and HumanEval-MT. Code Llama-**LB** models show consistent improvements over Code Llama across all underrepresented languages. Moreover, LANGBRIDGE models could match their larger baselines on the average scores, with the 9B model slightly underperforming Code Llama-13B by 0.2%, while 15B and 20B models surpassing Code Llama-34B. BLOOM trained on StarCoder data did not demonstrate noticeable improvements, re-emphasizing that the strengths of LANGBRIDGE models predominantly stem from the capabilities of original LMs.

4.4 Logical Reasoning

4.4.1 Experimental Setup

Evaluation Datasets We assess logical reasoning capabilities with **Big-Bench Hard (BBH)** (Suzgun et al., 2023), a collection of challenging subtasks where the application of chain-of-thought (CoT) reasoning has the potential to enhance performance substantially. From the 23 subtasks of BBH, we restrict the evaluation tasks to 14 subtasks that maintain relevancy even in another language.[10] We translate the selected subtasks to Bengali using GPT-4 to construct **BBH-BN**. We prompt GPT-4 with three human-translated examples for each subtask. Two native Bengali speakers iteratively curated the examples based on the translation quality of GPT-4 as feedback.

Language Models **Orca 2** (Mitra et al., 2023) was finetuned on Llama 2 with a collection of datasets augmented with reasoning traces of GPT-4 as well as fully synthetic datasets created with GPT-4. Orca 2 effectively improved the reasoning abilities of smaller LMs on complex tasks demanding advanced reasoning in zero-shot settings.

[10] List of selected subtasks is available in Appendix F.

	EN	BN
BLOOM-7B-OpenOrca	35.8	31.2
Orca 2-7B	**53.9**	36.7
Orca 2-**LB**-9B	46.9	**41.8**
Orca 2-13B	**57.9**	41.7
Orca 2-**LB**-15B	55.2	**45.4**
Orca 2-**LB**-20B	53.1	**45.4**

Table 3: Accuracy (%) on BBH (English) and BBH-BN (Bengali). We report the average accuracy across selected 14 subtasks. We compare LANGBRIDGE (**LB**) models to their original checkpoints and highlight the best-performing numbers in **bold**.

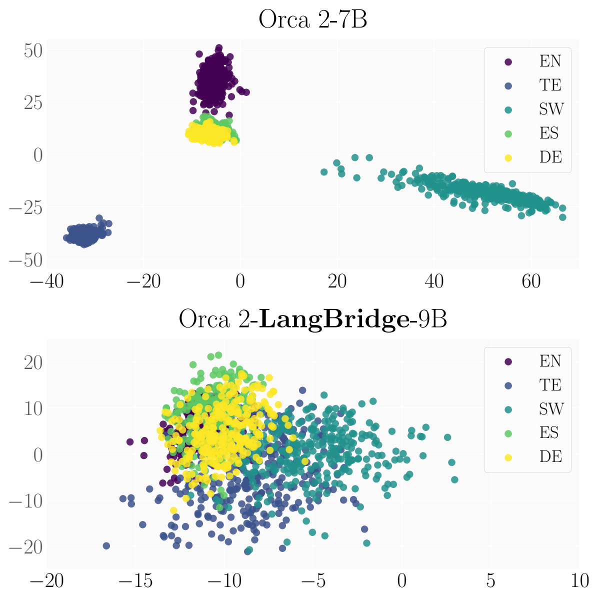

Figure 3: First two principal components of pooled representations obtained with FLORES. Note that the scales of the two subplots differ.

As the training dataset of Orca 2 is not publicly available, we sample the training data for LANGBRIDGE from OpenOrca dataset (Lian et al., 2023). OpenOrca follows the data distribution of the first iteration of Orca (Mukherjee et al., 2023). We employ CLD3[11] to filter any non-English data that mainly derives from translation datasets to ensure the *zero-shot* setting of our experiments. Examples were included if their input text had a 99% or greater probability of being English, while their target text also had a 95% or greater chance of being English. A slightly lower threshold was adopted for the target text to not falsely filter single-word responses, which CLD3 exhibits lower confidence.

Baselines In our evaluation of BBH, we assess whether Orca 2-**LB** models could acquire multilingual comprehension while retaining the zero-shot CoT capabilities of Orca 2. However, from our limited testing, we found that no existing open multilingual LMs could generate CoT reliably in a zero-shot setting. Consequently, they were not included as baselines. We do report the performance of **BLOOM-OpenOrca**, a BLOOM model further trained on the same training set as Orca 2-**LB**.

4.4.2 Results

Table 3 shows the average accuracy across the subtasks for BBH and BBH-BN. Notably, the Orca 2-**LB**-9B model shows considerable improvement in BBH-BN, surpassing the larger Orca 2-13B model. However, the Orca 2-**LB**-9B model has a discernible decline in performance in BBH English compared to Orca 2-7B. Nonetheless, the performance decrease in English is substantially mitigated in the larger-scale models. Another noteworthy observation is that the Orca 2-**LB**-20B model does not surpass the performance of the Orca 2-**LB**-15B model in either BBH or BBH-BN. In Appendix D, we analyze the impact of the encoder size on performance. Additionally, we present the evaluation results of individual subtasks in Appendix F and examples of generated CoT rationale in Appendix H.

5 Analysis of LANGBRIDGE

Based on the empirical evidence presented in the previous sections, we assert that LANGBRIDGE effectively enhances LMs' capability to address multilingual tasks without multilingual training, especially for low-resource languages. This section presents two fundamental observations that further reinforce the hypothesis outlined in Section 3.

The hypothesis primarily attributes the feasibility of LANGBRIDGE to the sufficiently language-agnostic representations of multilingual encoders. If the conjecture holds, given a LANGBRIDGE model, the LM's output representation of the soft prompt $\mathbf{H}_{enc}$ should also exhibit language-agnostic characteristics. It stands to reason that the LM would not arbitrarily introduce additional language-specific features to a language-neutral input.

Figure 3 shows the first two principal components of pooled representations obtained with 300 samples of FLORES (Goyal et al., 2022; Costa-jussà et al., 2022), a parallel corpora. For Orca 2, high-resource languages, English (EN), Span-

[11] github.com/google/cld3

ish (ES), and German (DE), are mapped closely together. Underrepresented languages, Telugu (TE) and Swahili (SW), exhibit a more distant mapping in the representation space, forming three clusters.[12] Conversely, Orca 2-LANGBRIDGE representations are mapped into a single cluster, indicating that the representations of $\mathbf{H}_{enc}$ maintain a relatively language-neutral status.

Figure 4 illustrates an example of "accidental translation" (Xue et al., 2021) by the Orca 2-LANGBRIDGE-15B model. Despite being given the options in Bengali, the LANGBRIDGE model perceives the options as Chinese and recites them in Chinese. With Bengali as input, accidental translation in a third language other than English or Bengali suggests that multiple languages may have similar representations in $\mathbf{H}_{enc}$ (Li and Murray, 2023). Nonetheless, accidental translations in languages other than English are uncommon for Orca 2-LANGBRIDGE models. We conduct a qualitative analysis on the CoTs generated by Orca 2-LANGBRIDGE-15B for BBH-BN SNARKS subtask, which we find is the combination with the most frequent accidental translations. Out of 178 CoT rationales generated, only eight examples contained accidental translations in Chinese, Danish, Hindi, Japanese, Marathi, and Turkish, with one or two examples per language. Additionally, seven examples had Bengali in their CoT rationales. The relatively high frequency of Bengali indicates that $\mathbf{H}_{enc}$ does not exhibit a perfectly language-agnostic behavior, and for some examples, the LM could discern the input language as Bengali. This is not ideal, as it suggests that the LM had to comprehend the input in Bengali, a language in which the LM lacks proficiency. We believe that LANGBRIDGE performance can be further enhanced by relieving the *zero-shot* constraint and adapting the mT5 encoder to have enhanced language-neutrality (Reimers and Gurevych, 2020; Feng et al., 2022) prior to alignment with the LM. However, we leave this exploration for future study.

6 Conclusion

In this paper, we propose LANGBRIDGE, a simple yet effective method of extending the capabilities of LMs to solve multilingual reasoning tasks without using multilingual supervision. We show that LANGBRIDGE is surprisingly effective for enhancing multilingual reasoning capabilities for low-resource languages. Additionally, we offer analysis explaining the efficacy of LANGBRIDGE and attribute it to the language-agnostic nature of multilingual representations. We hope our findings benefit the low-resource language users and spur further research advancing the development of LMs inclusive of the entire global community.

[12]Note that Flores Swahili is in Latin script, the same as the three high-resource languages.

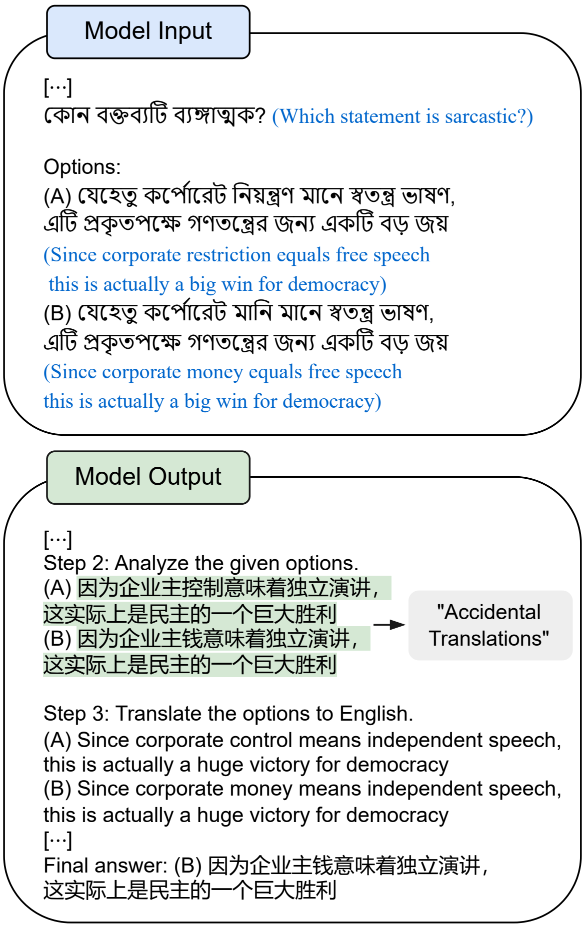

Figure 4: Example of accidental translation of an Orca 2-LANGBRIDGE model prompted with the SNARK subset of BBH-BN. Portions of the input prompt and several rational steps in the output are truncated for brevity. Translations are provided in (blue with parenthesis) wherever required.

Limitations

As LANGBRIDGE solely utilizes English data for training, LANGBRIDGE models may not be proficient in generating text in languages other than English. Although LANGBRIDGE successfully narrows the performance gap between high-resource and low-resource languages, a noticeable performance gap remains. Also, while multilingual representations are known to have language-agnostic characteristics to some degree, our analysis and previous works suggest that there is room for improvements (Libovický et al., 2020; Feng et al., 2022). While LANGBRIDGE has the potential to generalize to all languages supported by the multilingual encoder, the extent to which LANGBRIDGE enhances the reasoning capability of a specific language depends on two key factors: (1) The original proficiency of the LM in that particular language. (2) The proficiency of the encoder model in that particular language.

Ethical Considerations

While we share LANGBRIDGE models for open access, their terms for use or license adhere to those of the original LMs. The training datasets utilized in our research is primarily sourced from academic materials. As a result, we assess that the datasets likely contain a relatively fewer examples featuring offensive or personal information. Nevertheless, it is important to acknowledge that such content may be still present within the dataset. All annotators involved in this project received compensation higher than the minimum wage. Our work has been reviewed and received approval from the Institutional Review Board (IRB) at our institution.

References

Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, and Marine Carpuat. 2022. Quality estimation via backtranslation at the WMT 2022 quality estimation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 593–596, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.06786*.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Shikhar Vashishth, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. 2024. Llm augmented llms: Expanding capabilities through composition.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Loubna Ben Allal, Niklas Muennighoff, Logesh Kumar Umapathi, Ben Lipkin, and Leandro von Werra. 2022. A framework for the evaluation of code generation models. `https://github.com/bigcode-project/bigcode-evaluation-harness`.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed?

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023b. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

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

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2023. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5474–5490, Toronto, Canada. Association for Computational Linguistics.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 319–337, Toronto, Canada. Association for Computational Linguistics.

A Commonsense Reasoning

Table 9 shows the evaluation results of Orca 2 and Orca 2-LangBridge models on **COPA** (Roemmele et al., 2011) and **XCOPA** (Ponti et al., 2020), commonsense reasoning datasets. COPA is available in English, while XCOPA extends COPA to 11 languages. We do not include multilingual LMs as baselines, as COPA was included in the training set of Orca 2, making it challenging to draw meaningful comparisons.

Despite reaching near-perfect accuracy for COPA, Orca 2 models' performance drops closer to random chance (50%) on some of the underrepresented languages of XCOPA. LangBridge successfully decreases this performance degradation, except for Quechua (qu). The discrepancy is likely due to Quechua not being included in the 101 languages covered by mT5. This observation reemphasizes that the large-scale linguistic proficiency of LangBridge models is primarily derived from mT5.

B MSVAMP

MSVAMP (Chen et al., 2023) is a multilingual grade school level math word problem dataset translated from SVAMP (Patel et al., 2021) to 10 languages. We only evaluate MSVAMP in a zero-shot setting, as no CoT rationale examples are provided with the dataset.

Table 10 presents the evaluation results on MSVAMP. MetaMath and MathOctopus were not trained on SVAMP or MSVAMP, so MSVAMP can be seen as an out-of-domain test set to evaluate domain generalization (Chen et al., 2023). Performance of LangBridge-MetaMath models indicates our models can generalize to out-of-domain test sets successfully.

C General-domain Language Models

Table 11 probes the effect of LangBridge on general-domain English-centric LMs, **Llama 2** and **Mistral 7B** (Jiang et al., 2023), using MGSM. Consistent with the findings on specialized LMs, LangBridge enhances the performance of low-resource languages. As training corpora for Llama 2 and Mistral 7B are unavailable, we use a sample of the SlimPajama dataset (Soboleva et al., 2023) as the training set.

D Ablations

D.1 Freezing/Unfreezing

Target LM	**Trainable**		**AVG Score**
	Enc	LM	
MGSM			
Llama 2-7B			9.6
	✓		**11.3**
Llemma-7B			14.4
	✓		**20.4**
MetaMath-7B			**48.8**
	✓		43.9
HumanEval + HumanEval-MT			
Code Llama-7B			15.3
	✓		**19.4**
XCOPA			
Orca-7B			**76.6**
	✓		71.1
		✓	74.0
Orca-13B			**77.3**
	✓		65.1

Table 4: Ablations on the effect of freezing the encoder and the language model during aligning of LangBridge. ✓ denotes the module is trainable (not frozen) during aligning.

We strictly keep the embedding layers of mT5 trainable throughout our experiments as we extend the vocabulary and the embedding layer to incorporate whitespace characters. mT5 tokenizers do not have whitespace characters in their vocabularies, and their default behavior is to truncate any consecutive whitespaces to a single space. However, this could negatively affect understanding code or following instructions considering the frequent use of whitespaces as delimiters (\n, \t, and "four spaces"). Therefore, even when we freeze the encoder, we leave the embedding layer trainable for the added whitespace vocabulary.

Table 4 presents the ablation study on the impact of parameter freezing during the alignment process. We apply LangBridge with mT5-XL (2B) encoder on multiple LMs while varying the trainable modules. Notably, freezing the encoder appears beneficial when adapting finetuned LMs (MetaMath and Orca), whereas it negatively affects pretrained models (Llama, Llemma, and Code Llama). We speculate this divergence stems from differing entropy levels in the datasets: unlabeled corpora typically exhibit higher entropy than relatively well-formatted finetuning datasets. Consequently, we conjecture that for unlabeled data, keeping the encoder trainable enables the model to acclimate to the training data better. Nonetheless, we leave a thorough investigation for future research. Conversely, training the LM during the alignment phase does not improve performance. We hypothesize this is due to the training datasets being strictly in-domain of the LMs. As such, the models may not be learning additional information from the data.

D.2 Encoder Size

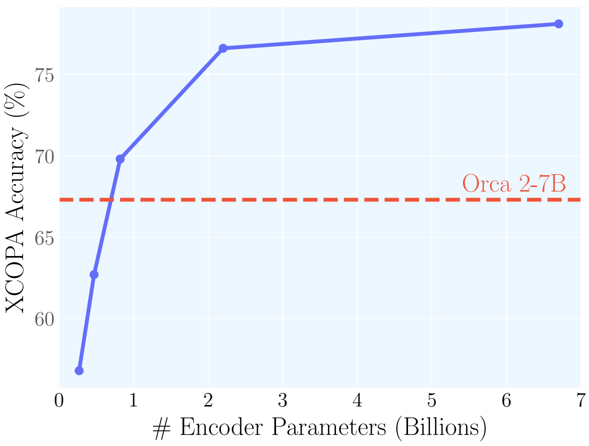

Figure 5: XCOPA accuracy (%) of Orca 2-7B models adapted with LangBridge using five different sizes of mT5 encoder. The dotted line shows the performance of base Orca 2-7B.

Figure 5 shows the effect of encoder size on LangBridge when applied to Orca 2-7B, measured with XCOPA. We test five different sizes of mT5 encoder: 270M (Small), 470M (Base), 820M (Large), 2.2B (XL) and 6.7B (XXL). We observe that LangBridge with the two smaller-sized encoders underperforms the base Orca 2-7B. Nonetheless, performance increases rapidly as the encoder size scales from 270M to 2.2B and saturates in the 2.2B to 6.7B range. These results, coupled with the main experiments of Section 4, highlight that scaling the encoder size past a certain point provides diminishing returns.

D.3 Training Set Size

While we fixed the training set size to 200,000 in our main experiments, Figure 6 shows that XCOPA performance peaks on 120,000 training examples, which is 60% of our experiment setting. This suggests that in practice, LangBridge can be applied more efficiently.

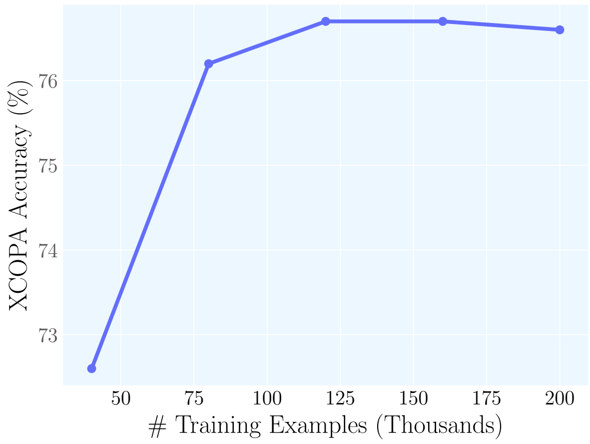

Figure 6: XCOPA accuracy (%) of Orca 2-7B models adapted with LangBridge using five different sizes of training datasets.

D.4 Different Architectures

	XCOPA (Acc. %)
Linear	76.6
MLP	72.7
Resampler	49.7

Table 5: Ablations of different LangBridge architectures using Orca 2-7B and mT5-XL encoder.

Table 5 shows the ablations of different LangBridge architectures. Throughout the main experiment, we adopt a single linear layer to connect the encoder and the language model. We ablate the effect of using an MLP layer following the second iteration of LLaVa (Liu et al., 2023a), and a "resampler" module that adopts the architecture of perceiver resampler employed by Flamingo (Alayrac et al., 2022). In contrast to the finding of Liu et al. (2023a), we find that using an MLP layer instead of a linear layer decreases the performance slightly. Using a resampler module results in random performance.

D.5 Different Encoder Models

	XCOPA (Acc. %)
umT5-XL	49.7
umT5-XL (F)	49.2
XGLM-1.7B	51.5
XGLM-1.7B (F)	49.7

Table 6: Ablations of different encoder models for LangBridge tested on Orca 2-7B model. (F) denotes that the encoder was frozen during alignment.

Table 6 presents the results of using different encoder models. We test out umT5 (Chung et al., 2023), which improves upon mT5 to include a higher proportion of low-resource languages during pretraining. As umT5 does not have "LM adapted" checkpoints, unlike mT5, we use the original checkpoints. However, we find using umT5 results in random performance. Since umT5 has a nearly identical architecture to mT5, except that it has relative position bias for every Transformer layer (Vaswani et al., 2017) in contrast to the very first, we speculate that using the encoder of a non-LM Seq2Seq model resulted in failed alignment. We also test XGLM (Lin et al., 2022), a decoder LM, as an encoder, in which we also observe similar results. In both cases, freezing the encoder made no difference. We leave further investigation for future research.

E Quality Estimation of HumanEval-MT

	EN	⟲SW	⟲BN	⟲PA	⟲TE	⟲UR
CL-7B	36.0	35.4	35.4	36.6	36.6	37.8
CL-34B	43.9	36.6	46.3	42.1	42.1	45.7
BLEU	-	88.3	90.5	93.4	90.5	96.1
crhF	-	89.9	93.1	94.6	93.1	96.5

Table 7: Quality estimation of HumanEval-MT using backtranslations. CL prefix denotes Code Llama Pass@1 score.

Table 7 presents the quality estimation for HumanEval-MT using backtranslation (Agrawal et al., 2022; Zhuo et al., 2023). We translate HumanEval-MT back to English and measure Code Llama Pass@1 scores. As explained in Section 4.3, we anonymize all function names. In addition, we report automatic evaluation metrics, BLEU (Papineni et al., 2002) and chrF (Popović, 2015), against the original HumanEval. We use the docstrings only for calculating the metrics.

Overall, Code Llama's performance on the backtranslations matches the original HumanEval benchmark, with the exception of Code Llama-34B's performance on Swahili backtranslation. As Code Llama-7B achieves comparable results on the Swahili backtranslation, this suggests that for Swahili, GPT-4 may struggle with accurately translating complex examples that only larger models can solve. Lower automatic evaluation scores in Swahili further support the idea.

```
def greatest_common_divisor(a: int, b: int) -> int:
    """ Rudi kipengele kikubwa zaidi cha pamoja cha
    integers mbili a na b
    >>> greatest_common_divisor(3, 5)
    1
    >>> greatest_common_divisor(25, 15)
    5
    """
```

```
def func(a: int, b: int) -> int:
    """ Rudi kipengele kikubwa zaidi cha pamoja cha
    integers mbili a na b
    >>> func(3, 5)
    1
    >>> func(25, 15)
    5
    """
```

Table 8: Comparison between original (**Top**) and anonymized (**Bottom**) prompts of HumanEval-MT Swahili.

F Additional Evaluation Results

F.1 BBH

Table 12 and 13 showcase the complete results for the 14 subtasks of BBH and BBH-BN. The selected subtasks are: CAUSAL JUDGEMENT, DATE UNDERSTANDING, DISAMBIGUATION QA, FORMAL FALLACIES, LOGICAL DEDUCTION (3, 5 AND 7), NAVIGATE, PENGUINS IN A TABLE, REASONING ABOUT COLORED OBJECTS, SNARKS, SPORTS UNDERSTANDING, TEMPORAL SEQUENCES, AND WEB OF LIES.

Orca 2-**LB** models show noticeable performance degradation in DATE UNDERSTANDING. From our qualitative analysis of the CoT, we observe that Orca 2-**LB** models frequently falsely assume an arbitrary date as the current date at the beginning of CoT (Figure 9), whereas the original Orca 2 models do not exhibit this behavior. Our exploration of the OpenOrca datasets reveals that examples often require the model to assume a specific current date. For example, an input text is given as "The current senate majority leader in the US is Chuck Schumer. Options: - yes - no", and the target text contains "...Today's date is October 12, 2021...". As we do not have access to Orca 2's original training dataset, a thorough ablation on the effect of such examples is challenging. Nevertheless, we speculate this problematic emergent behavior in Orca 2-**LB** models is partially due to the distribution shift of the training data from the original Orca 2 dataset to the OpenOrca dataset.

F.2 HumanEval

Table 8 compares the original and anonymized prompts of HumanEval-MT. **Top** can be solved without comprehending Swahili whereas **Bottom** is not. Table 14 presents the evaluation results on the non-anonymized version of HumanEval and HumanEval-MT. Compared to LANGBRIDGE models, Code Llama models show a sharper decline in performance when evaluated on the anonymized version, suggesting that they are less capable of comprehending natural text in another language.

G Experimental Details

We use a maximum input length of 1,024 and a maximum target length of 128 for training. However, as mentioned in Section 3, we vary the input length for every example within the 1,024 window to expose the LM to diverse lengths of soft prompts. Leveraging AdamW (Loshchilov and Hutter, 2019), we use a learning rate of $6e{-4}$ for the linear layer and $2e{-5}$ for the encoder. We do not use any learning rate scheduling. We use an effective batch size of 128. On a machine with 4 A100 80GB GPUs, the training takes under four hours for 9B models with the encoder frozen and under five hours with their encoder trainable. The maximum training time in our main experiments is ten hours for 20B models with encoder trainable. For further training BLOOM baselines, we keep all the hyperparameters the same as LANGBRIDGE models with two exceptions: (1) Learning rate, in which we used a uniform value of $2e{-5}$ across the entire model. (2) Sequence length, which was set to 1,152 $(1,024 + 128)$. For evaluations, we leverage LM Evaluation Harness (version 0.3.0) (Gao et al., 2023) and Bigcode Evaluation Harness (version 0.1.0) (Ben Allal et al., 2022) packages and report single run results with fixed seed.

For constructing HumanEval-MT and BBH-BN, we prompt GPT-4 with human-translated examples. The two examples used for HumanEval-MT were translated to Korean by a native Korean author. GPT-4 was then prompted to translate HumanEval to the target languages with the examples as reference. Note that the examples were provided to guide GPT-4 to keep the format of the data intact and only translate the natural language, not necessarily to enhance the translation quality itself. For BBH-BN, we hire two Bengali undergraduate students to translate three examples for each subtasks. The annotators were instructed to curate the examples and instructions until the GPT-4 translations were fluent and there were no unwanted artifacts in the output.

H CoT Examples

In this section, we provide three examples of CoT generated by LANGBRIDGE models. Figure 7 is from the MetaMath-LANGBRIDGE-15B model prompted with an example from MGSM Telugu. Figure 8 and 9 are from Orca 2-LANGBRIDGE-15B model prompted with examples from BBH-BN CAUSAL JUDGEMENT and BBH-BN DATE UNDERSTANDING. We select these two subtasks from BBH-BN as LANGBRIDGE substantially increased the performance of Orca 2-13B for the former (+13.4%) but caused a considerable decrease for the latter (−6.4%). We show a success case for the former and a failure case for the other. In Section F, we provide an explanation for the performance decrease in BBH-BN DATE UNDERSTANDING.

	COPA	AVG	HRL	URL	ZH	IT	VI	ID	ET	HT	QU	SW	TA	TH	TR
Lang. Freq. (Llama 2, %)	89.7	-	-	-	0.13	0.11	0.08	0.03	LESS THAN 0.005						
Orca 2-7B	**98.0**	67.3	**86.4**	63.0	**85.6**	**87.2**	**83.4**	82.4	54.4	52.2	49.8	54.2	58.0	62.0	71.0
Orca 2-**LB**-9B	90.0	**76.6**	83.4	**75.1**	85.4	81.4	79.8	**84.6**	**78.2**	**74.4**	**50.8**	**74.4**	**78.0**	**78.4**	**77.6**
Orca 2-13B	**99.0**	73.7	**93.1**	69.4	**92.4**	**93.8**	**87.0**	**86.8**	66.4	61.0	49.8	65.8	63.8	67.6	76.4
Orca 2-**LB**-15B	92.0	77.3	84.5	75.7	85.2	83.8	83.4	83.8	80.6	74.8	50.4	72.2	77.2	**79.8**	79.2
Orca 2-**LB**-20B	92.0	**79.8**	86.3	**78.4**	86.2	86.4	83.6	85.4	**82.8**	**76.4**	**54.2**	**77.8**	**82.8**	**79.8**	**82.4**

Table 9: Accuracy (%) on COPA and XCOPA. For XCOPA, we report the average accuracy across 11 languages. We compare LANGBRIDGE (**LB**) models to their original checkpoints and highlight the best-performing numbers in **bold**.

	AVG	HRL	URL	EN	DE	FR	ES	RU	ZH	JA	TH	SW	BN
Lang. Freq. (Llama 2, %)	-	-	-	89.7	0.17	0.16	0.13	0.13	0.13	0.10	LESS THAN 0.005		
MathOctopus-7B	39.2	41.5	34.0	39.8	42.4	44.0	43.3	41.6	40.4	38.7	35.1	29.7	37.2
MathOctopus-13B	45.1	47.2	40.0	44.8	46.7	50.6	49.9	47.6	47.1	43.9	40.0	36.3	43.6
BLOOM-7.1B-MM	24.9	29.5	14.1	40.7	28.8	33.8	32.0	27.7	27.7	15.7	3.0	21.2	18.0
MetaMath-7B	47.8	**61.1**	16.9	**66.3**	**63.5**	**64.1**	**64.9**	**60.3**	**55.0**	**53.4**	19.5	16.8	14.4
MetaMath-**LB**-9B	**52.0**	54.9	**45.1**	60.6	58.1	57.0	56.9	55.8	50.4	45.5	**46.3**	**42.1**	**46.8**
MetaMath-13B	50.6	**65.1**	16.8	**69.2**	**67.3**	**66.1**	**66.9**	**66.9**	**58.0**	**61.4**	18.6	14.4	17.3
MetaMath-**LB**-15B	57.0	60.4	49.1	64.1	61.9	65.9	64.8	61.0	55.0	50.1	51.3	42.1	54.0
MetaMath-**LB**-20B	**57.9**	60.4	**51.8**	65.3	63.0	62.5	62.7	60.9	55.4	53.3	**52.3**	**47.1**	**56.0**

Table 10: Accuracy (%) on MSVAMP. MM suffix denotes the model was finetuned on MetaMath. We compare LANGBRIDGE (**LB**) models to their original checkpoints and highlight the best-performing numbers in **bold**.

	AVG	HRL	URL	EN	DE	FR	ES	RU	ZH	JA	TH	SW	BN	TE
Lang. Freq. (Llama 2, %)	-	-	-	89.7	0.17	0.16	0.13	0.13	0.13	0.10	LESS THAN 0.005			
Llama 2-7B	9.1	12.1	3.9	15.2	11.6	**13.2**	11.2	**11.6**	**11.2**	**10.8**	7.2	5.2	3.2	0.0
Llama-**LB**-9B	**11.3**	**12.2**	**9.7**	**16.8**	**12.4**	12.8	**13.6**	9.2	10.0	**10.8**	**13.6**	**9.2**	**7.6**	**8.4**
Mistral-7B	**28.5**	**34.4**	12.8	**43.6**	**34.4**	**35.2**	**39.6**	**32.8**	**28.8**	**26.4**	20.4	7.6	16.0	7.2
Mistral-**LB**-9B	26.1	27.9	**20.7**	32.4	29.6	30.8	29.6	28.8	25.2	18.8	**24.4**	**18.8**	**22.4**	**17.2**

Table 11: Accuracy (%) on MGSM of general-domain LMs. We compare LANGBRIDGE models (**LB**) to their original checkpoints and highlight the best-performing numbers in **bold**.

	CAUSAL		DATE		DISAM.		FORMAL		LOGIC. 3		LOGIC. 5		LOGIC. 7	
	EN	BN	EN	BN	EN	BN	EN	BN	EN	BN	EN	BN	EN	BN
BLOOM-7B-OpenOrca	49.7	4.8	32.8	48.7	48.4	31.2	43.2	35.2	36.0	12.4	24.8	55.2	20.0	23.6
Orca 2-7B	**62.0**	47.1	**52.4**	**42.0**	62.8	42.4	**60.0**	50.4	**60.0**	37.2	**43.2**	25.6	**39.6**	20.8
Orca 2-**LB**-9B	57.2	**52.9**	26.8	24.4	**64.0**	**46.8**	55.2	**57.2**	52.4	**42.0**	36.0	**30.0**	38.8	**28.8**
Orca 2-13B	56.1	46.5	**64.0**	**50.0**	**66.8**	52.0	52.0	52.0	**68.4**	45.2	46.8	35.6	**49.2**	31.2
Orca 2-**LB**-15B	57.2	**59.9**	44.8	43.6	56.0	46.8	**55.6**	48.0	66.8	**56.4**	**47.2**	33.2	44.4	**34.4**
Orca 2-**LB**-20B	**64.7**	58.3	34.4	35.2	59.2	**56.0**	52.0	**54.8**	62.4	46.8	44.8	**38.0**	45.2	32.0

Table 12: Full results on BBH and BBH-BN (Part 1 of 2).

	NAVI.		PENGUINS		REASON.		SNARKS		SPORTS		TEMPO.		WEB	
	EN	BN	EN	BN	EN	BN	EN	BN	EN	BN	EN	BN	EN	BN
BLOOM-7B-OpenOrca	41.6	34.4	27.4	44.0	25.2	22.6	41.8	20.8	49.6	41.2	6.8	50.4	53.2	12.0
Orca 2-7B	**58.8**	46.4	**57.5**	24.0	**47.2**	25.6	**67.8**	42.9	**68.0**	51.6	**20.4**	13.6	**54.4**	43.6
Orca 2-**LB**-9B	48.8	**50.4**	44.5	**39.7**	41.6	**33.6**	56.5	**53.7**	66.4	**53.2**	17.6	**17.2**	50.8	**54.8**
Orca 2-13B	53.2	49.2	**59.6**	30.1	**61.6**	26.4	**65.5**	48.0	76.4	49.2	39.6	**22.8**	52.0	45.2
Orca 2-**LB**-15B	58.4	**62.0**	56.8	**43.2**	60.0	32.8	60.5	50.8	73.6	52.4	34.0	20.0	**57.6**	52.0
Orca 2-**LB**-20B	**60.0**	51.2	50.7	39.0	59.6	**37.2**	62.1	**54.8**	72.0	**53.6**	24.0	19.2	52.8	**59.2**

Table 13: Full results on BBH and BBH-BN (Part 2 of 2).

	AVG		**EN**		**SW**		**BN**		**PA**		**TE**		**UR**	
		An.		*An.*		*An.*		*An.*		*An.*		*An.*		*An.*
Llama2-7B	10.2	4.6	11.0	9.8	11.6	3.7	11.0	3.0	7.9	3.0	8.5	3.0	11.0	4.9
BLOOM-7.1B	6.7	4.9	8.5	7.3	6.1	2.4	6.1	5.5	6.7	4.3	6.7	6.1	6.1	3.7
BLOOM-7.1B-SC	8.4	4.5	11.0	6.7	9.8	3.0	7.9	4.3	6.7	4.3	7.9	3.7	7.3	4.9
Code Llama-7B	23.0	13.0	**36.0**	**36.0**	21.3	14.0	21.3	14.6	17.7	10.4	16.5	9.8	**25.0**	16.5
LB-Code Llama-9B	**24.9**	**19.4**	34.2	31.7	**27.4**	**17.1**	**23.2**	**15.2**	**23.2**	**18.3**	**19.5**	**15.2**	22.0	**18.9**
Code Llama-13B	26.0	19.6	**42.7**	40.2	24.4	15.2	**26.2**	17.7	17.7	12.2	17.7	12.8	**27.4**	19.5
LB-Code Llama-15B	**26.3**	**23.6**	36.6	**41.5**	22.6	20.1	23.8	**20.1**	**26.8**	19.5	**23.8**	**19.5**	24.4	**20.7**
LB-Code Llama-20B	26.2	23.5	35.4	36.0	**25.6**	**24.4**	22.6	**20.1**	25.6	**20.7**	20.7	18.9	27.4	20.7
Code Llama-34B	29.7	22.7	46.3	43.9	28.7	17.1	32.2	23.8	22.0	15.9	19.5	12.8	29.3	22.6

Table 14: HumanEval-MT Pass@1 performance. *An.* denotes the anonymized version.

Model Input

ప్రశ్న: జీనీ పాత సావనీర్ టి-షర్టులు ఉపయోగించి ఒక బొంత కుడుతున్నాడు. అతడికి ప్రతి విహారయాత్రకు ఒక చొక్కా ఉంది. ప్రతి షర్టుకు బొంతపై ఒక ప్రత్యేకమైన బ్లాక్ ఉంది. ప్రతి వరస కూడా విభిన్న సంవత్సరాల్లో వెళ్లిన విహారయాత్రల బ్లాక్‌లతో తయారు చేశాడు. అతడు సంవత్సరానికి నాలుగుసార్లు విహారయాత్రకు వెళతాడు, అతడు 23 సంవత్సరాల వయస్సు నుంచి విహారయాత్రకు వెళుతున్నాడు, అతడి వయస్సు ఇప్పుడు 34. అతడికి మొత్తం ఎన్ని క్విల్ట్ బ్లాక్‌లు ఉన్నాయి?

(Question: Christina is planning a birthday party and needs .75 gift bags per invited guest, because 1/4 of attendees don't show up. She invited 16 friends. Gift bags are $2 each. How much will she spend?)

Model Output

```
Kristin invited 16 friends, so she needs 16 x 0.75 = 12 gift bags.
Each gift bag costs $2, so she will spend 12 x $2 = $24 on gift bags.
#### 24
The answer is: 24
```

Figure 7: Example of zero-shot CoT generated by MetaMath-LANGBRIDGE-15B when prompted with an example from MGSM Telugu. For brevity, any instructional prompts are truncated. Translation of the input is provided in (blue with parenthesis).

Model Input

কিভাবে একজন সাধারণ ব্যক্তি কার্যকারণ সম্পর্কে নিম্নলিখিত প্রতিটি প্রশ্নের উত্তর দেবেন?
জো খুব অতিশীত অনুভব করছিল, সে স্থানীয় স্মুথি দোকানে স্থগিত হয়ে সর্বাধিক আকারের পানীয় কিনতে গেল। অর্ডার দেওয়ার আগে, ক্যাশিয়ার তাকে বললেন যে যদি সে একটি মেগা-সাইজড স্মুথি কিনে তাহলে সে এটি একটি বিশেষ স্মারকীয় কাপে পাবে। জো জবাব দিলেন, 'আমি স্মারকীয় কাপ এর ব্যাপারটা নিয়ে যতো চিন্তা করছি না, আমি শুধু আপনাদের সাথে থাকা সবচেয়ে বড় স্মুথি চাই।' আশনুরূপভাবে, জো স্মারক কাপে মেগা-সাইজড স্মুথি পেল। জো কি ইচ্ছাকৃতভাবে স্মারক কাপ নিল?
(How would a typical person answer each of the following questions about causation?
Joe was feeling quite dehydrated, so he stopped by the local smoothie shop to buy the largest sized drink available. Before ordering, the cashier told him that if he bought a Mega-Sized Smoothie he would get it in a special commemorative cup. Joe replied, 'I don't care about a commemorative cup, I just want the biggest smoothie you have.' Sure enough, Joe received the Mega-Sized Smoothie in a commemorative cup. Did Joe intentionally obtain the commemorative cup?)

```
Options:
(A) হ্যাঁ (Yes)
(B) না (No)
```

Model Output

```
Step 1: Understand the context of the question.
The question is asking whether Jo intentionally took the special commemorative
cup along with the mega-sized smoothie.

Step 2: Analyze the information provided.
Jo was overly excited and wanted the largest possible beverage. The cashier
offered a special commemorative cup for purchasing the mega-sized smoothie. Jo
did not express any concern about the cup and only wanted the largest smoothie.

Step 3: Determine the answer based on the information.
Since Jo did not express any concern about the cup and only wanted the largest
smoothie, it can be inferred that Jo did not intentionally take the special
commemorative cup.

Step 4: Choose the appropriate answer from the options.
The appropriate answer is (B) no.
```

Figure 8: Example of zero-shot CoT generated by Orca 2-LANGBRIDGE-15B when prompted with an example from BBH-BN CAUSAL JUDGEMENT. For brevity, any instructional prompts are truncated. Translation of the input is provided in (blue with parenthesis).

Model Input

আজ 9/7। জেন NFL 2003 দেখছে। MM/DD/YYYY-তে আজ থেকে এক সপ্তাহ আগের তারিখ কী?
(Today is 9/7. Jane is watching NFL 2003. What is the date one week ago from today in MM/DD/YYYY?)
Options:
(A) 09/05/2003 (B) 08/30/2003 (C) 08/31/2074 (D) 08/31/2003 (E) 06/30/2004

Model Output

```
Step 1: Identify the current date.
The current date is 7/9/2021.

Step 2: Determine the date one week before the current date.
To find the date one week before the current date, we need to subtract 7 days
from the current date.

Step 3: Calculate the date one week before the current date.\n7/9/2021 - 7 days
= 7/2/2021
[...]
Step 11: Match the calculated date with the given options.
The calculated date matches option (A) 05/18/2021.

Final answer: (A) 05/18/2021
```

Figure 9: Example of zero-shot CoT generated by Orca 2-LANGBRIDGE-15B when prompted with an example from BBH-BN DATE UNDERSTANDING. For brevity, any instructional prompts are truncated. Translation of the input is provided in (blue with parenthesis).