Optimizing Language Augmentation for Multilingual Large Language Models: A Case Study on Korean

ChangSu Choi[1‡], **Yongbin Jeong**[3‡], **Seoyoon Park**[2‡],
InHo Won[1], **HyeonSeok Lim**[1], **SangMin Kim**[1], **Yejee Kang**[3],
Chanhyuk Yoon[3], **Jaewan Park**[3], **Yiseul Lee**[3], **HyeJin Lee**[4],
Younggyun Hahm[3], **Hansaem Kim**[2] **and KyungTae Lim**[1†]
[1]SeoulTech, [2]Yonsei University, [3]Teddysum, [4]KISTI
choics2623@seoultech.ac.kr, ybjeong@teddysum.ai, seoyoon.park@yonsei.ac.kr
{wih1226, gustjrantk, sangmin6600, ktlim}@seoultech.ac.kr
khss@yonsei.ac.kr, {kangyj, chyoon, jwpark, yslee, hahmyg}@teddysum.ai, hyejin@kisti.re.kr

Abstract

Large language models (LLMs) use pretraining to predict the subsequent word; however, their expansion requires significant computing resources. Numerous big tech companies and research institutes have developed multilingual LLMs (MLLMs) to meet current demands, overlooking less-resourced languages (LRLs). This study proposed three strategies to enhance the performance of LRLs based on the publicly available MLLMs. First, the MLLM vocabularies of LRLs were expanded to enhance expressiveness. Second, bilingual data were used for pretraining to align the high- and less-resourced languages. Third, a high-quality small-scale instruction dataset was constructed and instruction-tuning was performed to augment the LRL. The experiments employed the Llama2 model and Korean was used as the LRL, which was quantitatively evaluated against other developed LLMs across eight tasks. Furthermore, a qualitative assessment was performed based on human evaluation and GPT4. Experimental results showed that our proposed `Bllossom` model exhibited superior performance in qualitative analyses compared to previously proposed Korean monolingual models.

Keywords: Large Language Model, Less-resourced Languages, Instruction-tuning

1. Introduction

A large language model (LLM) comprehends linguistic information and knowledge via pretraining to predict the subsequent word based on the given context (Zhao et al., 2023). However, the growth of LLMs increases the computing resources required for training, posing a challenge for smaller research groups to utilize them realistically (Hoffmann et al., 2022). To meet the demands of this era, numerous big tech companies and research institutes have been competing to launch multilingual LLMs (MLLMs) (Touvron et al., 2023a,b; Workshop et al., 2023). However, less-resourced languages (LRLs) are being overlooked (Gu et al., 2018). The recently launched Llama2 (Touvron et al., 2023b) is an MLLM trained in more than 28 languages; however, only 0.06% of the data was used for the Korean language. This leads to two significant syntactic and semantic challenges. First, during the MLLM training, LRLs use minimal vocabulary based on the scarce training data, which limits their expression owing to the inadequate lexicon. Second, greater semantic knowledge is required to employ LLMs for specific tasks, such as question-answering, thereby rendering the models inapplicable or prone to hallucinations (Zheng et al., 2023; Peng et al., 2023a).

Numerous methods have been proposed to enhance the LRL performance. These include expanding the vocabulary of word embeddings (Wang et al., 2019; Schuster et al., 2019), aligning multilingual embeddings by combining them with other languages (Artetxe et al., 2017, 2018), and reinforcing the utility of LLMs using minimal training data. The **L**ess **I**s **M**ore for **A**lignment (LIMA) study proposed a method to maximize the utility of LLMs using 1,030 high-quality instruction data(Zhou et al., 2023).

Based on the existing research, it remains to be investigated whether MLLMs can expand their vocabulary and enhance the semantic inferencing capability of a specific language using minimum additional data.

This study explores the aforementioned aspects by proposing a method to enhance the Korean language capabilities of a representative MLLM, i.e., Llama2. The specific language abilities of the MLLM are enhanced using the following strategies: (1) **Vocabulary expansion**: To enhance its fundamental vocabulary, the MLLM was augmented with the dictionary of a specific language. (2) **Knowledge enrichment**: The vocabulary and knowledge information of this language were enhanced in the MLLM via pretraining. (3) **Usability enhancement**: High-quality instruction data were generated in the Korean language to improve its

[‡]These authors contributed equally.
[†]Corresponding author.

LLM applicability. Korean is not an LRL as it comprises various language resources and evaluation data (Park et al., 2021). However, Korean is experimentally suitable as a relatively less-resourced language because the Llama2 model uses limited Korean data and vocabulary during training.

To enhance the vocabulary, knowledge reinforcement, and usability, 7,478 Korean vocabulary entries were added and pretraining was performed using a Korean–English corpus. The 1,030 English data, proposed by LIMA (Zhou et al., 2023), were restructured by three Korean linguistics to ensure their practical similarity with the Korean language, thus enhancing their usability.

The effectiveness of the three enhancement methods were validated by addressing the following questions: (1) What are the advantages of expanding the Korean vocabulary? (2) Is it effective to connect the knowledge of high- and low-resource languages via pretraining? (3) Do the actual usability and accuracy improve using the proposed Korean LIMA data? We propose the `Bllossom` model that applies the three aforementioned methods. Quantitatively, this model demonstrated an average performance improvement from 1.8% to 8% across eight tasks compared to the model without vocabulary expansion. For the qualitative evaluation of each model, the answers to 300 queries were compared using human and preference evaluations based on GPT4. Consequently, the `Bllossom` model outperformed the other Korean models of the same size by 93%. The contributions of this study are as follows:

- A method for enhancing the vocabulary, knowledge, and usability of LRLs using MLLMs was proposed.
- A method for constructing instruction data based on language-specific features was presented and demonstrated by constructing a Korean LIMA dataset.
- For easy utilization, the data, models, and services used to construct and evaluate the Korean LLM were made publicly accessible[1].

2. Related Work

2.1. MLLMs

LLMs are massive pretrained language models containing more than several hundred billions of parameters (Zhao et al., 2023). The generative LLMs based on the decoders of various transformers (Vaswani et al., 2017), including GPT series models (Radford et al., a,b; Brown et al., 2020), are pretrained using the causal language modeling (CLM) approach. This approach predicts the subsequent token based on the preceding token sequence, providing insights into language, grammar, and knowledge. However, the user intent is difficult to determine because CLM can only predict the subsequent token. Hence, a pretrained model should be tuned to accurately comprehend the user intent and achieve capabilities, such as instruction adherence (Ouyang et al., 2022). This tuning process is referred to as “instruction tuning”, which is commonly implemented using supervised fine tuning (SFT) that considers the two directives and input data as model inputs and predicts the output data (Wei et al., 2022a). SFT maximizes the training efficiency using a small amount of high-quality data rather than a large amount of low-quality data. LIMA (Zhou et al., 2023) uses a smaller amount of highly refined SFT training data of 1,030 to outperform models trained on large amounts of auto-generated or low-quality SFT data. Additionally, LIMA proposes qualitative performance evaluation methods including, human and GPT4 evaluation for LLMs to determine the superior model.

Multilingual LLMs are advantageous in accumulating vast training data from multiple languages. Owing to the increasing parameters and LLM training data, models such as FLAN-T5 (Wei et al., 2022a), BLOOM (Workshop et al., 2023), Falcon, Llama (Touvron et al., 2023a), and Llama2 (Touvron et al., 2023b), which are smaller but comprehensively acquire the multilingual knowledge, have attracted scholarly attention. Llama2 is a multilingual language model trained using large-scale publicly available data (including CommonCrawl, Github, Wikipedia, and ArXiv) for more than 28 languages. Thus, it possesses cross-language understanding capabilities. However, languages with non-Latin character systems, such as Korean and Chinese, exhibited inferior performances (Cui et al., 2023).

2.2. Open-source Korean LLMs

EleutherAI developed Polyglot-Ko, which is a monolingual Korean LLM pretrained on 1.2 TB of Korean data and contains models with sizes up to 12.8B (Ko et al., 2023). KoAlpaca[2] is a model based on Polyglot-Ko that automatically translates the SFT data of the English Alpaca (Taori et al., 2023) into Korean and performs SFT with a total of 21K data. Similarly, Kullm (Lab and research, 2023) was proposed based on Polyglot-Ko and tuned using additional instruction data. Kullm used 153K SFT training data by translating English SFT datasets, including the GPT-4-LLM (Peng et al., 2023b), Vicuna (Chiang et al., 2023), and Dolly from Databricks (Conover et al., 2023). Ad-

[1] github.com/Anonymous-introduce-paper/LR

[2] https://github.com/Beomi/KoAlpaca

Sentence: 햄버거를 먹는 공룡 (A dinosaur eating a hamburger)	
Model	**Tokenization results**
Llama2	'_', '<0xED>', '<0x96>', '<0x84>', '<0xEB>', '<0xB2>', '<0x84>', '<0xEA>', '<0xB1>', '<0xB0>', '를', '_', '<0xEB>', '<0xA8>', '<0xB9>', '는', '_', '공', '<0xEB>', '<0xA3>', '<0xA1>'
Proposed	'햄', '버', '거', '를', '_먹는', '_', '공', '룡'

Table 1: Comparison of tokenization results between Llama2 and the proposed model

ditionally, models that perform the Korean SFT based on multilingual models, such as Llama2, have been launched. Komt is an instruction-tuned model based on Llama2 using a total of 1,543K data processed from existing Korean natural language processing data. Ko-Platypus2 (Lee et al., 2023a) enhances the logic knowledge of LLMs using a translated dataset from English Open-Platypus into Korean. This model is tuned using Llama2 with 25K SFT data. The aforementioned Korean models were used in the experiment, and the access links and summary information for each model are listed in Table 4.

3. Enriching the MLLM vocabulary

This section introduces the following two approaches to the three language enhancement methods proposed in the Introduction: (1) vocabulary expansion and (2) knowledge enrichment. We propose a method to expand the Korean vocabulary in Llama2, which is a representative multilingual LLM, and reinforce the knowledge information between the Korean and English languages via CLM-based pretraining.

3.1. Vocabulary expansion

The training data of Llama2 consisted of 89.7% English words, and the tokenizer dictionary ($\mathcal{D}_L$) was composed of 90% English (or Latin) words. The majority of the remaining words were rare words, neologisms, and LRLs categorized as out-of-vocabulary (OOV). To address this issue, Llama2 employed the SentencePiece tokenizer (Kudo and Richardson, 2018) that uses a UTF-8 byte fallback mechanism to handle the OOV words by decomposing them to the UTF-8 byte level. Therefore, the words not in $\mathcal{D}_L$ are represented without expanding the tokenizer vocabulary. The Korean vocabulary was expanded despite having a method to represent the language. Table 1 compares the tokenizing results of Llama2 and the proposed model with an expanded Korean vocabulary for the sentence “햄버거를 먹는 공룡”. In the original Llama2, the Korean word “햄” was decomposed into the tokens “<0xED>”, “<0x96>”, and “<0x84>”, and “버” was decomposed into “<0xEB>”, “<0xB2>”, and “<0×84>” at the byte level. Contrastingly, the tokenizer with an expanded vocabulary tokenized “햄” and “버” in their original forms. The tokenizing results of the existing Llama2 model can lead to the following two problems, as indicated in the Chinese ALPACA (Cui et al., 2023):

1. **Increased token length**: The model cannot represent an OOV using a single token that requires three or four byte tokens. This reduces the possible input length of the model and increases the encoding and decoding times.

2. **Duplication of byte tokens**: “햄” and “버” are unrelated tokens; however, they are represented using the same byte token “<0x84>”. Therefore, the model may experience confusion while learning two semantically unrelated words with partially identical representations.

These limitations were overcome by introducing the Korean vocabulary, as shown in Equation 1. A new embedding was generated by combining the existing Llama2 vocabulary $\mathcal{D}_L$ and KoBERT [3] vocabulary $\mathcal{D}_K$.
The KoBERT vocabulary, designed by considering Korean morphemes, consists of $|\mathcal{D}_K| = 8,002$ words, whereas $|\mathcal{D}_L| = 32,000$. The union of the two vocabularies has a size of $|\mathcal{D}| = |\mathcal{D}_L \cup \mathcal{D}_K| = 39,478$. Therefore, the size of the newly added dictionary was $|\mathcal{D}_R| = 7,478$.

$$\mathcal{D} = [\mathcal{D}_L; \mathcal{D}_R] \tag{1}$$

Here, $\mathcal{D}_L$ used the word embeddings trained on the original Llama2, and the newly added word embeddings $\mathcal{D}_R$ were randomly initialized.

3.2. Enriching the knowledge information by MLLM pretraining

This section introduces methods to reinforce the word and knowledge information of MLLM via CLM-based pretraining. For queries in the Korean language, the publicly released Llama2 13b model responds in English or alternates between English and Korean (code-switching), indicating a limited Korean expression. However, the content is often accurate when Llama2 responds to a Korean query in English. When asked “이탈리아 수도에 대해 한국어로 소개해줘” (“Introduce me to the Italian capital in Korean”), the model replies, “로마 is the capital city of Italy and...” where the proper nouns “로마” (Rome) and “콜로세움” (Colosseum) are generated in Korean but

[3] https://github.com/SKTBrain/KoBERT

Category	LIMA dataset (huggingface.co/datasets/GAIR/lima)
NE change	(EN) I heard north ridge of **mount Stuart** from my friends, can you tell me more?
	(KO) I heard north ridge of **'Bukhansan Mountain'** from my friends, can you tell me more?
NE change	(EN) How to claim tax back (in USA)?
	(KO) How to claim tax back **in Korea**?
topic change	(EN) What are the primary objections **Democrats** have to a **border wall**?
	(KO) What is the **Korean Democratic Party's** opinion on **voting rights** for overseas Koreans?
topic change	(EN) How to make **creepy** food??
	(KO) How to make **bizarre** food??

Table 2: Instances of modifications in the English LIMA dataset to reflect the Korean cultural context

Language	Source	Size(GB)	Content
Korean	Public	22.41	news, web
	WIKI-ko	0.76	wikipedia
English	WIKI-en	9.92	wikipedia
Total		33.09	

Table 3: The composition of the pretraining data. The Public data is in (www.aihub.or.kr)

the detailed explanations are provided in English. This is because the knowledge acquired through pretraining was predominantly in English.

This limitation can be overcome by aligning the knowledge of the Korean and English languages in the MLLM by further pretraining it on a small amount of data. The MLLM expanded with Korean vocabulary was trained on the English and Korean Wikipedia, thereby bridging the extensive English knowledge (accounting for 89.7%) and limited Korean knowledge (0.06%). This method aligns with that of the multilingual BERT approach, which was pretrained on the Wikipedia data from 104 languages (Pires et al., 2019).

Equation 2 shows that the proposed model was pretrained using CLM. Given an input token sequence $x_{<i} = (x_0, x_1, \dots, x_{i-1})$ the model predicts the next token x_i, computes the loss by taking the negative log-likelihood of the predicted token probability, and minimizes this loss.

$$L_{CLM}(\theta)=\mathbb{E}_{x\sim\mathscr{D}_{PT}}\left\{-\sum_{i} logP(x_i|x_{<i};\theta,\mathcal{D})\right\} \quad (2)$$

Here, $L_{CLM}(\theta)$ represents the loss function of the language model over the pretraining dataset $\mathscr{D}_{PT}$, θ signifies the model parameters, x_i is the target token for prediction, and $\mathcal{D}$ refers to the dictionary expanded using the Korean vocabulary. Table 3 lists the specific compositions and sources of $\mathscr{D}_{PT}$. The loss function accounts for the prediction accuracy of each token within the pretraining dataset. The Korean and English bilingual corpora were adopted in the training method and the model was trained on 33 GB of pretraining data for one epoch with a batch size of 8.

4. Instruction Tuning on LIMA

The Korean language capability was enhanced by pretraining and the existing knowledge between the English and Korean languages was bridged. However, models trained during pretraining have limited applicability because they are specialized for predicting only the subsequent token. Consequently, high-quality Korean SFT data are required to accurately understand the user intent and generate desired responses. This section describes the method for reconstructing Korean SFT data based on English LIMA and introduces an instruction-tuning approach using these data.

4.1. Building the Korean LIMA

The Korean LIMA dataset for SFT was constructed based on a version that underwent machine translation using the English LIMA dataset. Consequently, post-processing was required to address the following issues: (1) discrepancies within the authentic Korean linguistic styles owing to machine translation and (2) exclusion of the Korean cultural context stemming from the characteristics of raw sources in the English LIMA dataset. The Korean LIMA dataset used in this study underwent a human review of the initial machine-translated text and modifications to reflect the Korean cultural context, which involved replacing the named entities and changing the main topic. For the human review process, we recruited the reviewers with Korean as their native language, ensuring that they calibrated all the translated data to the most natural Korean linguistic style.

The raw sources for the English LIMA dataset were posts from the English-speaking community forums, such as Stack Exchange and WikiHow, which reflected the cultural context of English speakers. The cultural context refers to a broad spectrum encompassing everything from daily consciousness to political, economic, and social systems. For instance, a sample from the English LIMA dataset, “How to make banana muffins?” may be irrelevant to the Korean culture because “banana muffins” are neither a popular consumable nor a frequently baked item in Korea. To reflect the Korean cultural context in the dataset, we modified the instances in the English

LIMA data that featured Western cultural contexts, particularly the American contexts. These modifications ranged from narrow changes, such as renaming the entities, to broader adjustments, such as entirely altering the dataset topic to fit the Korean context (see examples in Table 2).

4.2. MLLM Training using the Korean LIMA

The Stanford Alpaca (Taori et al., 2023) is an instruction tuned model based on the Llama trained on 52k instruction data. The corresponding training code is open-source[4]. We adapted the training script to instruct our model using the Korean LIMA dataset. Instruction-tuning follows the SFT method, wherein prompts are provided as inputs to the model which is subsequently trained to produce the user-desired responses. While this process is similar to pretraining, it differs in that only the output of the prompt is used to compute the loss. This can be mathematically represented as follows:

$$L_{SFT}(\theta)=\mathbb{E}_{x\sim\mathscr{D}_{SFT}}\left\{-\sum_{i\in out} logP(x_i|x_{<i};\theta,\mathcal{D})\right\} \quad (3)$$

Where θ represents the model's parameters, $\mathscr{D}_{SFT}$ denotes the SFT dataset, and $x = (x_0, x_1, x_2...)$ signifies the token sequence of the template containing the instruction and output.

Pretraining and instruction-tuning require substantial GPU resources. Recent research proposals have focused on training only specific portions of the model that require minimum GPU resources. LoRA (Hu et al., 2022) is a representative method that involves freezing a pretrained model and infusing each of its layers with trainable rank-decomposition matrices for further training. To apply LoRA, one must choose which parts of the entire model to train. This study trained only the linear layers of the transformer attention, including the query, key, and value, along with the expanded word embedding (as shown in Figure 1). Consequently, 5.977% of the total Llama2 parameters were used for training, thereby facilitating the training of our model on a single A6000 GPU.

Figure 1 shows the three proposed enhancement methods. The final model underwent the following sequence: (1) vocabulary expansion, (2) bilingual pretraining, and (3) instruction tuning (SFT). Within this context, “Trainable” (red) and “Frozen” (blue) refer to regions where the parameters were updated and not updated during training, respectively. SFT was concurrently performed using the constructed Korean and English LIMA datasets.

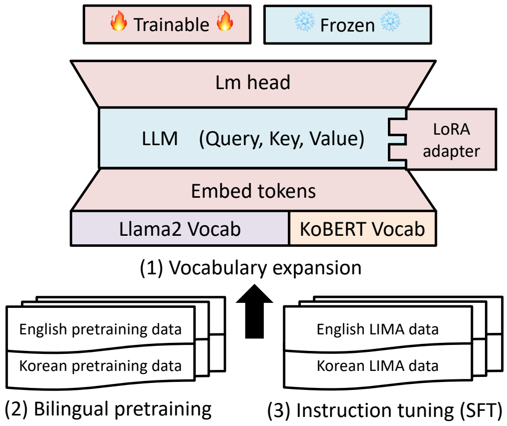

Figure 1: Architecture of the proposed model. During training, the LoRA adapter takes Query, Key, and Value from LLM and trains it with new parameters.

5. Quantitative evaluation

This section describes the following experiments to explore the three research objectives presented in the introduction: (1) comparison between the models with and without the expanded Korean vocabulary; (2) comparison between the model pretrained on the Korean–English bilingual training data and that trained only on the Korean data; and (3) variations in performance owing to instruction-tuning using the LIMA dataset.

5.1. Evaluation Environment

To ensure a fair comparison of LLMs, it is essential to define the task selection for evaluation and specify the LLM model used in the evaluation. To quantitatively evaluate the problem-solving capability of LLMs from various perspectives, the tasks involve language comprehension and inference, sentiment analysis, etcetera (Park et al., 2021; Zhou et al., 2023). The Korean LLMs were comprehensively evaluated based on eight datasets. The benchmarks for the evaluation tasks were KLUE’s NLI, STS, and YNAT, and Naver AI’s Ko-SBI (Lee et al., 2023b), and KOBEST’s BoolQ, HellaSwag, SentiNeg, COPA (Jang et al., 2022), which are described as follows:

- **Natural Language Inference** (**NLI**): A classification dataset predicting the relationship between two sentences.
- **Semantic Textual Similarity** (**STS**): A classification dataset measuring the semantic equivalence between two sentences.
- **YNAT**: A classification dataset that infers the topic of a given sentence.
- **SBI**: A classification dataset aimed at identifying social stereotypes or biases.

[4] `github/tatsu-lab/stanford_alpaca`

Model	Features	Backbone	Instruction	Pre-training
polyglot-ko-12.8b	**Monolingual** model	None	None	1.2TB
KoAlpaca-Polyglot-12.8b	+/mono SFT (21K)	polyglot-ko-12.8b	21K	None
kullm-polyglot-12.8b-v2	+/mono SFT (153K)	polyglot-ko-12.8b	153K	None
Llama2	**Multilingual** model	Llama-2-13b-hf	27K	2 trillion-token
Ko-Platypus2-13B	+/ mono SFT (25K)	Llama-2-13b-hf	25K	None
komt-Llama-2-13b-hf	+/ mono SFT (154K)	Llama-2-13b-chat-hf	1,543K	None
Llama2-koSFT (ours)	+/ mono SFT (1K)	Llama-2-13b-chat-hf	1K (Ko LIMA)	None
Llama2-ko (ours)	+/ mono PT (33GB)	Llama-2-13b-chat-hf	None	33Gb (Ko)
Bllossom-ko (ours)	+/ expand_vocab	Llama-2-13b-chat-hf	None	33Gb (Ko)
Bllossom-bi (ours)	+/ bilingual PT, expand_vocab	Llama-2-13b-chat-hf	None	33Gb (Ko:En=7:3)
Bllossom-bi-koSFT (ours)	+/ mono SFT(1K)	Bllossom-bi(ours)	1K (Ko LIMA)	None
Bllossom-bi-biSFT (ours)	+/ bilingual SFT(2K)	Bllossom-bi(ours)	2K (Ko-En,LIMA)	None

Table 4: Overview of the Korean LLMs (The model is from https://huggingface.co)

Prompt
질문: 문장 1과 문장 2는 서로 유사한 의미를 가지나요? (Question: Do sentence 1 and sentence 2 have similar meanings?) 문장 1: 습도기 보면 안된다고 경고했어 (Sentence 1: I warned not to look at the humidifier.) 문장 2: 습도기 자꾸 보려고 하지마 (Sentence 2: Don’t keep trying to look at the humidifier.) 정답: (Answer:)

Table 5: Evaluation prompt of STS task

- **BoolQ**: A question answering dataset for yes/ no questions.
- **HellaSwag**: A commonsense NLI dataset.
- **SentiNeg**: A sentiment classification data.
- **COPA**: A classification dataset determining the cause/effect based on a paragraph.

The experiments were performed using the Polyglot team' s public branch of EleutherAI' s lm-evaluation harness (Gao et al., 2021) to ensure reproducibility and compare the models. Each model was evaluated using the same data and prompt commands. Table 5 lists the STS evaluation prompts for which each system generated an answer.

For fair evaluation, the model to be evaluated must accurately represent the backbone of the training model and size of the used data, which are defined for the proposed model in Table 4. The Model column is structured in the format "Model-Pretrain Language-Option". The Pretrain Language value is 'bi' denotes a model that simultaneously uses the Korean and English languages for pretraining. The Option field denotes the application of SFT, where "biSFT" represents the implementation of the Korean and English LIMA data, whereas "koSFT" denotes the usage of only the Korean LIMA data. The Bllossom model refers to the model with vocabulary expansion applied to Llama2. For instance, Llama2-ko is a model pretrained on Llama2 in Korean and Bllossom-bi is a model pretrained in Korean and English after vocabulary expansion. The Bllossom-bi-koSFT is a Bllossom-bi model tuned using the Korean LIMA data. Polyglot-ko and KoAlpaca, are presented in Section 2.

5.2. Experiment Results

(**Overall**) Table 6 shows the performances of various models proposed in Table 4. Compared to the monolingual models (such as Polyglot-Ko, KoAlpaca, and Kullm), the proposed multilingual Bllossom models (referred to as "ours") exhibited an average performance with an increment of approximately 4.57 points. The MLLM performance was affected by the presence or absence of pretraining. The difference between the performances of Llama2-ko, which underwent only pretraining, and Llama2-koSFT, which underwent only SFT, was a substantial 6.2 points.

(**The influence of vocabulary expansion**) In Table 6, Bllossom-ko outperformed Llama2-ko by approximately 1.8 points. For NLI and STS which infer the relationship between two sentences, the Bllossom-ko model with an expanded vocabulary outperformed by 9.15 points. Contrastingly, Llama2-ko, which did not undergo vocabulary expansion, performed better on SBI by 8.8 points. Thus, vocabulary expansion improves the overall comprehension, reasoning, cognition, and causal understanding of the Korean language.

(**The influence of bilingual pretraining**) The Bllossom-ko and Bllossom-bi models in Table 6 differ on the usage of English and Korean bilingual training data during pretraining. The models exhibited similar performances with scores of 58.9 and 58.6, respectively. However, the following observations were made: (1) In contrast to Bllosson-bi, Bllossom-ko exhibited a bias issue wherein the model responded in Korean even when queried in English. (2) For the SBI tasks, Bllossom-bi outperformed by 11.6 points than Bllossom-ko. And it underperformed 11.2 and 10.6 points on the STS and HellaSwag tasks, respectively. Quantitatively, the impact of

Model	**NLI** ACC	**STS** ACC	**SBI** F1	**YNAT** ACC	**BoolQ** F1	**H-Swag** F1	**S-Neg** F1	**COPA** F1	AVG
polyglot-ko-12.8b	35.5	50.1	48.6	31.0	59.4	**48.8**	**95.7**	**81.0**	56.2
KoAlpaca-Polyglot-12.8b	38.0	42.7	48.4	26.0	66.4	44.4	84.8	80.0	53.8
kullm-polyglot-12.8b-v2	33.9	44.8	52.5	24.6	44.2	48.3	89.8	79.3	52.1
Llama2	44.0	45.8	56.0	25.4	73.8	40.7	78.1	60.9	53.1
Ko-Platypus2-13B	50.5	59.9	37.1	28.9	72.0	41.4	85.1	63.8	54.8
Komt-Llama-2-13b-hf	33.4	51.6	48.7	24.2	52.6	39.7	62.4	64.2	47.1
Llama2-koSFT (ours)	44.5	50.6	38.5	23.1	71.7	41.2	77.3	60.5	50.9
Llama2-ko (ours)	41.5	47.4	61.7	32.6	72.8	43.5	89.1	68.4	57.1
Bllossom-ko (ours)	49.4	57.8	52.9	33.1	73.0	48.6	87.9	69.0	**58.9**
Bllossom-bi (ours)	48.8	46.6	**64.5**	32.8	74.0	38.0	93.2	71.2	58.6
Bllossom-bi-koSFT (ours)	**49.6**	**54.9**	55.0	33.9	**74.2**	40.0	92.0	68.4	58.5
Bllossom-bi-biSFT (ours)	45.7	46.4	63.4	**36.0**	69.4	39.1	89.9	70.0	57.5

Table 6: Benchmarking Korean LLMs: Accuracy (ACC) and F1 score metrics across tasks

bilingual pretraining was minimal; however, a significant performance difference was qualitatively observed owing to bilingual pretraining.

(**The influence of SFT on the Korean LIMA**) This experiment evaluated the impact of 1K Korean LIMA data by comparing Llama2 and Llama2-koSFT, which performed SFT on Llama2. In Table 6, Llama2, the backbone, outperformed by an average of 2.2 points. Similarly, the performance of Komt, which underwent an extensive SFT on Llama2, was approximately reduced by six points. The models based on Polyglot-ko, such as KoAlpaca and Kullm, exhibited lower performance than that of the backbone. Therefore, SFT may not significantly influence the quantitative evaluations of classification tasks. However, Llama2-koSFT empirically produced better responses than Llama2 based on qualitative factors, such as the quality of the generated responses, vocabulary, and completeness. Therefore, the following section analyzes the effects of SFT based on qualitative evaluations performed by humans (Lee et al., 2023c) and GPT.

6. Qualitative evaluation

Based on LIMA (Zhou et al., 2023), the qualitative evaluation was performed by humans and GPT. The former involved posing the same question to LLMs A and B and the evaluators subsequently deciding among the responses based on the following three options: Model A is better, Model B is better, or neither is significantly better. Contrastingly, GPT4-based evaluation enabled the GPT to decide among these options. We translated 300 entries from the LIMA human evaluation test dataset into Korean and proceeded with evaluation. According to the LIMA study, the 1k LIMA training data and LIMA human evaluation test dataset were designed to have completely different topics, styles, and tasks. Therefore, tuning the model using LIMA training data negligibly improves the performance (Zhou et al., 2023).

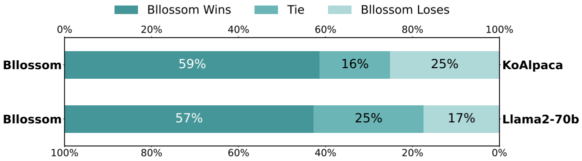

Figure 2: Preference evaluation results by human

(**Overall**) Figure 2 and Figure 3 show the results of human and GPT4 evaluations, respectively. When comparing Bllossom to Koalpaca and Kullm models of the same size, Bllossom outperformed them in human and GPT4 evaluations and even outperformed the larger Llama2-70b-chat model. Another interesting point is the qualitative evaluation results for human and GPT4 were similar. This was also observed in LIMA.

6.1. Human-assisted preference evaluation

In Figure 2, the number of times Bllossom won both Koalpaca and Llama2 in human evaluation was 124. The 124 tasks included answering real-world user requests for recommendations, answering questions requiring imagination or creativity, organizing travel itineraries and writing code. In contrast, there were 40 instances where Bllossom gave inferior answers compared to both models, mostly for factual QA. This suggests that Bllossom is not yet as knowledgeable as the larger models, which is likely due to the difference in data size from the pre-training phase.

6.2. Preference Evaluation using GPT4

Using the methodology proposed by LIMA, GPT4 was used to compare the performance of Bllossom with six other models. The evaluation was conducted on the Korean LIMA test data.

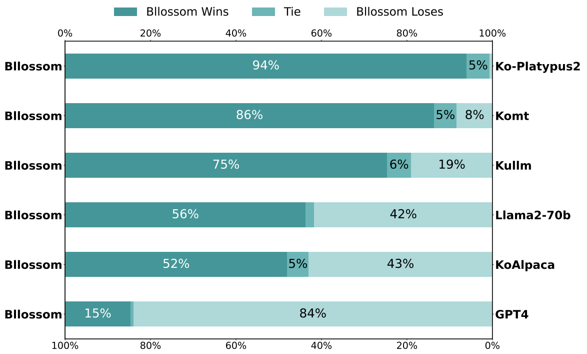

Figure 3: Preference evaluation results by GPT4

(**Comparing Bllossom with Korean models based on Llama2**) Figure 3 shows that Komt and Ko-Platypus2 are models based on Llama2-13b that underwent SFT, similar to Bllossom. However, these two models were exclusively subjected to SFT without vocabulary expansion or pretraining. They were fine-tuned using extensive datasets that were either autogenerated or translated, and thus, of lower quality. Qualitatively, Bllossom exhibited superior performance with a margin exceeding 40%, indicating that pretraining significantly influences the Korean proficiency. During the human evaluation of history-related questions, Komt and Ko-Platypus2 either failed to provide answers or exhibited hallucinations more frequently compared to Bllossom. This can be attributed to Bllossom gaining additional knowledge during pretraining.

(**Comparing Bllossom with Polyglot-ko-based Korean models**) We discuss whether the proposed Bllossom model exhibits a better qualitative evaluation than Korean monolingual LLMs. Polyglot-ko is a representative Korean monolingual model pretrained on vast Korean datasets and KoAlpaca and Kullm are models trained based on Polyglot-ko. Figure 3, shows that the Bllossom model has a 9~56% higher probability of producing superior answers than the two monolingual models utilizing Polyglot-Ko as their backbone. During pretraining, Llama2 incorporated a relatively limited set of Korean data; however, its training dataset significantly expanded when English data was included, surpassing the dataset size of Polyglot-ko. This suggests that the bilingual pretraining, which was carried out to augment the deficient proficiency in Korean, has somewhat assisted in bridging the knowledge between Korean and English (Cui et al., 2023).

(**Comparing Bllossom with GPT4 and Llama**) We discuss the Korean-language proficiency of the proposed Bllossom model. The Llama2-70b model, which has significantly more parameters, was evaluated. Based on the results in Figure 3, the Bllossom model was selected for approximately 14% of the answers than Llama2-70B. Therefore, in case of an extreme difference in the number of parameters, the differences in performance can be fairly compensated via techniques such as word expansion and pretraining. The qualitative evaluation results for OpenAI' s much larger GPT4 model indicated its superiority in frequent answering.

(**The effect of bilingual dataset for SFT**) In Figure 4, the Bllossom-bi-koSFT model and the Bllossom-bi-biSFT model differ based on whether bilingual data was utilized for SFT. We conducted a comparative evaluation of the two models using both Korean and English. For the Korean and English LIMA test data, the win ratio for the Bllossom-bi-biSFT model was overwhelmingly high at 67% and 95%, respectively. This indicates that contrary to qualitative evaluation, the effect of bilingual SFT in quantitative evaluation is significant.

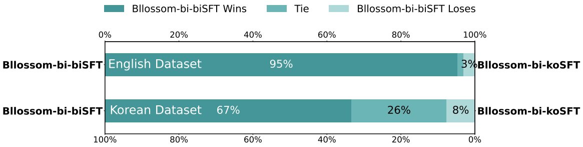

Figure 4: Comparing **bi**SFT and **ko**SFT models

(**The effect of English language**) We investigated whether a model tuned to Korean based on Llama2 would perform poorly in English. As shown in Figure 5, all the models appear to have significantly lost their English proficiency compared to the original Llama2 model. Nevertheless, Bllossom showed much better performance compared to Komt and Ko-Platypus2. From this, we can infer that while acquiring Korean proficiency, the Bllossom model has a lesser reduction in English capabilities compared to other Korean LLMs.

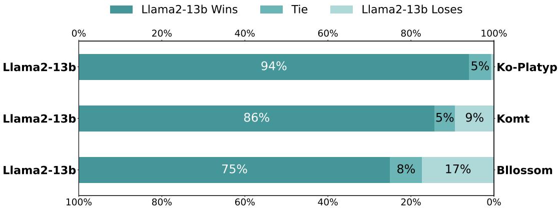

Figure 5: Comparing English performance in Llama2-backboned models

7. Conclusion

This study proposed three methods for enhancing the MLLM capability of LRLs. First, to improve the Korean vocabulary capability of the existing Llama2 model, the vocabularies from KoBERT and Llama2 were merged to create a new embedding. Second, pretraining was performed using bilingual data to enhance knowledge information by aligning high- and low-resource languages. Third, instruction-tuning was performed using the English and refined Korean LIMA datasets to accurately understand the user intent and produce the desired response. Quantitative assessments were performed using eight benchmark datasets and qualitative assessments were conducted using humans and the GPT4 model to investigate the proposed model. The experimental results revealed that the proposed `Bllossom` model outperformed the pre-existing Korean monolingual models that require vast computing resources and supervised data.

8. Ethical Considerations

While we have no ethical concerns regarding the current work, our commitment to upholding the highest ethical standards in all our activities and human evaluations remains unwavering.

9. Limitations

In this paper, we proposed a method to enhance Korean language in MLLM. However, to apply the same method to other languages, the following efforts are required. (1) For building LIMA data, one needs to translate 1,030 data, (2) one also need to translate 300 training data for testing.

Acknowledgements

This research was supported by the National Research Foundation of Korea for KyungTae Lim (2021R1F1A1063474), Institute of Information & communications Technology Planning & Evaluation (IITP) by the Korea government(MSIT) (2022-0-00078, Explainable Logical Reasoning for Medical Knowledge Generation) and the Yonsei University Research Fund of 2023 (2023-22-0410).

10. Bibliographical References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine

Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Myeongjun Jang, Dohyung Kim, Deuk Sin Kwon, and Eric Davis. 2022. KoBEST: Korean balanced evaluation of significant tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3697–3708, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiwung Hyun, Sungho Park, and Kyubyong Park. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

NLP AI Lab and Human-Inspired AI research. 2023. Kullm: Korea university large language model project. https://github.com/nlpai-lab/kullm.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023a. Platypus: Quick, cheap, and powerful refinement of llms.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023b. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023c. Qasa: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Lucy Park, Alice Oh, Jung-Woo Ha (NAVER AI Lab), Kyunghyun Cho, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023a. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023b. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. a. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. b. Language models are unsupervised multitask learners.

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. *J. Mach. Learn. Res.*, 21(1).

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hai Wang, Dian Yu, Kai Sun, Janshu Chen, and Dong Yu. 2019. Improving pre-trained multilingual models with vocabulary expansion.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo

Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers?

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.