LREC-COLING 2024

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

MLP Lab

Contents

Bllossom

01 Previous Research and Research Necessity
02 Research Methodology
03 Analysis Results (Quantitative, Qualitative)
04 Discussion and Implications

Author Introduction

Presenter

ChangSu Choi

M.S. degree in Artificial Intelligence Applications at SeoulTech

KyungTae Lim, Prof.
SeoulTech

Younggyun Hahm, Ph.D.
SeoulTech

Hansaem Kim, Prof.
Yonsei University

Seoyoon Park, Ph.D. candidate
Yonsei University

HyeonSeok Lim, M.S.
SeoulTech

InHo Won, M.S.
SeoulTech

SangMin Kim, M.S.
SeoulTech

Yongbin Jeong, M.S.
TeddySum

Yejee Kang, Ph.D.
TeddySum

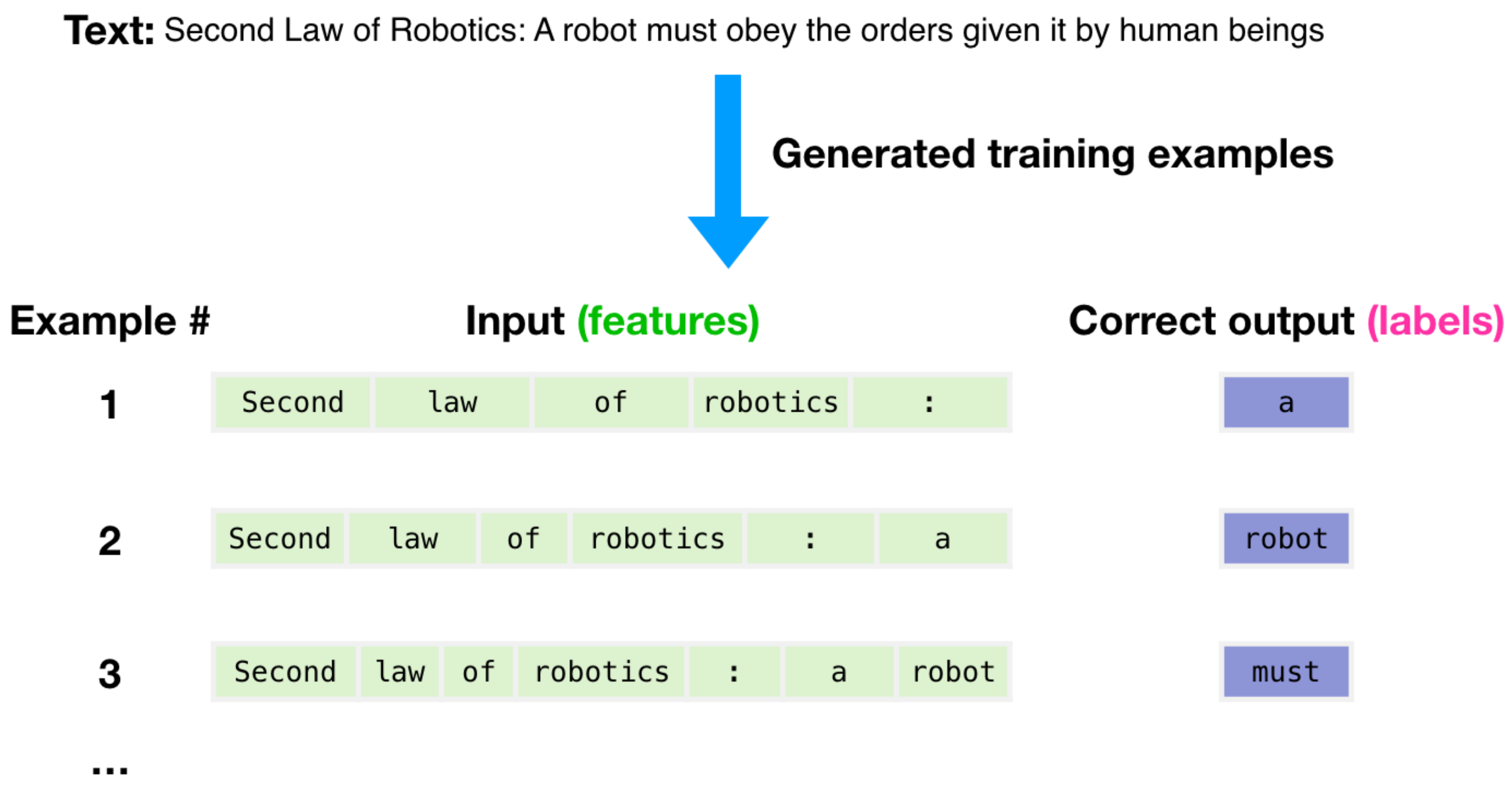
Chanhyuk Yoon
TeddySum

01. Previous Research and Research Necessity

01 Previous Research and Research Necessity

LLMs predict the next word based on large-scale text data to understand linguistic information and knowledge.

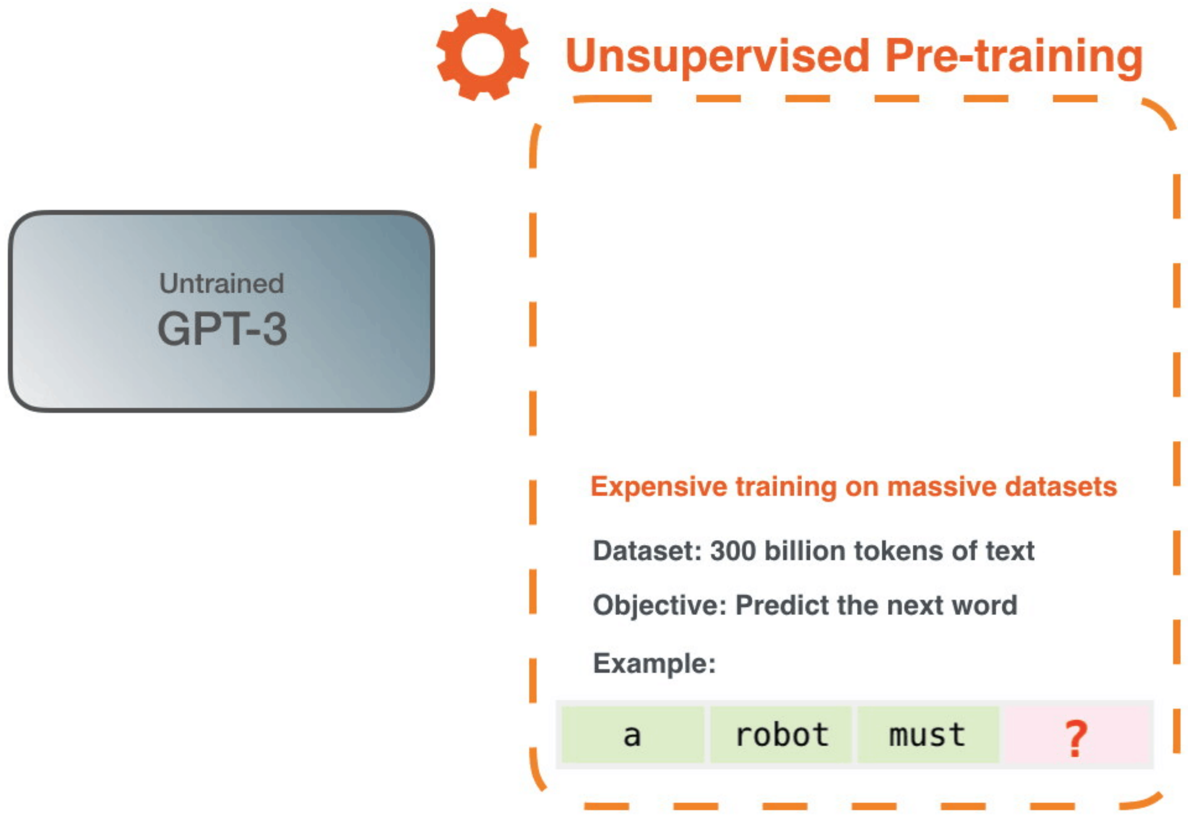
- Strong generalization capabilities (e.g., ChatGPT, Gemini).
- Training is conducted on general plain text, followed by unsupervised pretraining.



Text: Second Law of Robotics: A robot must obey the orders given it by human beings

Generated training examples

Example #	Input (features)							Correct output (labels)
1	Second	law	of	robotics	:			a
2	Second	law	of	robotics	:	a		robot
3	Second	law	of	robotics	:	a	robot	must
...								



Untrained GPT-3

Unsupervised Pre-training

Expensive training on massive datasets

Dataset: 300 billion tokens of text

Objective: Predict the next word

Example:

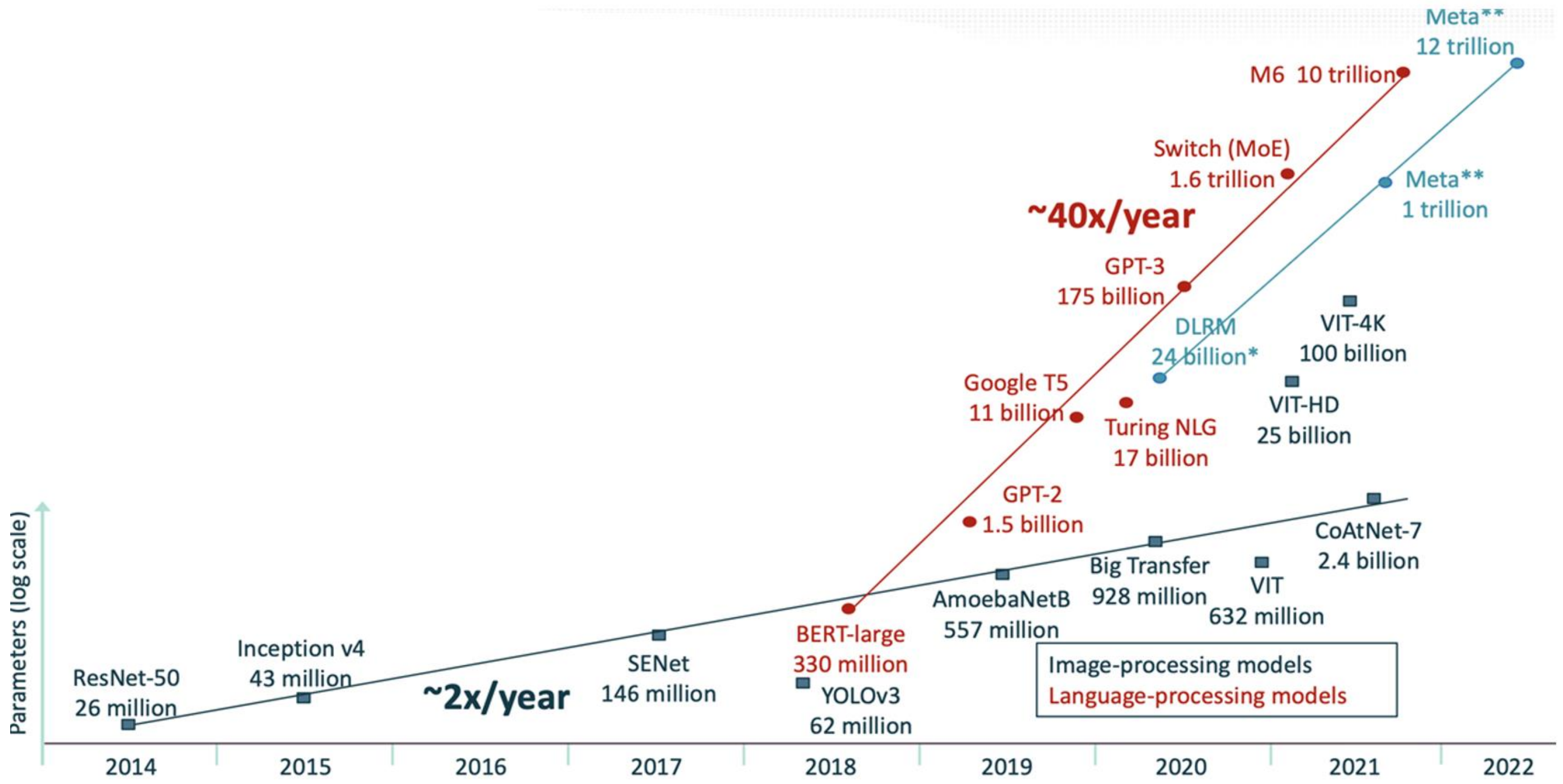
a	robot	must	?

https://jalammar.github.io/how-gpt3-works-visualizations-animations/

01 Previous Research and Research Necessity

- Training an LLM requires significant resources! Therefore, there is a critical need for publicly available models.

- Due to this demand, many companies and research groups have released multilingual LLMs.

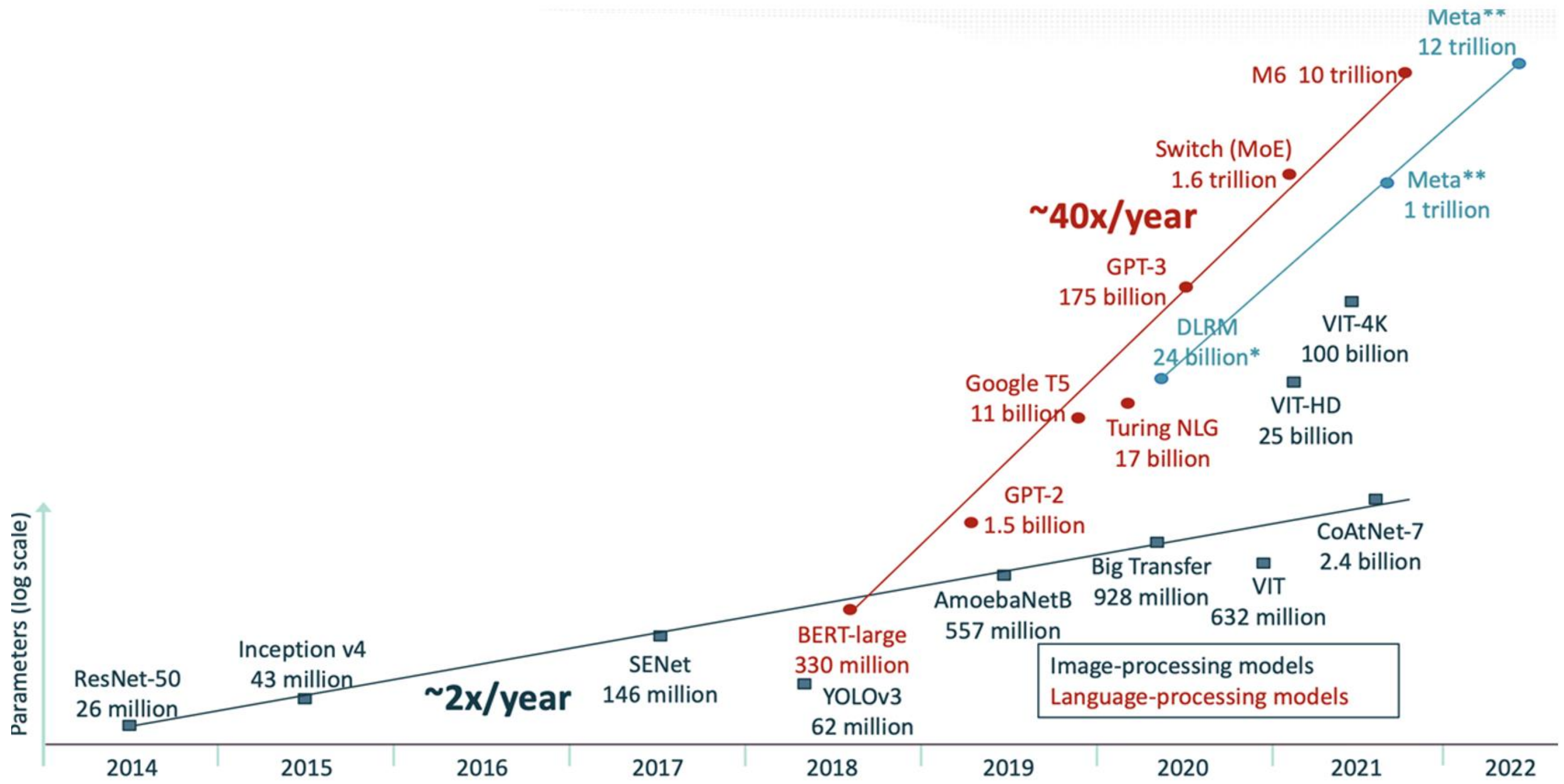


01 Previous Research and Research Necessity

- Training an LLM requires substantial resources! Therefore, there is a critical need for publicly available models.

- Due to this demand, many companies and research groups have released multilingual LLMs.

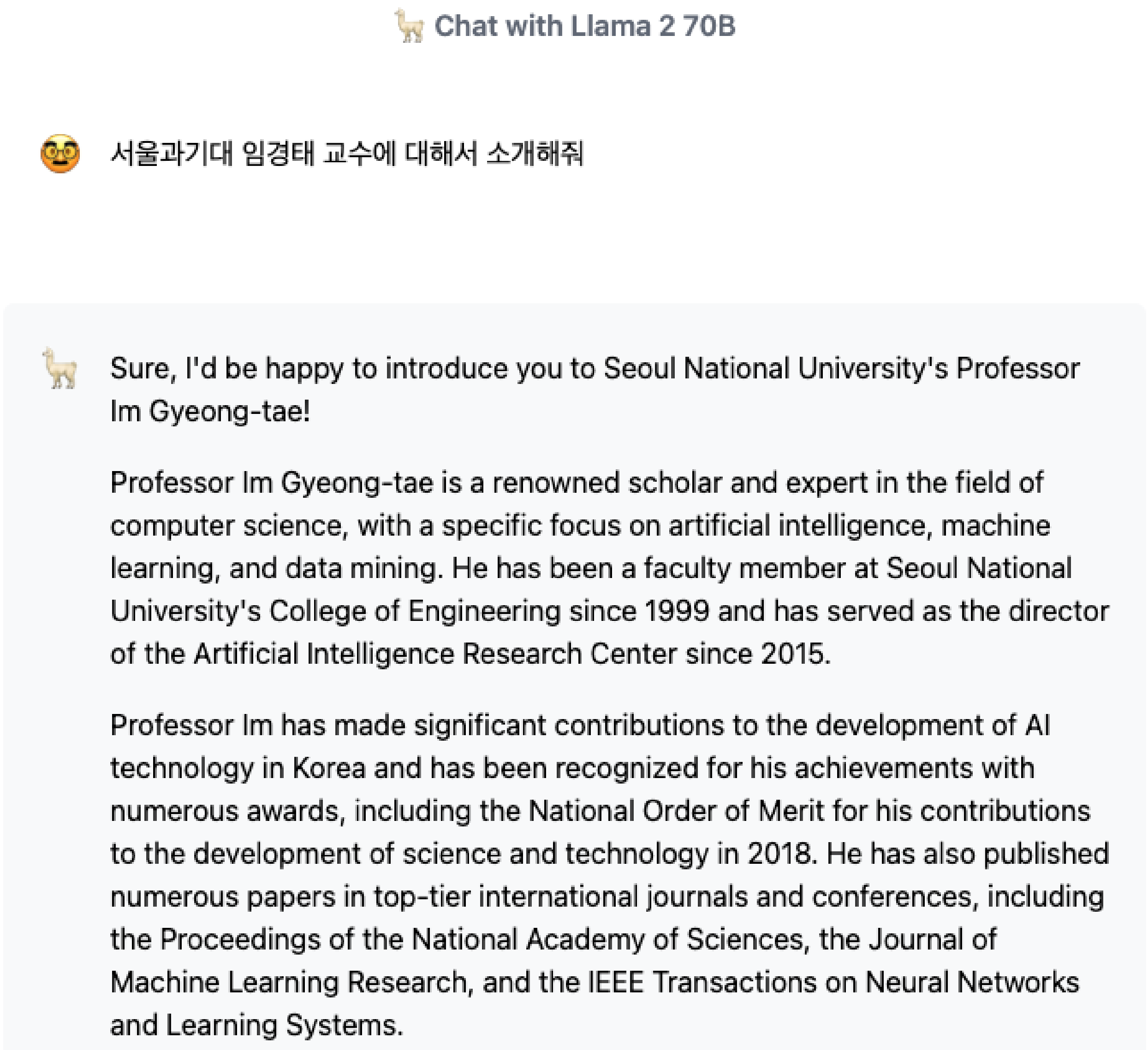
However, the released multilingual models have relatively less focus on low-resource languages



01 Previous Research and Research Necessity

If you ask a question in Korean to META's LLAMA2, one of the most advanced multilingual LLMs, it responds in English!

- Even though it understands Korean, it doesn't know enough words and stuff about the language to respond like a natural speaker would



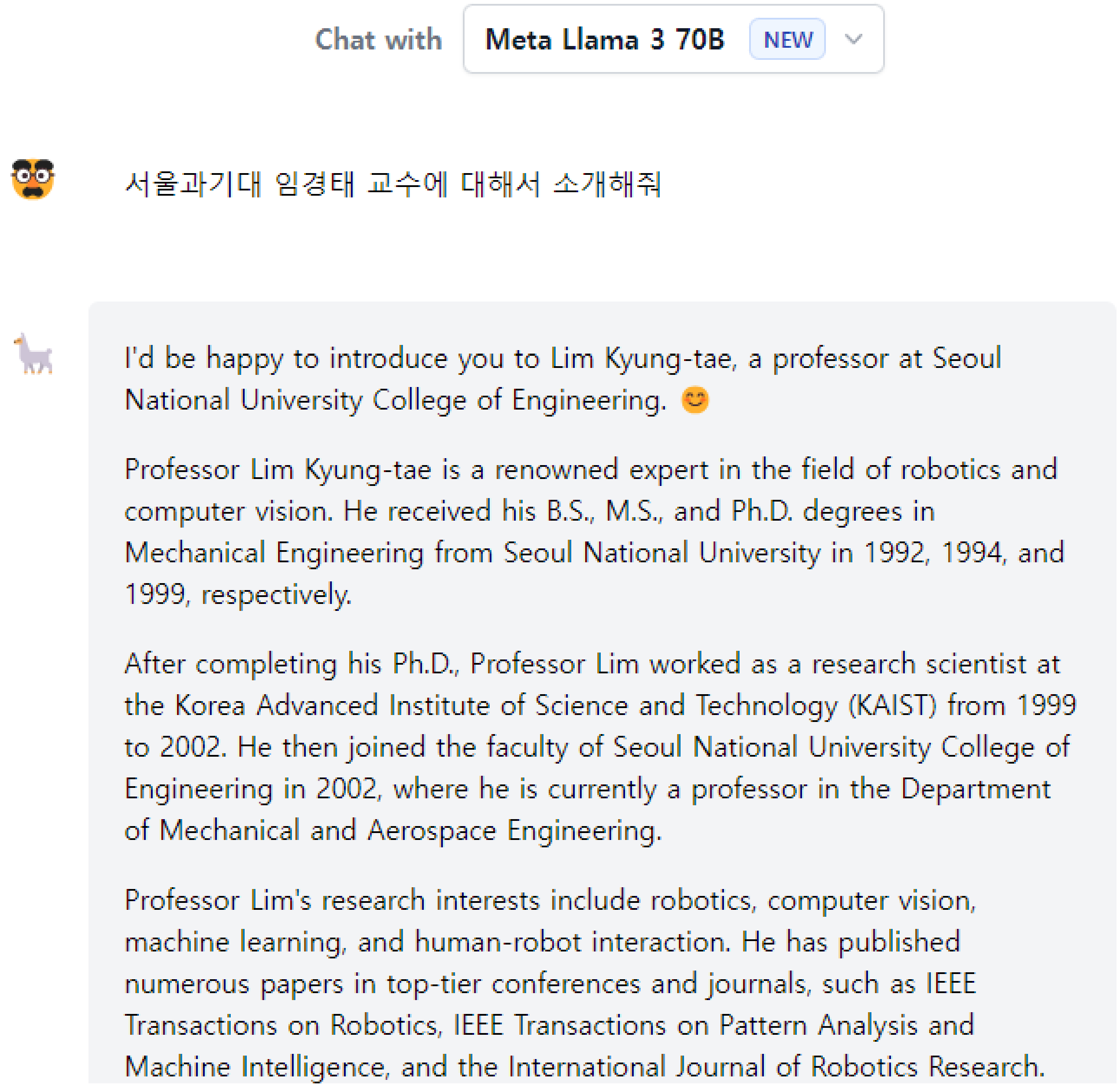
Chat with Llama 2 70B

서울과기대 임경태 교수에 대해서 소개해줘

Sure, I'd be happy to introduce you to Seoul National University's Professor Im Gyeong-tae!

Professor Im Gyeong-tae is a renowned scholar and expert in the field of computer science, with a specific focus on artificial intelligence, machine learning, and data mining. He has been a faculty member at Seoul National University's College of Engineering since 1999 and has served as the director of the Artificial Intelligence Research Center since 2015.

Professor Im has made significant contributions to the development of AI technology in Korea and has been recognized for his achievements with numerous awards, including the National Order of Merit for his contributions to the development of science and technology in 2018. He has also published numerous papers in top-tier international journals and conferences, including the Proceedings of the National Academy of Sciences, the Journal of Machine Learning Research, and the IEEE Transactions on Neural Networks and Learning Systems.



Chat with **Meta Llama 3 70B** NEW

서울과기대 임경태 교수에 대해서 소개해줘

I'd be happy to introduce you to Lim Kyung-tae, a professor at Seoul National University College of Engineering. 😊

Professor Lim Kyung-tae is a renowned expert in the field of robotics and computer vision. He received his B.S., M.S., and Ph.D. degrees in Mechanical Engineering from Seoul National University in 1992, 1994, and 1999, respectively.

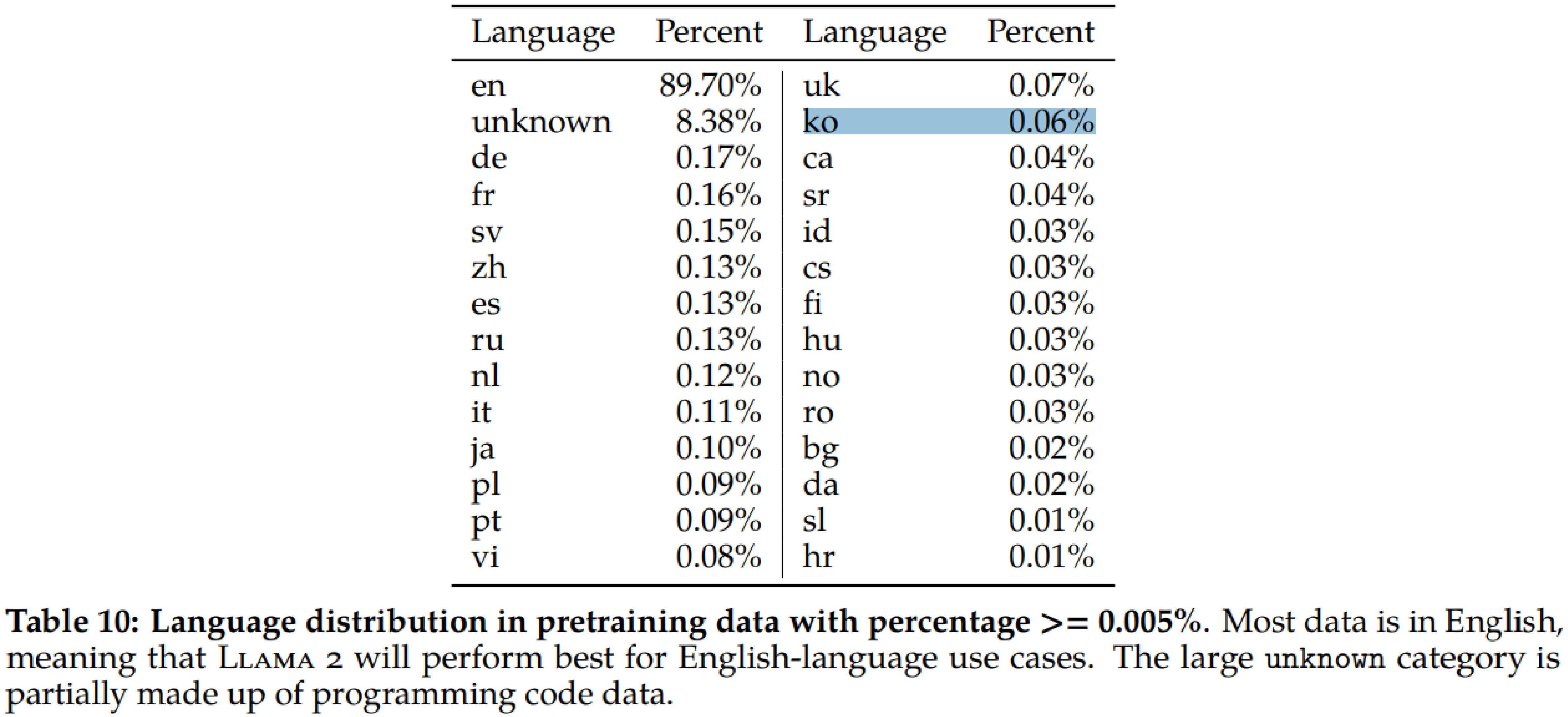
After completing his Ph.D., Professor Lim worked as a research scientist at the Korea Advanced Institute of Science and Technology (KAIST) from 1999 to 2002. He then joined the faculty of Seoul National University College of Engineering in 2002, where he is currently a professor in the Department of Mechanical and Aerospace Engineering.

Professor Lim's research interests include robotics, computer vision, machine learning, and human-robot interaction. He has published numerous papers in top-tier conferences and journals, such as IEEE Transactions on Robotics, IEEE Transactions on Pattern Analysis and Machine Intelligence, and the International Journal of Robotics Research.

01 Previous Research and Research Necessity

The reason META's LLAMA2 does not respond in Korean is because only 0.06% of the model's training data is in Korean!

- Due to a lack of vocabulary and semantic knowledge, the model's utility can be limited.

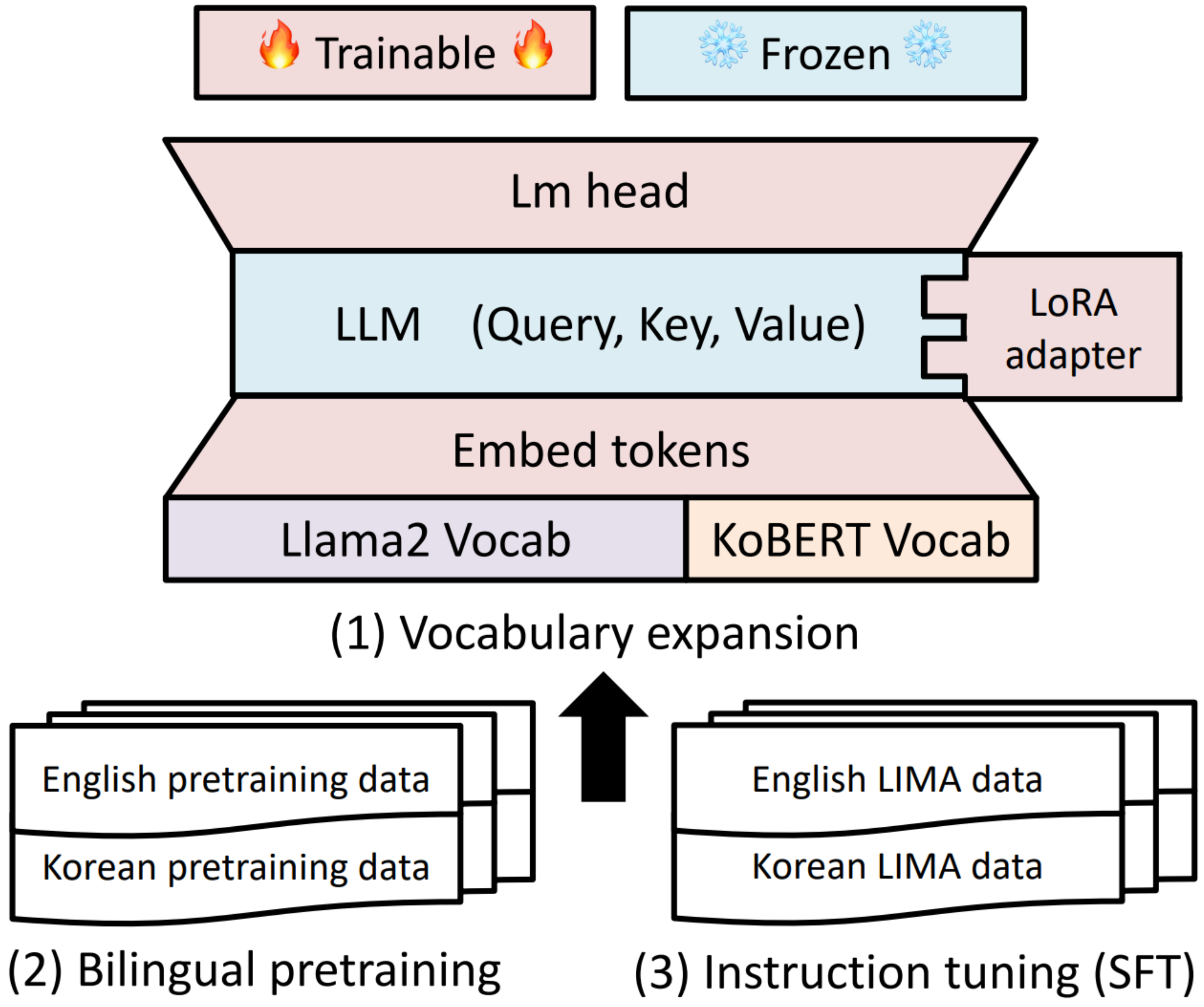


Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Table 10: Language distribution in pretraining data with percentage >= 0.005%. Most data is in English, meaning that LLAMA 2 will perform best for English-language use cases. The large `unknown` category is partially made up of programming code data.

01 Previous Research and Research Necessity

This study proposes the Bllossom model, which introduces three performance enhancement methods for Korean, a Less-Resourced Language (LRL) in LLAMA2

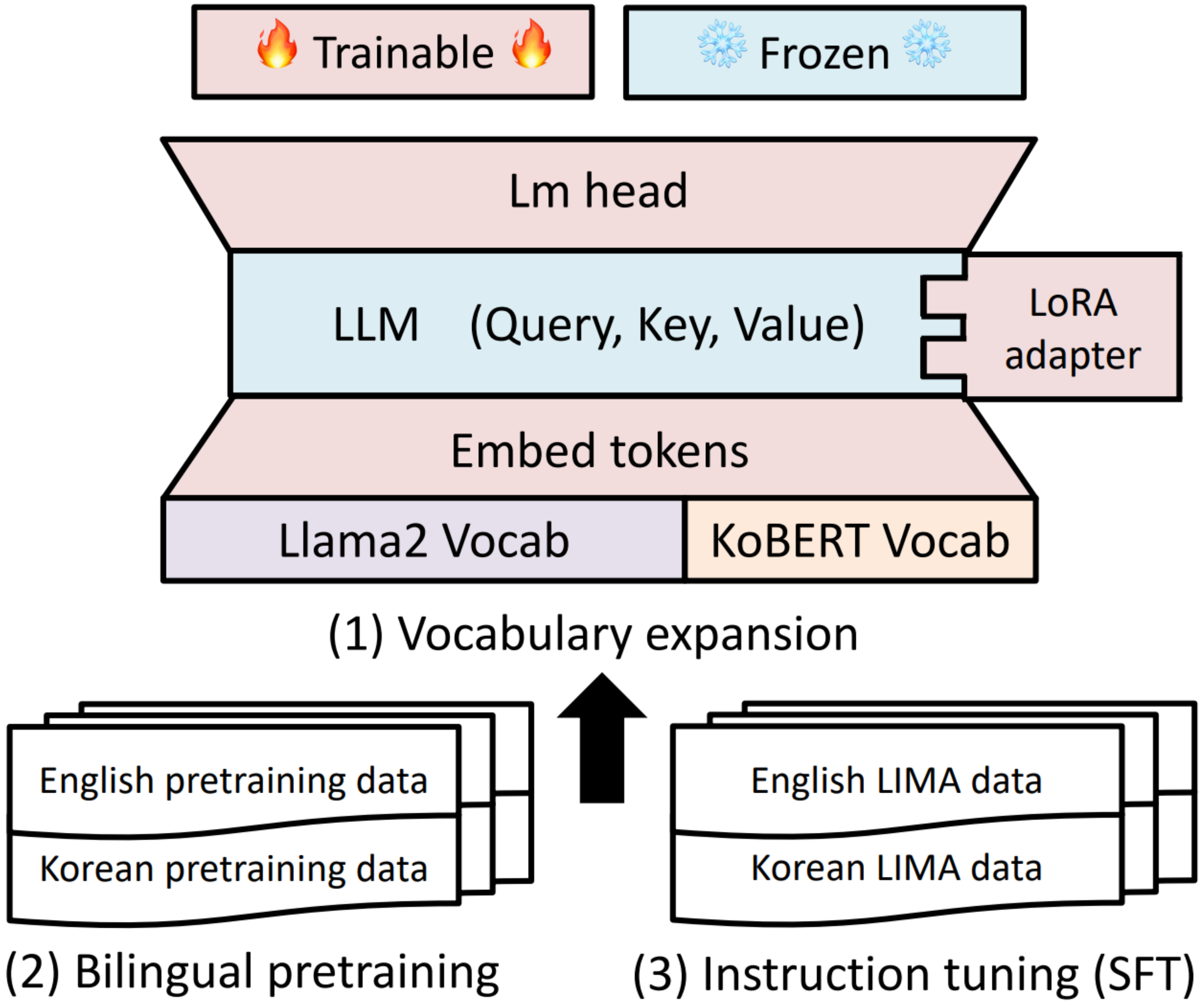


02. Research Methodology

02 Research Methodology

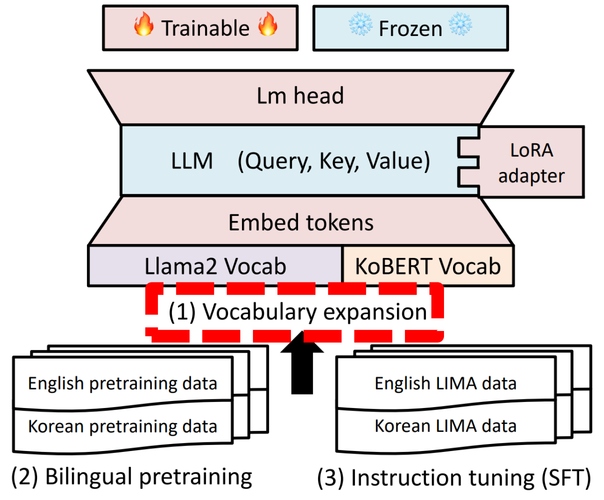
Three Strategies to Enhance Korean Language Performance in Multilingual LLMs

1. Vocabulary expansion
2. Knowledge enrichment
3. Usability enhancement



02 Research Methodology

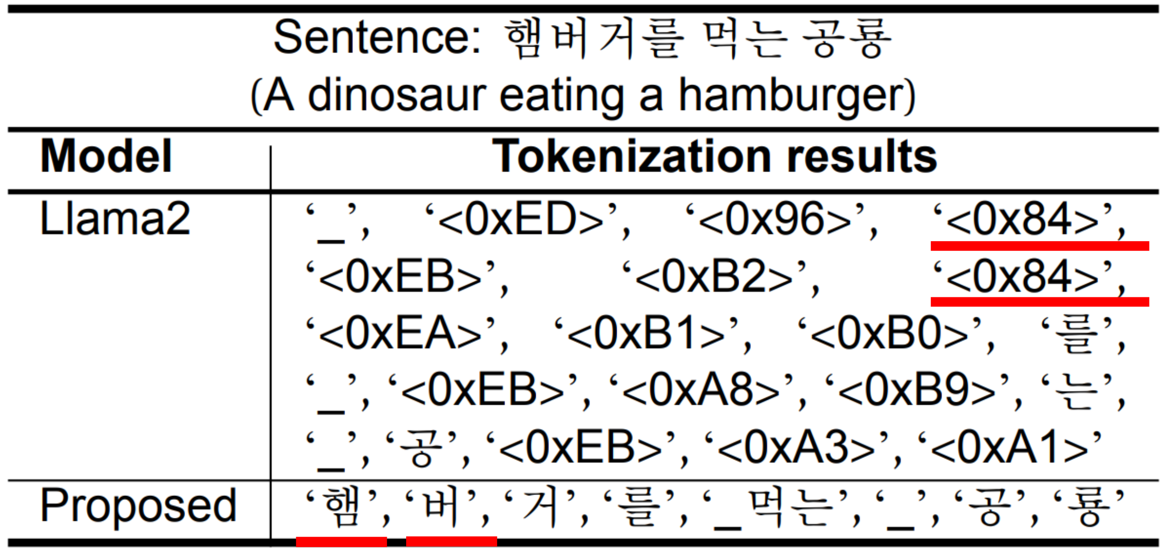
1. Vocabulary expansion: Enhancing Korean Vocabulary



Challenges of Utilizing Korean Language in the Existing LLama2 Model:

1. Increased Token Length: The model must represent Out-Of-Vocabulary (OOV) words, which cannot be expressed with a single token, using three or four byte tokens. This reduces the number of characters that can be inputted into the model and increases encoding and decoding times.

2. Redundancy of Byte Tokens: Unrelated tokens such as "햄" (ham) and "버" (ver) are represented using the same byte token "<0x84>". Consequently, the model may learn these semantically unrelated words as partially identical, leading to confusion during training.

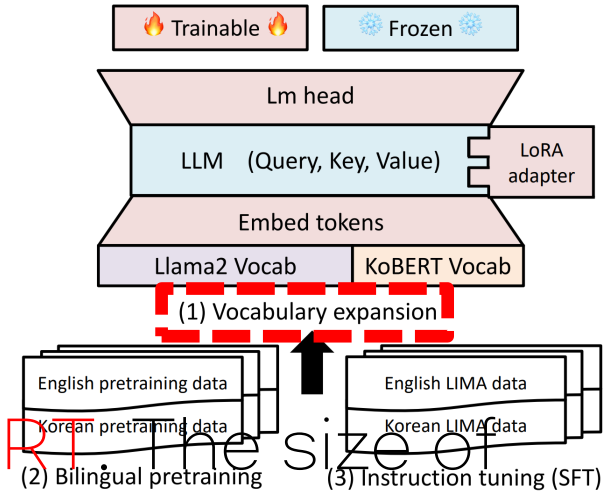


Sentence: 햄버거를 먹는 공룡 (A dinosaur eating a hamburger)	
Model	**Tokenization results**
Llama2	‘_’, ‘<0xED>’, ‘<0x96>’, ‘<0x84>’, ‘<0xEB>’, ‘<0xB2>’, ‘<0x84>’, ‘<0xEA>’, ‘<0xB1>’, ‘<0xB0>’, ‘를’, ‘_’, ‘<0xEB>’, ‘<0xA8>’, ‘<0xB9>’, ‘는’, ‘_’, ‘공’, ‘<0xEB>’, ‘<0xA3>’, ‘<0xA1>’
Proposed	‘햄’, ‘버’, ‘거’, ‘를’, ‘_먹는’,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

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

02 Research Methodology

1. Vocabulary expansion: Enhancing Korean Vocabulary



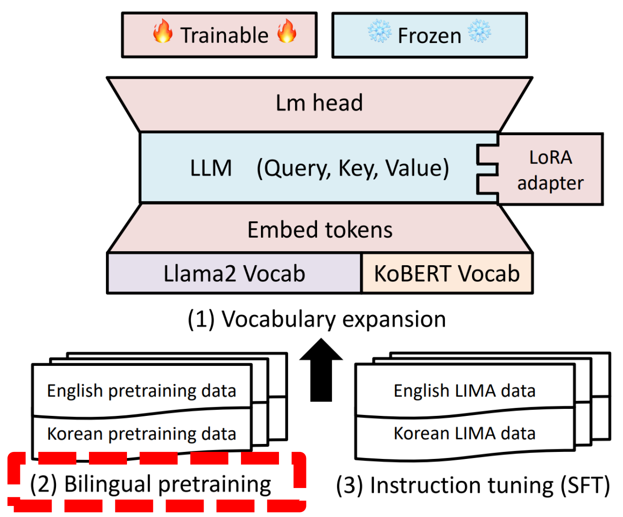
Vocabulary expansion is achieved by merging the dictionaries of Llama2 and KoBERT. The size of Llama2's tokenizer dictionary is $|D_L| = 32{,}000$, with over 90% consisting of English or Latin characters. KoBERT's tokenizer dictionary is sized at $|D_K| = 8{,}002$.

When merging these two dictionaries, excluding duplicates, the resulting combined dictionary size is $|D| = |D_L \cup D_K| = 39{,}478$.

Thus, Bllossom will train both the pre-existing word embeddings from Llama2 and the newly initialized 7,478-sized word embeddings.

02 Research Methodology

2. Knowledge enrichment: Enhancing Vocabulary and Knowledge through Pretraining



Challenges with the existing Llama2: Although the model understands Korean, it tends to respond in English due to its extensive English knowledge base.

Input	“이탈리아 수도에 대해 한국어로 소개해줘 (Introduce the capital of Italy in Korean.)”
Output	“로마 is the capital city of Italy and …”

Proper nouns such as “로마 (Rome)” and “콜로세움 (Colosseum)” are generated in Korean, but the detailed content is produced in English.

➔This is because the knowledge from pretraining is primarily in English.

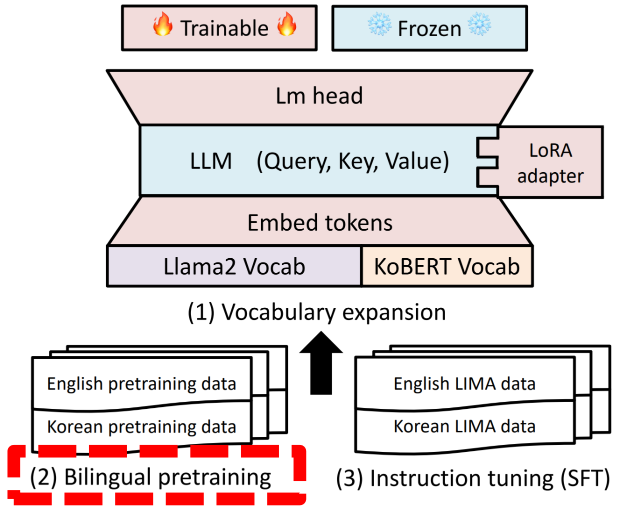
➔To address this, additional Bilingual Pretraining can be used to align Korean and English knowledge.

After Bilingual Pretraining:

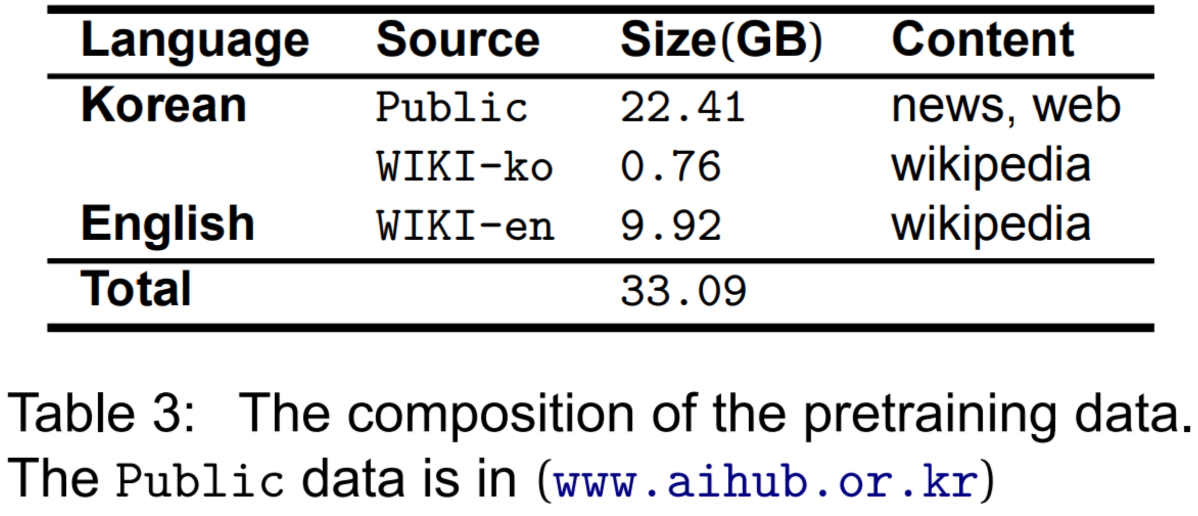
Input	“이탈리아 수도에 대해 한국어로 소개해줘 (Introduce the capital of Italy in Korean.)”
Output	“로마는 이탈리아의 수도이며…(로마 is the capital city of Italy and …)”

02 Research Methodology

2. Knowledge enrichment: Enhancing Vocabulary and Knowledge through Pretraining



Used Bilingual Pretraining Data



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	

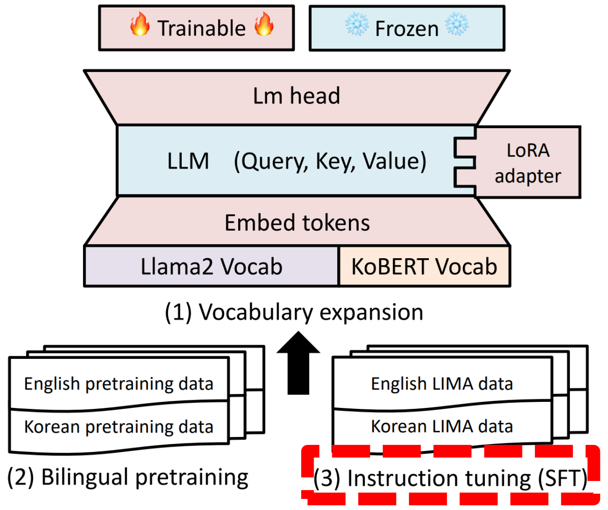
Table 3: The composition of the pretraining data. The `Public` data is in (`www.aihub.or.kr`)

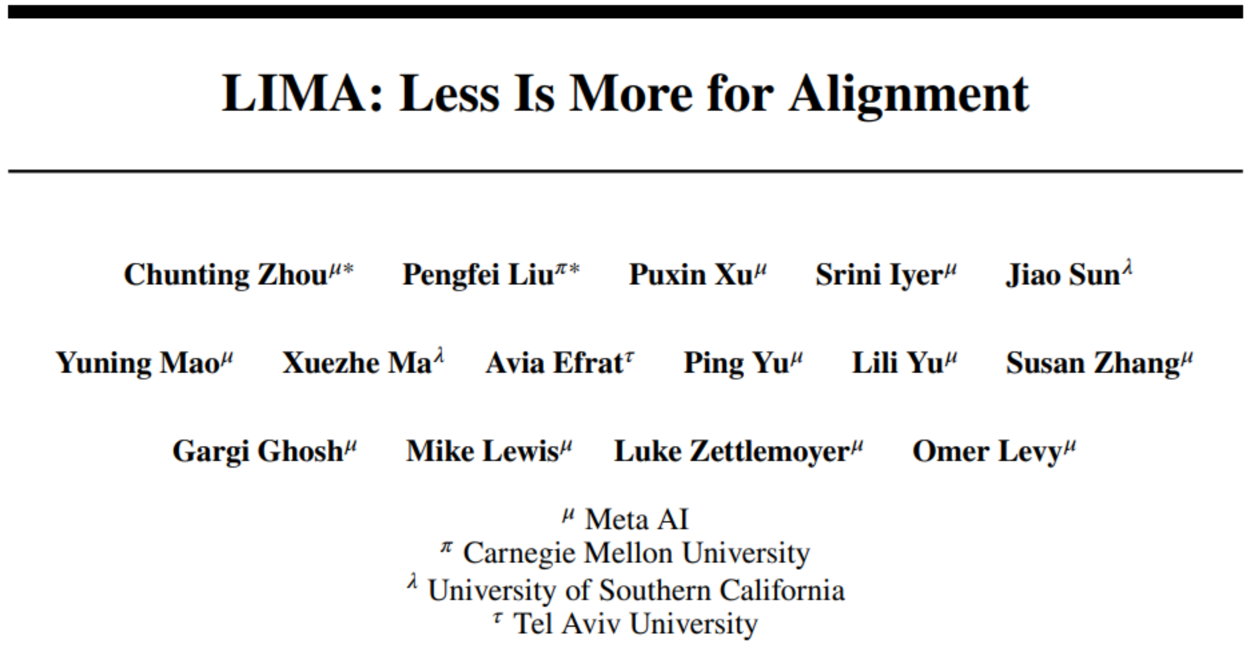
Loss Function

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

02 Research Methodology

3. Usability enhancement: Creating Instruction Data to Enhance the Utilization of Korean





LIMA: Less Is More for Alignment

Chunting Zhou[μ*] **Pengfei Liu**[π*] **Puxin Xu**[μ] **Srini Iyer**[μ] **Jiao Sun**[λ]
Yuning Mao[μ] **Xuezhe Ma**[λ] **Avia Efrat**[τ] **Ping Yu**[μ] **Lili Yu**[μ] **Susan Zhang**[μ]
Gargi Ghosh[μ] **Mike Lewis**[μ] **Luke Zettlemoyer**[μ] **Omer Levy**[μ]

[μ] Meta AI
[π] Carnegie Mellon University
[λ] University of Southern California
[τ] Tel Aviv University

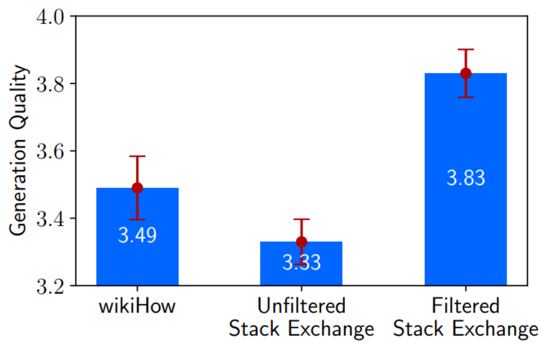


Figure 5: Performance of 7B models trained with 2,000 examples from different sources. **Filtered Stack Exchange** contains diverse prompts and high quality responses; **Unfiltered Stack Exchange** is diverse, but does not have any quality filters; **wikiHow** has high quality responses, but all of its prompts are "how to" questions.

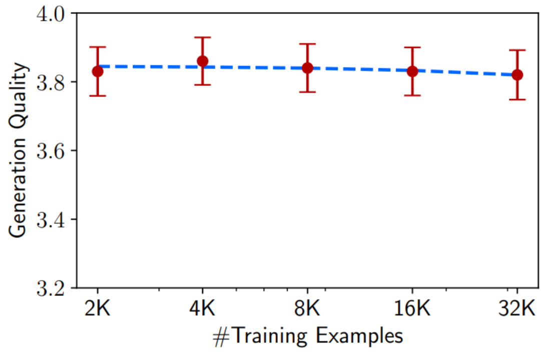


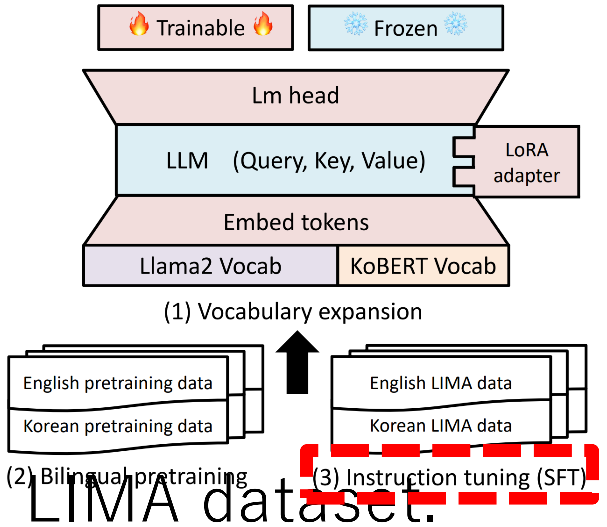
Figure 6: Performance of 7B models trained with exponentially increasing amounts of data, sampled from (quality-filtered) Stack Exchange. Despite an up to 16-fold increase in data size, performance as measured by ChatGPT plateaus.

The LIMA paper asserts that the quality of answers from Foundation Models is largely independent of additional pretraining, and varies according to the quality of the data

In other words, significant effects can be achieved through Instruction Tuning with small-scale but high-quality data

02 Research Methodology

3. Usability enhancement: Creating Instruction Data to Enhance the Utilization of Korean



The Korean LIMA dataset for SFT is built based on a machine-translated version of the English LIMA dataset.

Machine Translation Side Effects:

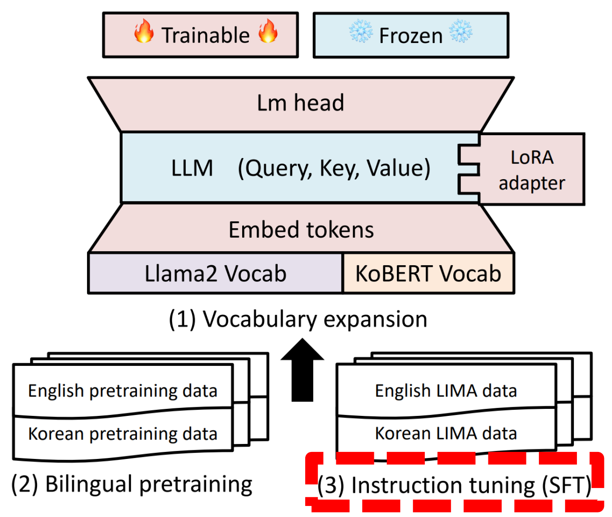
1. Inconsistency in language style between English and Korean.
2. Exclusion of Korean cultural context due to the characteristics of the source data.

Post-processing:

1. Replace named entities, change themes.
2. At this point, adjustments are made to the LIMA data to align it with Western cultural contexts, encompassing a wide spectrum from political to economic and social systems.

02 Research Methodology

3. Usability enhancement: Creating Instruction Data to Enhance the Utilization of Korean



Example of Post-Processed LIMA Korean Data

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??

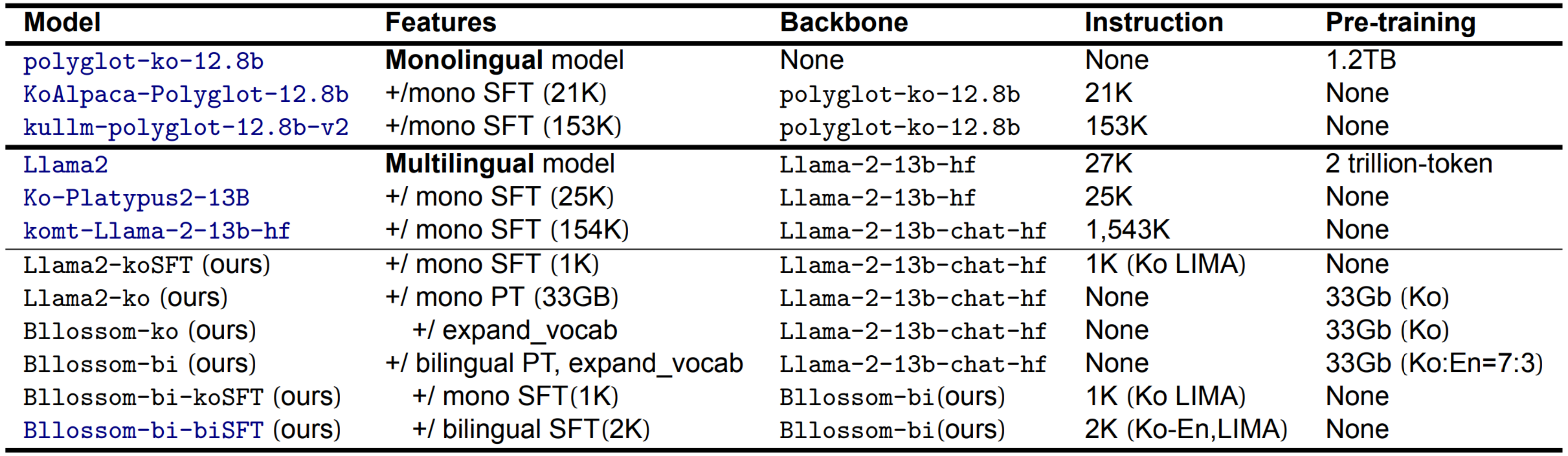
Table 2: Instances of modifications in the English LIMA dataset to reflect the Korean cultural context

Loss Function

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

03-1. Analysis Results (Qualitative)

03 Analysis Results (Quantitative)



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

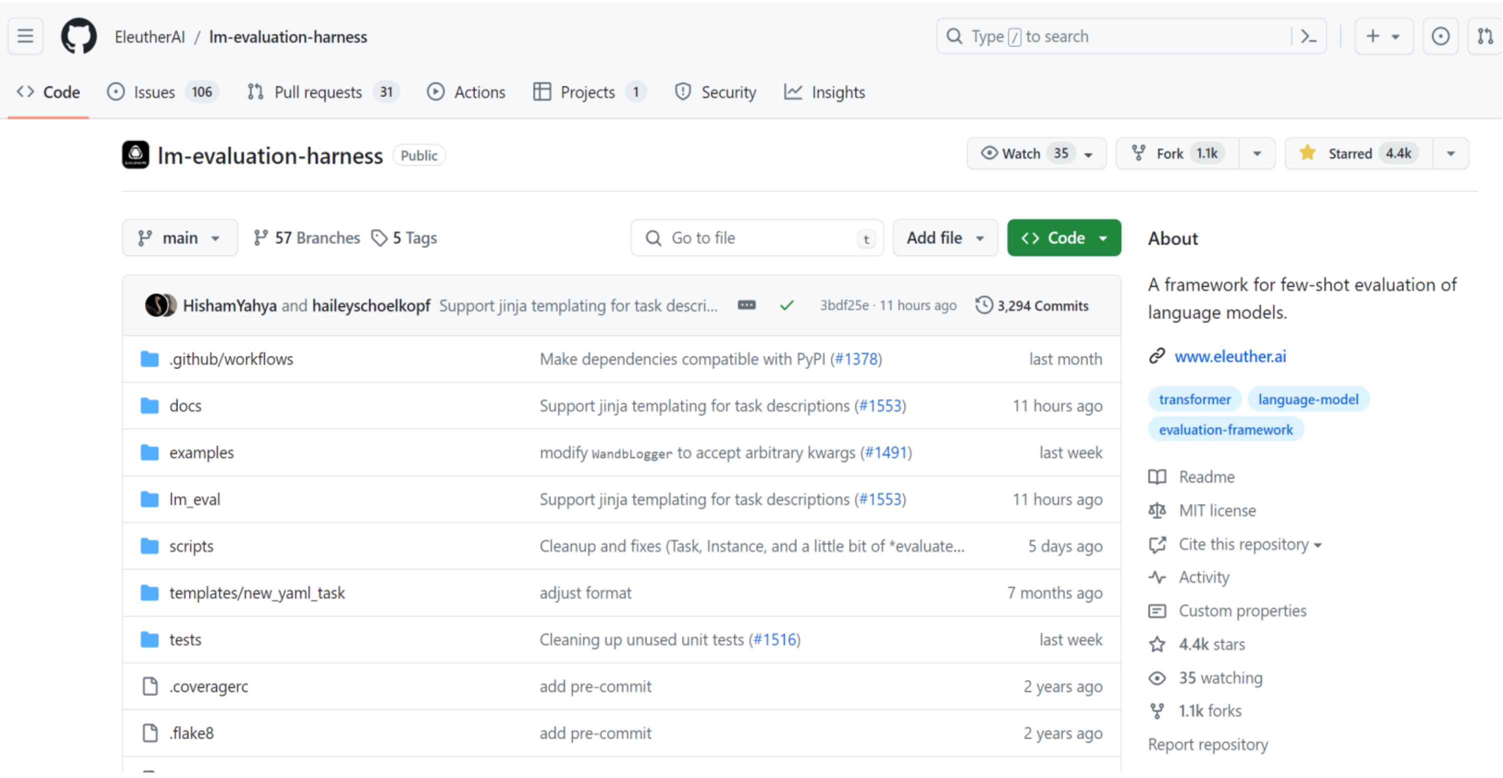
Table 4: Overview of the Korean LLMs (The model is from `https://huggingface.co`)

Three Experimental Objectives

1. Compare models with and without an expanded Korean vocabulary (Impact of Vocabulary Expansion).
2. Compare models pre-trained with Korean-English Bilingual data to those pre-trained with Korean Monolingual data (Impact of Bilingual Pretraining).
3. Performance changes according to Instruction Tuning using the proposed Bilingual LIMA dataset (Impact of Korean-English LIMA SFT).

03 Analysis Results (Quantitative)

Model comparison and reproducibility are ensured using the public branch of Polyglot team of EleutherAI's lm-evaluation-harness.



HishamYahya and haileyschoelkopf Support jinja templating for task descri... 3bdf25e · 11 hours ago 3,294 Commits

.github/workflows	Make dependencies compatible with PyPI (#1378)	last month
docs	Support jinja templating for task descriptions (#1553)	11 hours ago
examples	modify `WandbLogger` to accept arbitrary kwargs (#1491)	last week
lm_eval	Support jinja templating for task descriptions (#1553)	11 hours ago
scripts	Cleanup and fixes (Task, Instance, and a little bit of *evaluate...	5 days ago
templates/new_yaml_task	adjust format	7 months ago
tests	Cleaning up unused unit tests (#1516)	last week
.coveragerc	add pre-commit	2 years ago
.flake8	add pre-commit	2 years ago

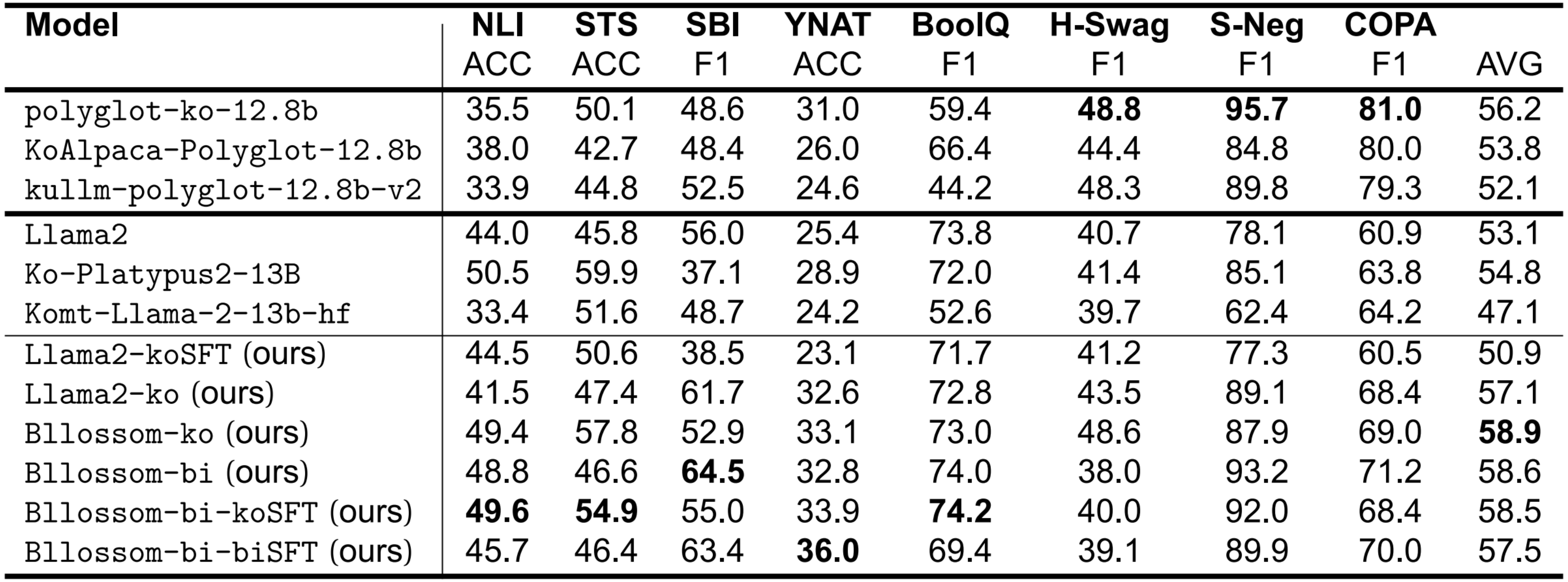
A framework for few-shot evaluation of language models.

www.eleuther.ai

Evaluation Benchmark Datasets

NLI	A classification dataset for predicting the relationship between two sentences.
STS	A classification dataset for measuring semantic equivalence between two sentences.
YNAT	A classification dataset for inferring the topic of a given sentence.
SBI	A classification dataset aimed at identifying social stereotypes or biases.
BoolQ	A QA dataset for answering yes/no questions.
HellaSwag	A commonsense NLI dataset.
SentiNeg	Emotion classification data.
COPA	A classification dataset for determining cause/effect based on paragraphs.

03 Analysis Results (Quantitative)



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

03 Analysis Results (Quantitative)

(Overall) Compared to single-language models, the multilingual Blossom model showed an average performance improvement of 4.57 points.

(Impact of Vocabulary Expansion)

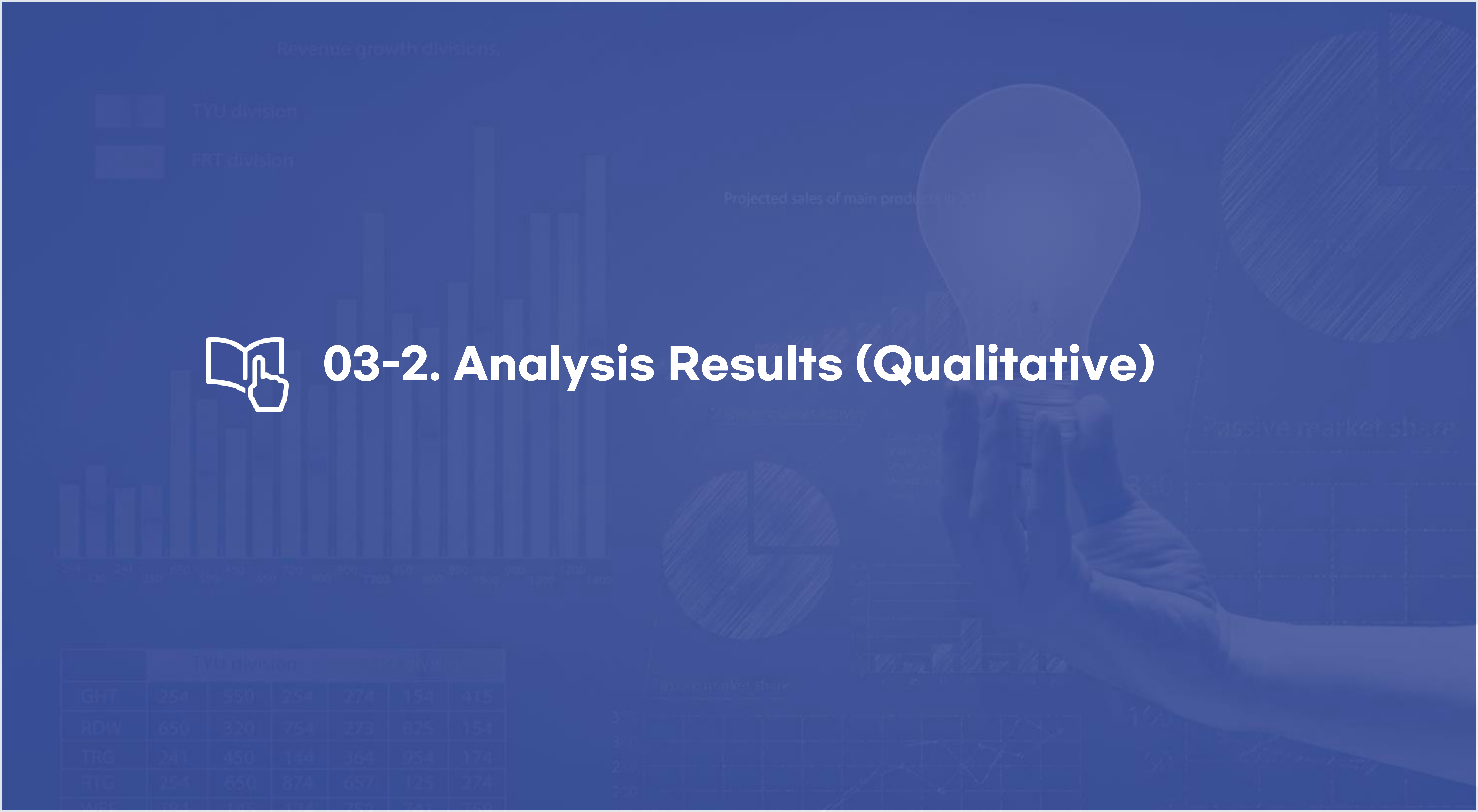
- Vocabulary expansion improved overall Korean comprehension, reasoning, perception, and causal understanding.

(Impact of Bilingual Pretraining)

- No significant difference was observed in quantitative assessments, but a notable difference was evident in qualitative evaluations.

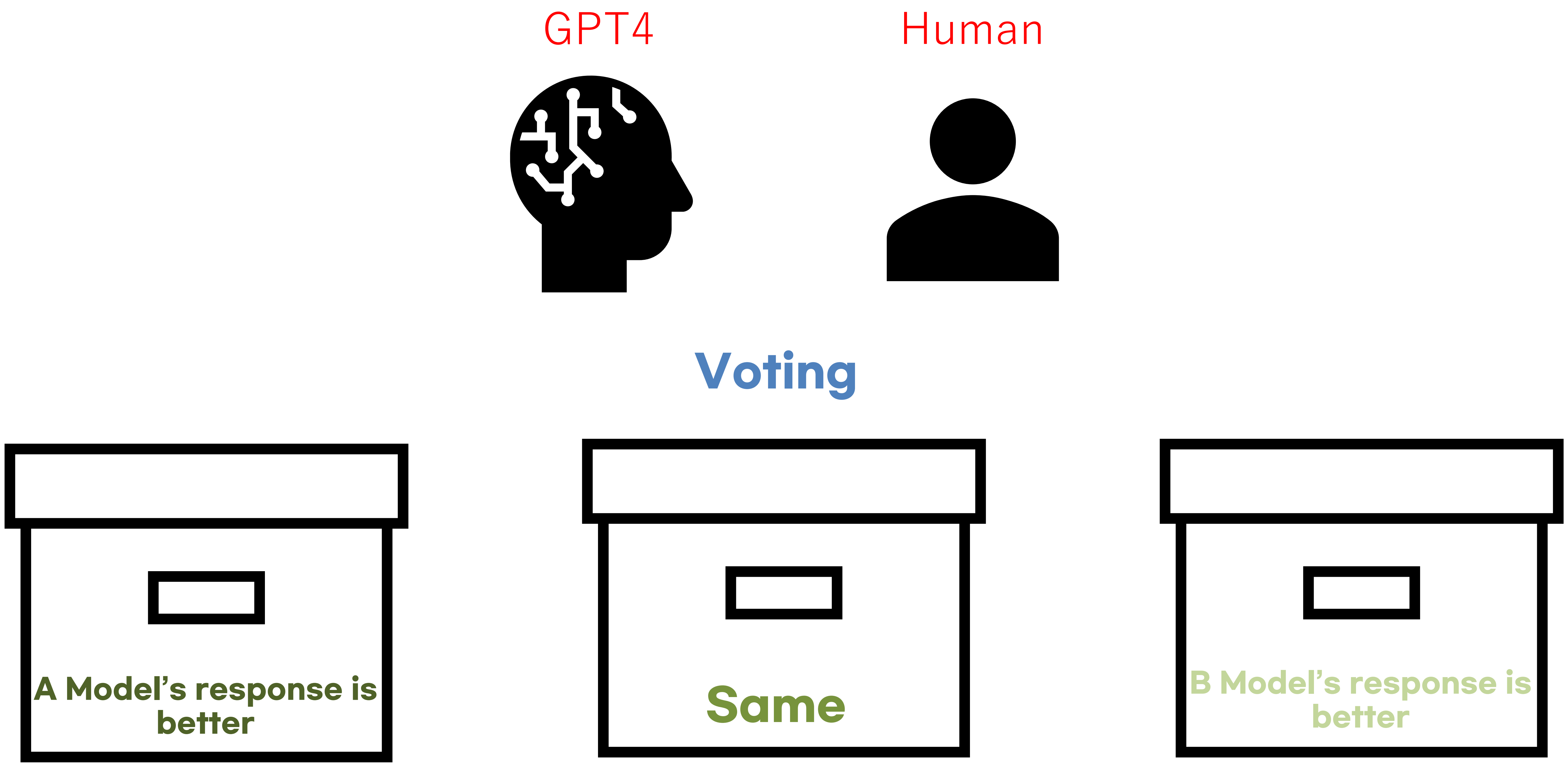
(Impact of Korean-English LIMA SFT)

- SFT did not significantly affect the quantitative assessment of classification tasks.
- However, based on qualitative factors such as response quality, vocabulary, and completeness, it generated better answers than Llama2.



03-2. Analysis Results (Qualitative)

03-2 Analysis Results (Qulitative)



GPT4

Human

Voting

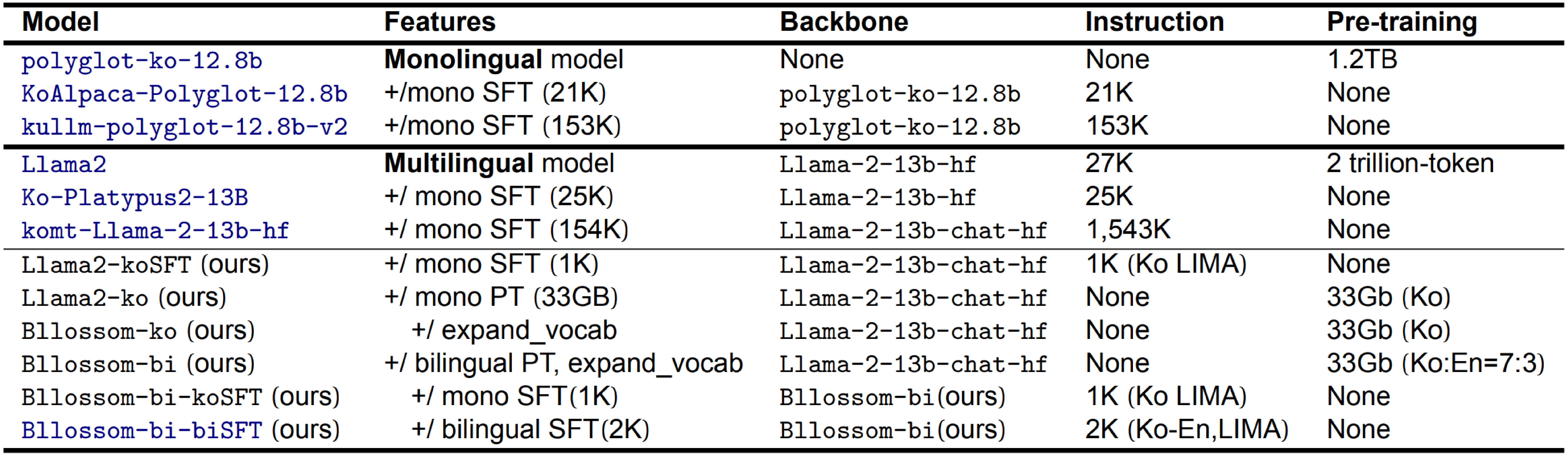
A Model’s response is better

Same

B Model’s response is better

For the evaluation, participants were asked to choose one of three options for each of the 300 Korean LIMA dataset questions, which included completely different topics, styles, and tasks.

03-2 Analysis Results (Qulitative)



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`)

03-2 Analysis Results (Qulitative)

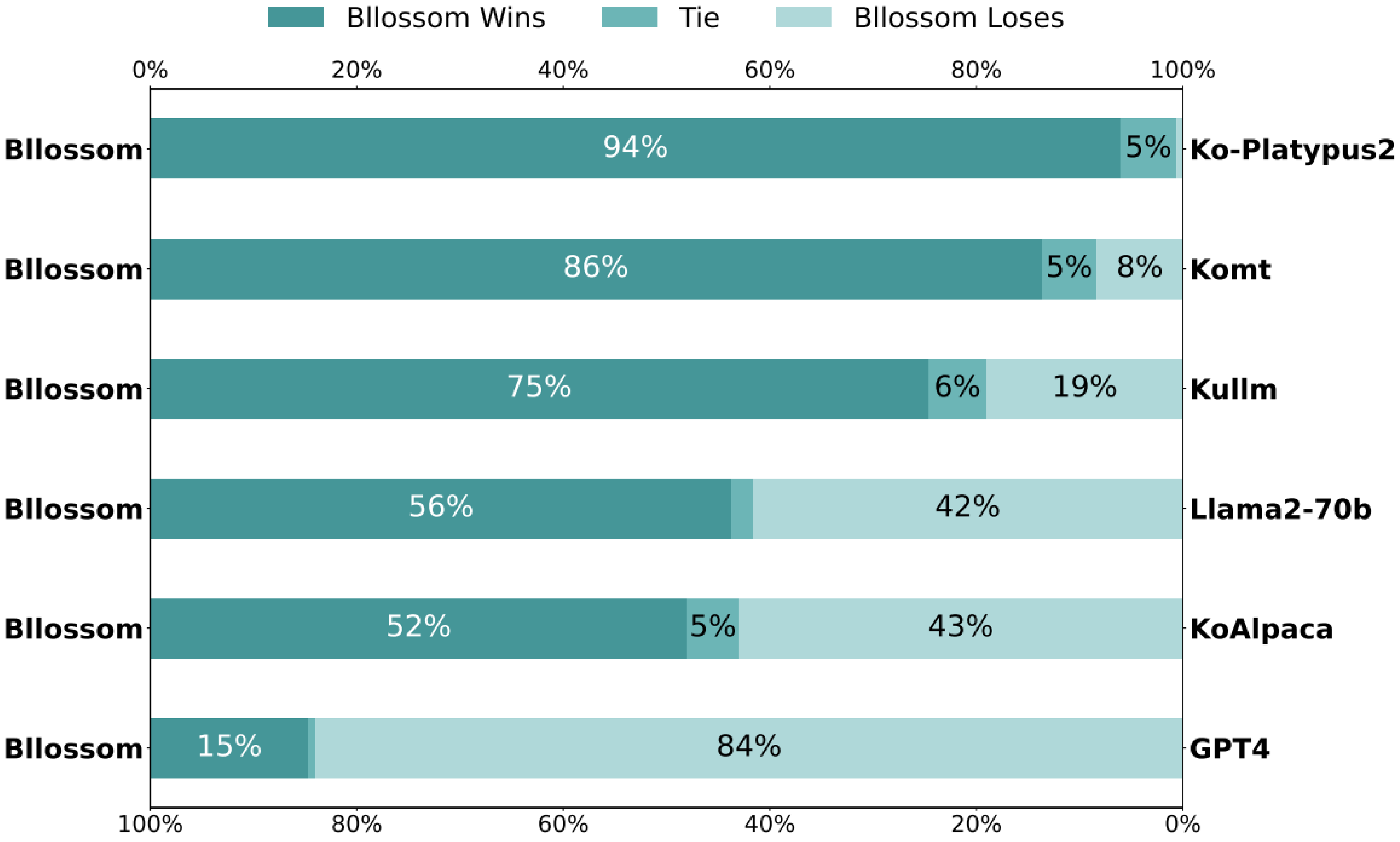


Figure 3: Preference evaluation results by GPT4

03-2 Analysis Results (Qulitative)

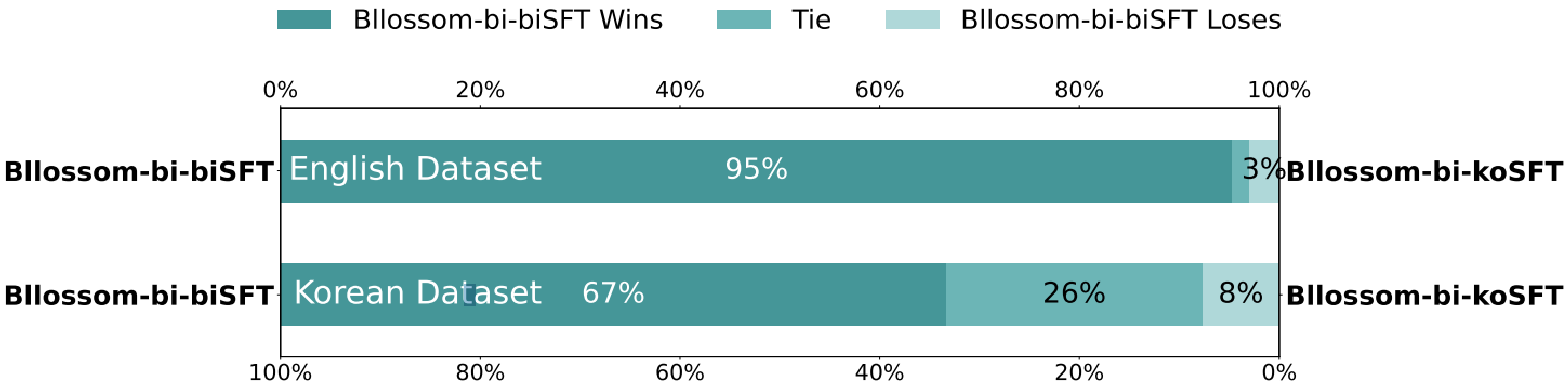


Figure 4: Comparing **bi**SFT and **ko**SFT models

03-2 Analysis Results (Qulitative)

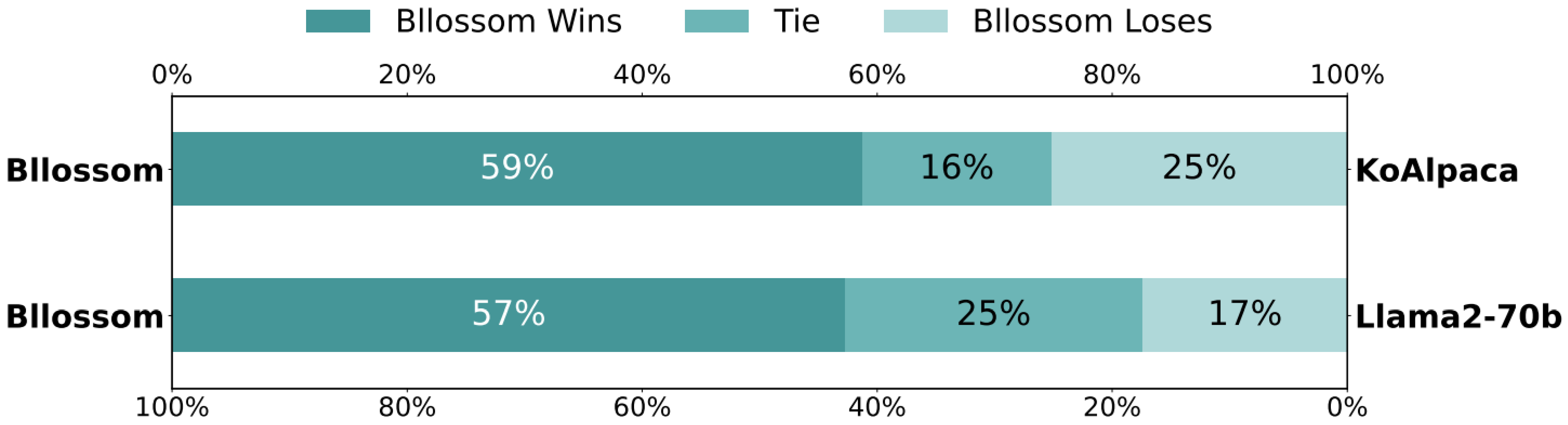


Figure 2: Preference evaluation results by human

03-2 Analysis Results (Qulitative)

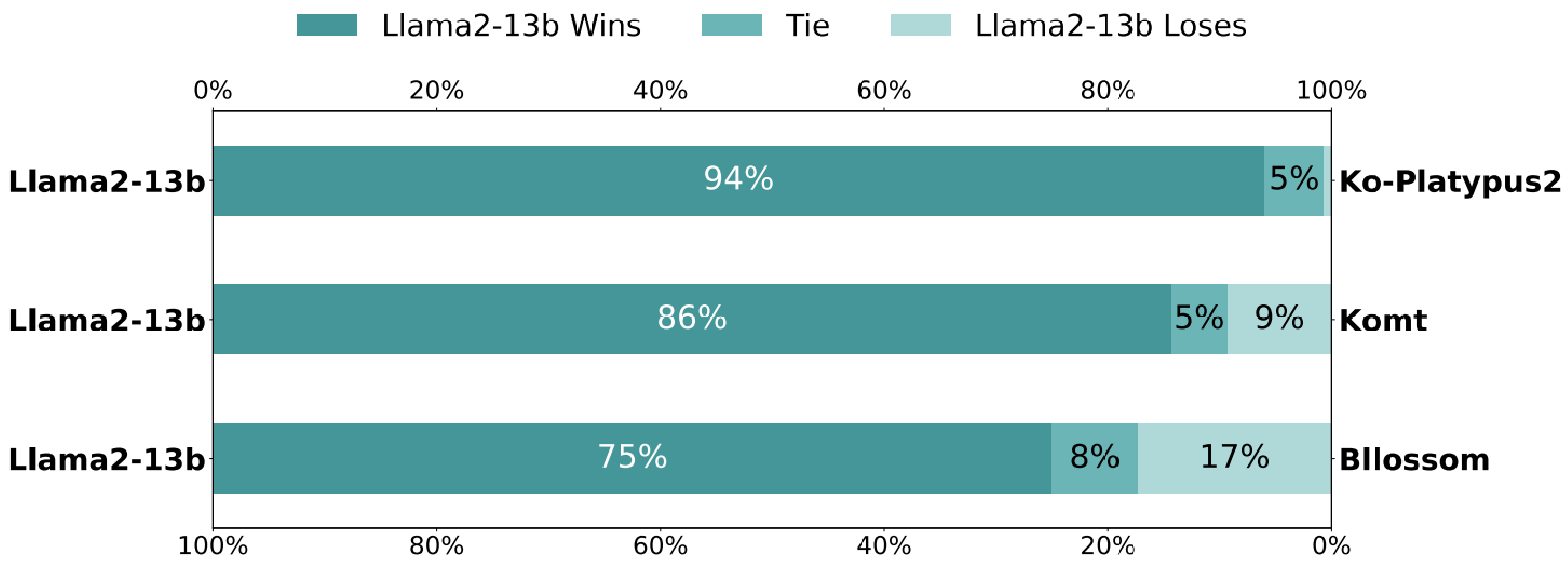


Figure 5: Comparing English performance in Llama2-backboned models

03-2 Analysis Results (Qulitative)

(Overall) Human evaluations and qualitative evaluation with GPT-4 show similar results, with Blossom outperforming models of similar size, and even surpassing the larger Llama2-70b-chat model.

(Comparison between Korean models based on Llama2 and Bllossom)

- **Pretraining has shown to significantly impact Korean proficiency.**

(Comparison between Bllossom and Monolingual Foundation Model-based models)

Bilingual Pretraining performed to compensate for poor performance in Korean significantly helps bridge the knowledge gap between Korean and English.

(Comparison of Bllossom with GPT-4 and Llama2)

- When there is an extreme difference in model size, techniques such as vocabulary expansion and pretraining can significantly mitigate performance disparities.

(Impact on English Proficiency)

- Training with bilingual data not only improves Korean proficiency but also results in less decline in English proficiency.

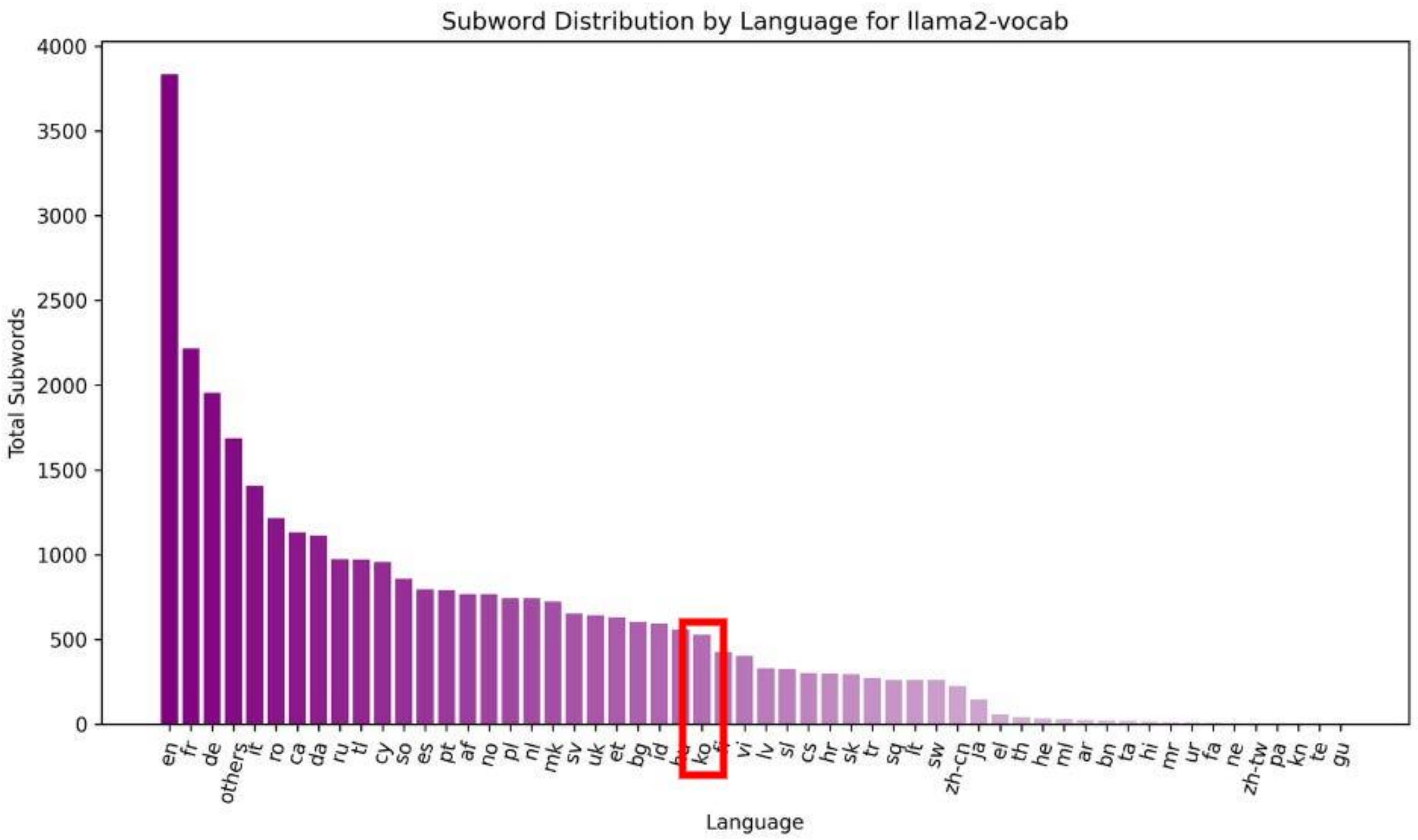
04. Discussion and Implications

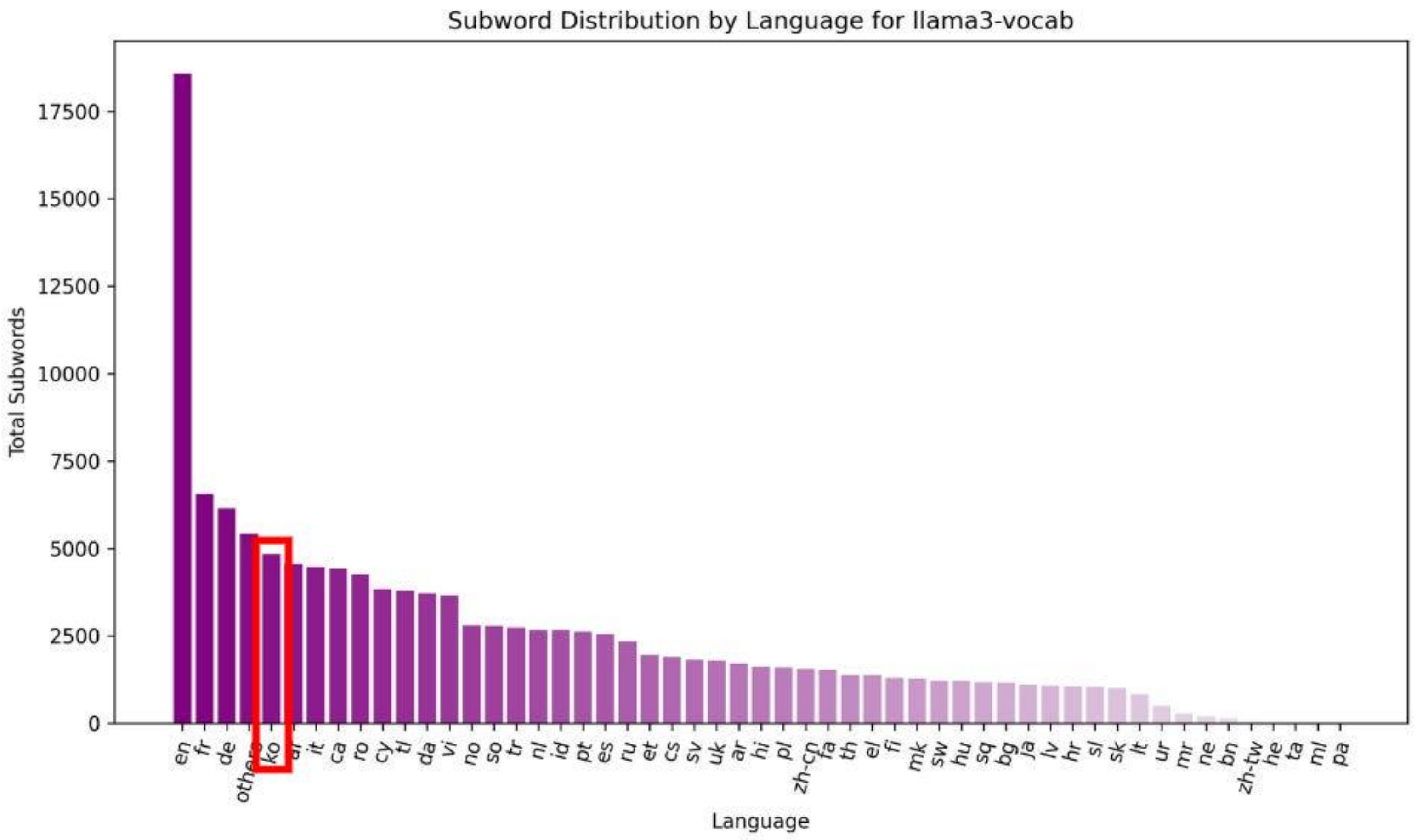
04 Discussion and Implications

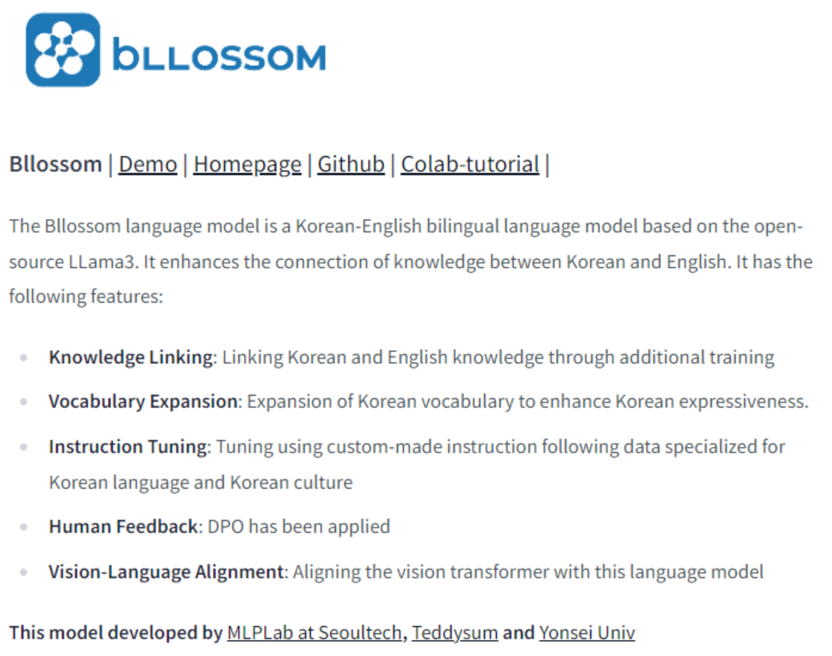
To apply the methodology used in this study to other languages, the following conditions are necessary:

- **To construct the LIMA dataset, 1,030 data points must be translated and post-processed.**
- **For testing, 300 data points must be translated and post-processed.**

04 Discussion and Implications





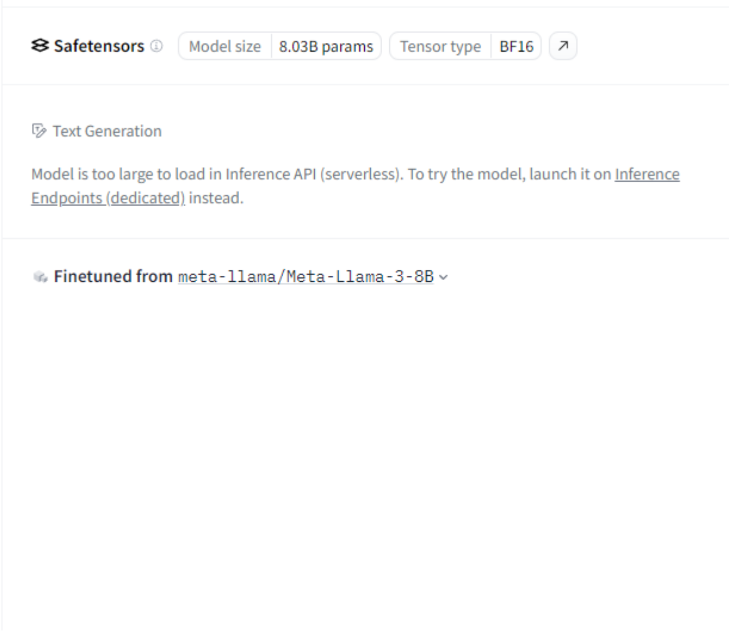


Bllossom | Demo | Homepage | Github | Colab-tutorial |

The Bllossom language model is a Korean-English bilingual language model based on the open-source LLama3. It enhances the connection of knowledge between Korean and English. It has the following features:

- **Knowledge Linking**: Linking Korean and English knowledge through additional training
- **Vocabulary Expansion**: Expansion of Korean vocabulary to enhance Korean expressiveness.
- **Instruction Tuning**: Tuning using custom-made instruction following data specialized for Korean language and Korean culture
- **Human Feedback**: DPO has been applied
- **Vision-Language Alignment**: Aligning the vision transformer with this language model

This model developed by MLPLab at Seoultech, Teddysum **and** Yonsei Univ

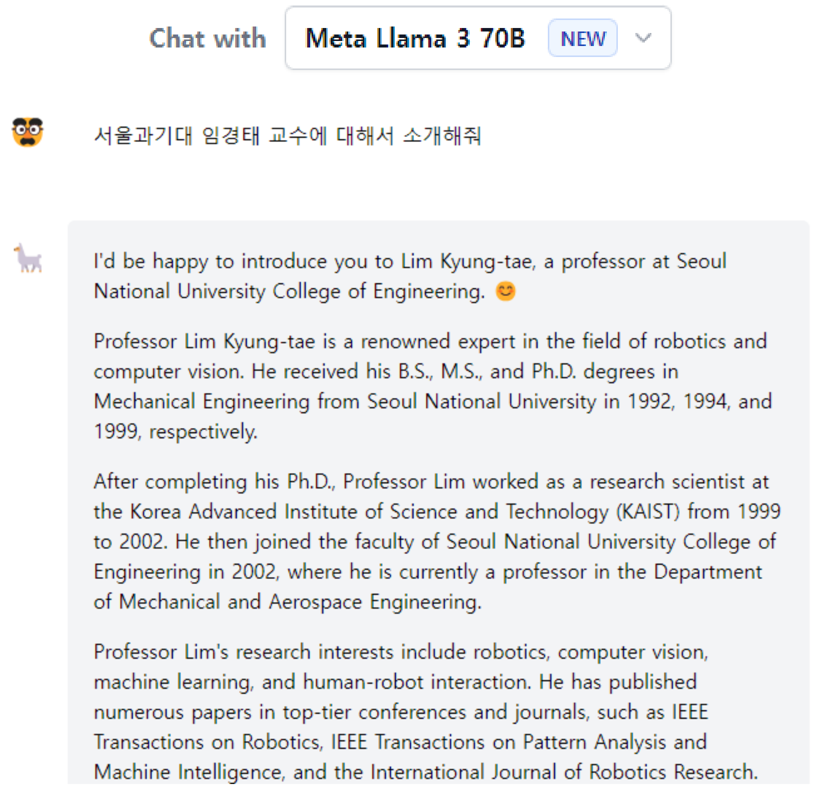


Safetensors | Model size | 8.03B params | Tensor type | BF16

Text Generation

Model is too large to load in Inference API (serverless). To try the model, launch it on Inference Endpoints (dedicated) instead.

Finetuned from `meta-llama/Meta-Llama-3-8B`



Chat with **Meta Llama 3 70B** NEW

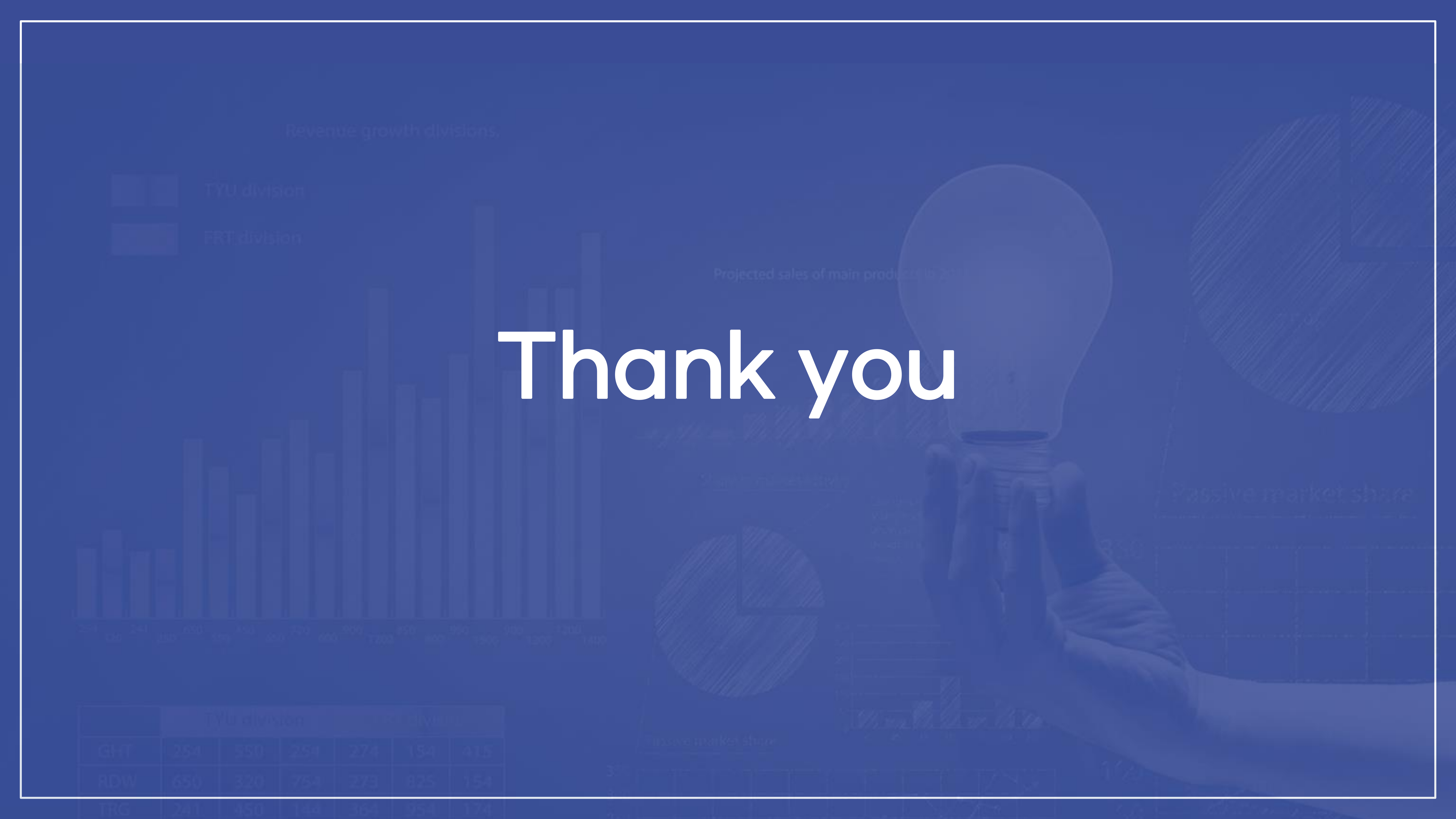
서울과기대 임경태 교수에 대해서 소개해줘

I'd be happy to introduce you to Lim Kyung-tae, a professor at Seoul National University College of Engineering. 😊

Professor Lim Kyung-tae is a renowned expert in the field of robotics and computer vision. He received his B.S., M.S., and Ph.D. degrees in Mechanical Engineering from Seoul National University in 1992, 1994, and 1999, respectively.

After completing his Ph.D., Professor Lim worked as a research scientist at the Korea Advanced Institute of Science and Technology (KAIST) from 1999 to 2002. He then joined the faculty of Seoul National University College of Engineering in 2002, where he is currently a professor in the Department of Mechanical and Aerospace Engineering.

Professor Lim's research interests include robotics, computer vision, machine learning, and human-robot interaction. He has published numerous papers in top-tier conferences and journals, such as IEEE Transactions on Robotics, IEEE Transactions on Pattern Analysis and Machine Intelligence, and the International Journal of Robotics Research.


Revenue growth divisions
TYU division
FRT division
Projected sales of main products
Passive market share

Thank you