

Investigating linguistic bias in large language models: A case study of Korean text generation

Narae Park

Informatics: Language Technology
60 ECTS study points

Department of Informatics
Faculty of Mathematics and Natural Sciences

Narae Park

Investigating linguistic bias in large
language models: A case study of
Korean text generation

Supervisors:
Andrey Kutuzov
David Samuel

Abstract

This study explores the linguistic biases in language models trained on English-centric data when applied to non-English languages, using Korean text generation as a case study. By prompting various types of language models (Korean monolingual models, multilingual models, Korean continually pre-trained models, and models not trained on Korean) to generate text using Korean input prompts, the study investigates the differences in the generated texts and the differences in the Korean sentences filtered from the generated texts, from a linguistic perspective. The generated texts and Korean sentences are evaluated across multiple dimensions, such as surface-level, lexical, syntactic, semantic and English translationese aspects. The influences of input prompts, model size, and pre-training data on the generation are also analyzed.

The results demonstrate that Korean monolingual models and Korean continually pre-trained models generate more linguistically appropriate Korean texts compared to multilingual models across various linguistic aspects. Models not trained on Korean struggled to generate Korean in the zero-shot setting but exhibited rapid learning ability for Korean in the few-shot setting. Furthermore, the results suggest that learning the formal aspects of a language can be achieved with smaller-sized models or limited training data, while learning the semantic aspects or reasoning with them requires larger models and extensive pre-training data. The results also show that language models trained on English-centric data can generate good-quality Korean sentences when further trained with sufficient, high-quality Korean data, providing a practical strategy for applying large language models to Korean tasks without serious concerns about linguistic bias. The methodology employed in this study can be extended to other languages and contribute to deepening the understanding of linguistic biases in large language models.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research Questions	2
1.3	Contributions	3
1.4	Structure	4
2	Background	5
2.1	A Brief history of language models	5
2.2	Overview of various MLLMs	8
2.3	Benefits of MLLMs	13
2.3.1	Cross-Lingual Transfer	13
2.3.2	Capturing relationships between Languages	16
2.4	Issues with MLLMs	17
2.4.1	Performance gap	17
2.4.2	Biases	19
3	Methods	22
3.1	Models	22
3.1.1	Korean Monolingual Models	22
3.1.2	Multilingual Models	23
3.1.3	Korean Continually Pre-trained Models	24
3.1.4	Non-Korean-trained Models	25
3.2	Tasks	26
3.2.1	Task 1: Text generation without reasoning (“zero-shot”)	27
3.2.2	Task 2: Text generation with reasoning (“few-shot”)	28
3.3	Dataset	29
3.4	Hyperparameter search	30
3.4.1	Hyperparameter configurations	30
3.4.2	Text generation for hyperparameter search	31
3.4.3	Evaluation of the generated text quality	31
3.5	Text Generation	35
3.6	Analysis: Generated texts	36
3.6.1	Basic Statistics	37
3.6.2	Surface-level Evaluation	37
3.6.3	Semantic Evaluation	39
3.7	Analysis: Generated Korean sentences	40
3.7.1	Basic Statistics	40
3.7.2	Lexical Evaluation	41
3.7.3	Syntactic Evaluation	41
3.7.4	English Translationese	42
3.7.5	Semantic Evaluation	44

4	Results	45
4.1	Analysis for generated texts	45
4.1.1	Basic Statistics	45
4.1.2	Surface-level Evaluation	53
4.1.3	Semantic Evaluation	59
4.2	Analysis for generated Korean sentences	64
4.2.1	Basic Statistics	64
4.2.2	Lexical Evaluation	66
4.2.3	Syntactic Evaluation	68
4.2.4	English translationese	75
4.2.5	Semantic Evaluation	82
5	Discussion	85
5.1	Answers to Research Questions	85
5.1.1	Differences in generated texts	85
5.1.2	Differences in generated Korean sentences	87
5.2	Further Findings	89
5.2.1	Influence of Few-Shot Prompts	89
5.2.2	Influence of Model Size	90
5.2.3	Influence of Pre-training Data	91
5.3	Limitations & Future Work	94
5.3.1	Limitations in Evaluation Methodology	94
5.3.2	Limitations in Scope of the Study	94
6	Conclusion	96
A	Appendix	99

List of Figures

2.1	Transformer - model architecture. Image taken from Vaswani et al., 2017.	8
3.1	Top 5 hyperparameter configurations for each model based on semantic similarity scores of the texts generated in hyperparameter search.	33
4.1	Average length of texts generated by language models. The four segments in the plot represent Korean monolingual models, multilingual models, Korean continually pre-trained models, and non-Korean-trained models, respectively. Blue indicates results in the zero-shot setting, while orange indicates results in the few-shot setting. The height of the bars represents the mean value, and the vertical lines denote the standard deviation. The horizontal lines in each segment represent the mean of each language model group, with the numeric value on the line indicating the mean and standard deviation (in parentheses).	46
4.2	Average number of tokens generated by language models.	46
4.3	Character type distribution of the texts generated language models in the zero-shot task. The proportion of each character type is represented by a different color in the stacked bar graph. Note that the total ratio for each model does not equal, 1 because only the major types are shown.	47
4.4	Character type distribution of the texts generated language models in the few-shot task.	48
4.5	Token type distribution of the texts generated by each model in the zero-shot task.	49
4.6	Token type distribution of the texts generated by each model in the few-shot task.	49
4.7	Proportion of the number of sentences to the total length of the texts generated by language models	51
4.8	Average text length per sentences generated by language models.	52
4.9	Average number of words per sentence generated by language models.	52
4.10	Distribution of Korean character ratio per sentence in the texts generated by the language models. It is represented using box plots. The box shows the range of the middle 50% of the data (from Q1 (25th percentile) to Q3 (75th percentile)), and the horizontal line inside the box represents the median of the data. The whiskers extending from the top and bottom of the box indicate the typical range of the data (from $Q1 - 1.5 \times (Q3 - Q1)$ to $Q3 + 1.5 \times (Q3 - Q1)$). Points outside the whiskers are marked as outliers.	53
4.11	Proportion of sentences containing unusual patterns to the total sentences generated by language models.	54
4.12	Proportion of sentences containing the headline pattern "<...>" provided in examples to the total sentences generated by language models.	55

4.13	Proportion of sentences containing the headline pattern “<” or “>” provided in examples to the total sentences generated by language models.	56
4.14	Proportion of the number of sentences that do <i>not</i> have typical Korean sentence-ending formats to the total number of generated sentences.	57
4.15	Proportion of the number of formally complete Korean sentences to the total number of sentences generated by language models.	58
4.16	Cosine similarity scores of semantic embeddings between the original texts and the generated texts by language models.	60
4.17	Visualization using t-SNE of semantic embedding distribution for the original texts and the texts generated by language models in the zero-shot task. . .	62
4.18	Visualization using t-SNE of semantic embedding distribution for the original texts and the texts generated by language models in the few-shot task. . .	63
4.19	Number of Korean sentences generated by language models.	65
4.20	Average text length of Korean sentences generated by language models. . .	65
4.21	Average number of words per sentence generated by language models. . . .	66
4.22	Root Type-Token Ratio (RTTR) of Korean corpora and Korean sentences generated by language models.	67
4.23	Root Type-Token Ratio (RTTR) of Korean corpora and Korean sentences generated by language models, considering only tokens included in the Korean dictionary entry morpheme set.	67
4.24	UPOS distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.	68
4.25	UPOS distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.	69
4.26	XPOS distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.	70
4.27	XPOS distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.	70
4.28	XPOS distribution (separated tag) in Korean corpora and Korean sentences generated by language models in the zero-shot task.	71
4.29	XPOS distribution (separated tag) in Korean corpora and Korean sentences generated by language models in the few-shot task.	71
4.30	Dependency relation distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.	72
4.31	Dependency relation distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.	72
4.32	Direction and lengths of the dependency arcs in Korean corpora and Korean sentences generated by language models in the zero-shot task.	74
4.33	Direction and lengths of the dependency arcs in Korean corpora and Korean sentences generated by language models in the few-shot task.	74
4.34	Proportion of the word ‘han(한)’ as a translationese of ‘a’ to the nouns in Korean corpora and Korean sentences generated by language models.	75
4.35	Proportion of the word ‘geu(그)’ as a translationese of ‘the’ to the nouns in Korean corpora and Korean sentences generated by language models. . . .	76
4.36	Proportion of the word ‘geu(그)’ as a translationese of ‘the’ to the nouns in Korean corpora and Korean sentences generated by language models, without the spoken language corpus.	76
4.37	Proportion of the word ‘deul(들)’ as a translationese of plurals to the nouns in Korean corpora and Korean sentences generated by language models. . . .	77

4.38	Proportion of the passive subjects in Korean corpora and Korean sentences generated by language models.	78
4.39	Proportion of sentences without subjects in Korean corpora and Korean sentences generated by language models.	79
4.40	Proportion of sentences without objects in Korean corpora and Korean sentences generated by language models.	80
4.41	Proportion of sentences without both subjects and objects in Korean corpora and Korean sentences generated by language models.	81
4.42	Proportion of Object-Verb order sentences to the sentences with object and verb in Korean corpora and Korean sentences generated by language models.	81
4.43	Proportion of Verb-Object order sentences to the sentences with object and verb in Korean corpora and Korean sentences generated by language models.	82
4.44	Distribution of sentiment classification in Korean corpora and Korean sentences generated by language models.	83
A.1	Input prompt provided to LLM as a judge in hyperparameter search.	99
A.2	Visualization using PCA of semantic embedding distribution for the original texts and the texts generated by language models.	110
A.3	Visualization using UMAP of semantic embedding distribution for the original texts and the texts generated by language models.	111

List of Tables

1.1	Proportion of English data in pre-training data for widely used large language models.	2
3.1	Summary of language models used in this study.	27
3.2	Examples of the input prompts for text generation tasks.	29
3.3	Composition of the dataset for text generation.	30
3.4	Hyperparameter configurations for Hyperparameter search.	31
3.5	Average semantic similarity scores of models for texts generated in hyperparameter search.	32
3.6	Hyperparameter configurations for each model derived from hyperparameter search.	36
4.1	Number of sentences from text generated by language models after sentence segmentation. Since the total text length generated differs across models, the proportion relative to the total text length (%) is also shown.	51
4.2	Number of sentences filtered at each step in general Korean sentence format filtering process. The proportion of the final filtered sentences relative to the initial total number of sentences is also shown.	58
4.3	Cosine similarity semantic embeddings between the original texts and the generated texts by language models, presented in descending order of similarity scores.	60
A.1	Statistics on the length of texts generated by language models.	100
A.2	Statistics on the number of tokens generated by language models.	101
A.3	Character type distribution in texts generated by language models in the zero-shot task.	102
A.4	Character type distribution in texts generated by language models in the few-shot task.	103
A.5	Token type distribution in texts generated by language models in the zero-shot task.	104
A.6	Token type distribution in texts generated by language models in the few-shot task.	104
A.7	Average text length and average number of words per sentence generated by language models.	105
A.8	Total sentences, sentences with unusual patterns, and their ratio in texts generated by language models.	106
A.9	Total sentences, sentences with the headline pattern "<...>", and their ratio in texts generated by language models.	107

A.10	Total sentences, sentences with the headline pattern “<” or “>”, and their ratio in texts generated by language models.	108
A.11	Total sentences, sentences with formally complete endings, and their ratio in texts generated by language models.	109
A.12	Number of Korean sentences in the Korean corpora and Korean sentences generated by language models.	112
A.13	Average text length and average number of words per Korean sentences in the Korean corpora and Korean sentences generated by language models.	113
A.14	Lexical diversity metrics of Korean corpora and Korean sentences generated by language models.	114
A.15	Lexical diversity metrics of Korean corpora and Korean sentences generated by language models, considering only tokens included in the Korean dictionary entry morpheme set.	114
A.16	UPOS distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.	115
A.17	UPOS distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.	115
A.18	Universal Part-of-Speech tags and descriptions.	116
A.19	XPOS distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.	117
A.20	XPOS distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.	118
A.21	Korean Part-of-Speech tags and descriptions.	119
A.22	XPOS distribution (separated tag) in Korean corpora and Korean sentences generated by language models in the zero-shot task.	120
A.23	XPOS distribution (separated tag) in Korean corpora and Korean sentences generated by language models in the few-shot task.	121
A.24	Dependency relation distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.	122
A.25	Dependency relation distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.	123
A.26	Universal dependency relations tags and descriptions.	124
A.27	Dependency arc direction and lengths in Korean corpora and Korean sentences generated by language models in the zero-shot task.	125
A.28	Dependency arc direction and lengths in Korean corpora and Korean sentences generated by language models in the few-shot task.	126
A.29	Detection of English translationese artifacts in Korean corpora and Korean sentences generated by language models in the zero-shot task.	127
A.30	Detection of English translationese artifacts in Korean corpora and Korean sentences generated by language models in the few-shot task.	128
A.31	Sentiment classification distribution in Korean corpora and Korean sentences generated by language models.	129

Preface

I would like to express my gratitude to the many people who have helped me complete this research.

First and foremost, I must mention Andrey and David, the supervisors of this study. Andrey boldly encouraged me to conduct research on the Korean language and always inspired me with his open-minded attitude towards new ideas. The resources he provided and his incisive feedback were of great help to the research. David, with his smartness behind a friendly smile, offered a lot of invaluable advice in carrying out the research. Whenever I reached out to him, he always welcomed me, and no matter what questions I brought to him, he always had answers. Every time I left a meeting with them, I felt that I had learned something new, and I truly enjoyed every meeting with them.

I also extend my thanks to Nikolay, who often lent an ear to my thesis progress and provided precious advice in the LTG kitchen late at night. Thanks to Petter, I could stay more comfortably in the LTG space. I am grateful for the many warm conversations and meals he shared with me. To Jakoba, I am thankful for the kind advice and encouragement she offered whenever I met her. Her thoughtfulness allowed me to focus on my thesis in the final stages.

I also send my gratitude to those I met in the LTG space and to those who shared the master's room with me, whether they have already graduated, are graduating now, or will graduate in the future. I am thankful to my friends in Korea and Norway who believed in me and encouraged me even when I doubted myself, and to my family, who have always been there for me with unwavering support.

I am grateful to the countless individuals who share their knowledge on the web. I also received much help with translations and coding from LLMs, which have evolved as an aggregation of that collective knowledge. I hope that this research can contribute, even in a small way, to making LLMs better serve all the people from whom they have originated.

* The code used in this study is available on GitHub. ¹

* The cover image was generated by DALL-E 3.

¹https://github.com/n4r4e/uio_ms_thesis

Chapter 1

Introduction

1.1 Motivation

Over the past few years, research on language models has been highly active (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020, etc.) and the performance of language models has significantly improved, leading to the expansion in the use of language models worldwide. Additionally, as the size of language models and the volume of training data have increased (Brown et al., 2020; Kaplan et al., 2020), and as the cross-lingual transfer ability in language models is discussed (Artetxe and Schwenk, 2019; Pires et al., 2019), interest in language models supporting various languages has also grown substantially (Lin et al., 2022; Shliazhko et al., 2022; Scao et al., 2023). The development of language models that support various languages has several benefits. Multilingual language models make language models available in many different languages, increasing the accessibility of language models to people around the world and allowing more people to benefit from the use of language models. They also can reduce the computational resources and time required to develop and maintain separate models for each language, thereby increasing overall efficiency.

However, are these language models sufficiently and appropriately multilingual? The common usage of language models today often involves using large-scale language models as base models and fine-tuning them with specific datasets suitable for downstream tasks, or eliciting appropriate answers from them through prompting without additional fine-tuning. The large-scale language models that serve as the base models require very large training data, and these training data are generally heavily skewed towards English, as shown in Table 1.1. This language bias in training data of language models could bring potential concerns or issues in the multilingualism of language models and the multilingual downstream tasks based on these models. Some of recent studies are actively investigating these concerns, from various perspectives: For example, large language models may perform better in English (Huang et al., 2023), their responses might carry grammatical nuances of English (Papadimitriou et al., 2023), they culturally reflect values of English-speaking countries (Masoud et al., 2023; Naous et al., 2023), or they “think” in English as a pivoting language (Etxaniz et al., 2023; Zhang et al., 2023; Wendler et al., 2024). Such English bias in widely used language models can potentially

Language Model	English data in pre-training data
LLaMA 2 (Touvron, Martin, et al., 2023)	89.70%
Claude 2 (Anthropic, 2023)	Roughly 90%
PaLM (Chowdhery et al., 2022)	77.98%
LaMDA (Thoppilan et al., 2022)	Over 90%
GPT-3 (Brown et al., 2020)	93%

Table 1.1: Proportion of English data in pre-training data for widely used large language models.

and gradually affect the diversity of languages, the diversity of cultures behind them, and further, the diversity of thought. Therefore, as the use of large language models is becoming more widespread, understanding the language bias of these language models is an important and timely research topic.

1.2 Research Questions

The core question driving this research is as follows:

“Are there any potential issues, concerns, or side effects when using language models trained on English-centric data to perform tasks in non-English languages?”

To investigate this, the study will explore linguistic biases in language models using Korean as an individual non-English language. This decision is primarily based on the constraints of time and resources, and the fact that Korean is the author’s native language, which provides the author with the linguistic and cultural understanding for in-depth exploration. However, since Korean has many differences from English, this study is not only limited to Korean but also aims to provide a extensive understanding and implications for the application of language models to languages significantly different from English.

Furthermore, the study focuses on text generation as the task at hand. Text generation is a comprehensive task that encompasses the language understanding and generative capabilities of language models. Although evaluating text generation tasks is challenging since there is no one right answer, the generated text provides rich information that help understand the operation of language models, which is why this study aims to closely examine and analyze the results of text generation.

For the language models to be explored in this study, we focus on pre-trained models that are commonly used as base models in various language applications. This is to investigate the biases in the base models without the influence of factors such as the dataset or training tasks used in fine-tuning. The study covers a diverse range of language model types, including:

1. Korean monolingual models trained on Korean data
2. Multilingual models trained on data consisting of various languages, including Korean

3. Models that are not explicitly trained on Korean, such as models trained primarily on English or English and Chinese data
4. Models that are further trained on Korean data using the aforementioned models as base models

These models vary in size and have different compositions of Korean training data, enabling analysis from multiple angles.

In this context, the research questions that further specify the core question and the experimental designs for them are as follows:

RQ1. When various types of language models are prompted to generate text in Korean, what texts do the language models generate?

Experimental Design: Perform Korean text generation tasks using various types of language models using Korean input prompts, and examine the results.

RQ2. Are there differences in the Korean sentences generated by various types of language models?

Experimental Design: Compare and analyze the Korean sentences generated by various types of language models

The study investigates these research questions by varying several factors that could influence the outcomes:

1. **Type of text generation task:** Text generation task that do not require reasoning (“zero-shot”) and text generation task that require reasoning (“few-shot”).
2. **Size of the language model:** Different sizes of language models within the same model family.
3. **Size/ratio of Korean training data:** Different sizes of Korean training data on which the language models were trained.

Through these research questions and corresponding experiments, this study aims to comprehensively explore the linguistic biases when language models are applied to individual languages.

1.3 Contributions

This study provides the following findings and contributions:

This study contributes to the understanding of language biases in various language models from the perspective of a specific language, Korean. By comparing and analyzing the results of generating Korean texts using various types of language models, the study provides empirical findings on the Korean text generation capabilities of different language models. The results demonstrate that Korean monolingual models and Korean continually pre-trained models generate more appropriate Korean texts compared to multilingual models across multiple linguistic dimensions.

Furthermore, the study offers insights into the language learning abilities of language models. The findings, which take into account factors such as model size, pre-training data, and input prompts, suggest that learning the formal aspects of a language can be achieved with small-sized models and limited training data, while learning the semantic aspects of a language or acquiring reasoning abilities requires relatively large-sized models and extensive training data.

Moreover, the results provide practical implications for using language models in Korean tasks. The findings show that even language models trained on English-centric data can generate good-quality Korean texts without serious concerns about language biases, if they are further trained with sufficient Korean data. This suggests that fine-tuning a base language model, which has been trained on large, high-quality data, with Korean data is a strategy worth considering in Korean applications.

The methods employed in this study to explore linguistic biases in language models can be extended and applied to other languages. The results presented in this research can also be analyzed in combination with the application results from other studies. The approach and insights from this study can serve as a starting point for future research investigating language biases across different languages.

1.4 Structure

The structure of this thesis is as follows.

Chapter 2 (Background) provides the background and context for this study through a review of previous research and related literature.

Chapter 3 (Methods) describes in detail the experimental design and the process of the experiments in this study.

Chapter 4 (Results) presents and analyzes the results of the experiments following the methods in Chapter 3.

Chapter 5 (Discussion) discusses the various findings derived from the experimental results and analyses in Chapter 4.

Chapter 6 (Conclusion) summarizes the findings of this study and concludes the research.

Chapter 2

Background

This section provides an extensive background on multilingual large language models (MLLMs) to facilitate a profound comprehension of this study. We begin by introducing the historical development of language models and then present various types of MLLMs based on the transformer-based architecture that many recent language models are built upon. After that, we discuss the benefits of MLLMs, as well as the potential issues in these models. Through this, we aim to provide a comprehensive landscape of MLLMs in the current trend of their rapid advancement, highlighting their unique and noteworthy characteristics as well as the potential issues that warrant attention.

2.1 A Brief history of language models

The emergence of multilingual large language models (MLLMs) is inseparable from the evolution of language models. A brief overview of the progression from language models to large language models, and their expansion into MLLMs, will provide a good starting point for understanding MLLMs.

A language model, as the name suggests, is a model of human language aiming to represent and reproduce human language. One way to represent and reproduce a language is to model the distribution of the language, capturing the probabilistic distribution of words and their sequences that appear in that language. A language model aims to probabilistically map out the sequential and interrelated distribution space of language. To achieve this, a language model computes and assigns probabilities to sequences of words. (Manning et al., 2009; Jurafsky and Martin, 2023)

In the 1910s, Markov (1913)¹ laid the mathematical foundation for computing the probability of sequences in a language, using the concept of Markov chain. In the 1940s-50s, Shannon (1948, 1951) expanded upon this, developing the n -gram Markov model targeting sequences of n consecutive characters. Based on these ideas, the probability of a sequence of words ' $w_1 w_2 \dots w_n$ ' can be expressed as a product of conditional probabilities using the chain rule, as shown in Equation 2.1.

¹For the English translation of this work, refer to Markov, 2006

$$P(w_1w_2\dots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\dots P(w_n|w_1w_2\dots w_{n-1}) = \prod_{k=1}^n P(w_k|w_1\dots w_{k-1}) \quad (2.1)$$

However, in the n -gram based approach, as the vocabulary size increases and sequences become longer, the model parameters grow exponentially, and more and more valid sequences may not appear in the training corpus at all, leading to the sparsity problem. Moreover, in the 1950s, Chomsky (1956, 1957) argued that the finite-state Markov process was insufficient as a comprehensive model for human language. This led to a period of stagnation in statistical language modeling until the 1980s (Nadas, 1984; Church and Gale, 1991), when the field was revived with the development of techniques such as smoothing, clustering, and caching that could mitigate the limitations of n -gram models (Goodman, 2001; H. Li, 2022; Jurafsky and Martin, 2023).

In the 2000s, with the application of word embeddings and neural network models, language model research underwent a significant transformation. Word embeddings refer to representations of words as numerical vectors, and are obtained by mapping words into a continuous vector space. Bengio et al. (2000) introduced a neural network model using distributed word embedding vectors, which addressed the issue of sparsity in traditional n -gram models. This shift also allowed for the relationships between words to be captured in their embeddings. Additionally, while the number of parameters in n -gram models increased exponentially based on the context length (n), neural network models using fixed-size distributed word vectors allowed the number of model parameters to increase linearly with the context length (the size of the input window), making the models more efficient. As advanced word embedding techniques such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017) were developed and architectures like long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) and gated recurrent units (GRU; Cho et al., 2014) were introduced to address the vanishing gradient problem in long-term backpropagation, which was a major issue in recurrent neural networks (RNNs; Elman, 1990), the performance of language modeling improved substantially. In 2017, with the introduction of the Transformer architecture (Vaswani et al., 2017) that adopted the attention mechanism, enabling parallel processing, the performance of language models advanced even further. Many of the language models in use today are based on this transformer architecture.

Neural network models require large amounts of training data, and obtaining this data is often costly and time-consuming. As an attempt to address this, transfer learning was introduced, leveraging already learned knowledge from pre-training for new tasks. By pre-training a language model with general data related to the target task and utilizing the knowledge gained from this, the model can handle the task well even with minimal additional fine-tuning on a limited number of samples. The knowledge to be transferred has evolved from pre-trained word embeddings (Turian et al., 2010) to contextualized embeddings (Peters et al., 2018), and as the scale of the models has grown, pre-trained language models have become capable of capturing lexicon, syntax, and semantics in a language. With the method of fine-tuning using pre-trained language models demonstrating high performance in downstream tasks, this approach has gradually become popular. BERT (Devlin et al., 2019), which is primarily designed

for natural language understanding, and GPT (Radford et al., 2018, 2019; Brown et al., 2020), which is mainly aimed at natural language generation, are well-known pre-trained language models.

During this process, it was observed that as the parameter size and training data size of the pre-trained model increased, the performance in downstream tasks also improved (Devlin et al., 2019; Brown et al., 2020). This led to a research trend of continuously increasing the size of the pre-trained language models, also known as the base models. While the early pre-trained model, BERT, was about 340M parameters (BERT-large) in size, recent language models have grown to as large as 1.2T (W. X. Zhao et al., 2023). Especially recently, the emergent capabilities of language models to significantly improve performance beyond a certain scale has been observed (Wei, Tay, et al., 2022), leading to increased interest in large language models. Language models characterized by such large parameter sizes and vast training data are commonly referred to as “Large Language Models” (LLMs).²

These pre-trained large language models are assumed to have knowledge about language. To apply them to downstream tasks, they generally require further adaptation to specific tasks. This is typically achieved through fine-tuning (Howard and Ruder, 2018; Devlin et al., 2019), where the model is additionally trained with a dataset relevant to the specific task. However, as the size of the model increases, fine-tuning the entire model becomes challenging and inefficient (Hao et al., 2019; Kovaleva et al., 2019), leading to research on more efficient fine-tuning techniques (Bapna and Firat, 2019; Houlsby et al., 2019; Pfeiffer et al., 2020; Radiya-Dixit and Wang, 2020; Pfeiffer et al., 2021; Ben Zaken et al., 2022) or using prompt-based tuning approaches (Radford et al., 2019; Brown et al., 2020; Lester et al., 2021; X. Liu et al., 2022), where the model leverages its pre-trained knowledge to perform the desired tasks without altering its weights. In addition, the development and deployment of instruction-fine-tuned models (Chung et al., 2022; Wei, Bosma, et al., 2022), which are further trained with instruction datasets to facilitate prompt-based approaches, have also increased.

Multilingual large language models have emerged as an extension of the development of these large language models. The pre-training of multilingual language models is not fundamentally different from that of monolingual language models. It uses the same model architecture as monolingual models but employs multilingual data composed of various languages for pre-training, and the training tasks are also slightly modified accordingly. For example, mBERT, a well-known multilingual large language model, is trained on Wikipedia dump data from the top 100 languages using the BERT model. mBART (Y. Liu et al., 2020) is trained on a dataset of 25 languages using the BART (Lewis et al., 2020) model, and mGPT (Shliazhko et al., 2022) is trained on a 30-language dataset utilizing the GPT3 (Brown et al., 2020) architecture. In the following section, various multilingual large language models will be introduced based on their

²There is no agreed-upon standard for defining the scale of a ‘large’ language models. What is considered ‘large’ is recognized at a given point in time, and the size of models has rapidly increased over recent years. Some research include language models from the era of the enlarged pre-trained model like BERT as LLMs (Shanahan, 2023), while others consider models around or exceeding 100B parameters based on their emergent capabilities (W. X. Zhao et al., 2023). In this study, we will refer to the language models of the pre-trained model era, including BERT, as large language models. This is to distinguish between the era before and after the introduction of the transformer architecture and the use of large-scale pre-trained language models, and to use the terminology widely used today.

architectures.

2.2 Overview of various MLLMs

To introduce multilingual large language models alongside their architecture types, it is necessary to briefly introduce the Transformer architecture, from which these architectures originated. The transformer architecture, introduced in Vaswani et al., 2017, serves as the foundation upon which many language models in use today are based. The transformer model was proposed as an attempt to overcome the limitations of prevalent RNN-based and CNN-based language models at the time. RNNs, due to their sequential nature, faced challenges with parallelization, making it difficult to process large datasets efficiently. Meanwhile, CNNs had difficulties managing long-range dependencies. The transformer model tackled these issues by utilizing the attention mechanism that allows it to simultaneously access and learn the relationships between all tokens in an input sequence. The Transformer's architecture, with its parallelized processing, offers significant advantages in quickly processing large datasets while also demonstrating excellent performance in language tasks, leading to its widespread adoption in subsequent large language models.

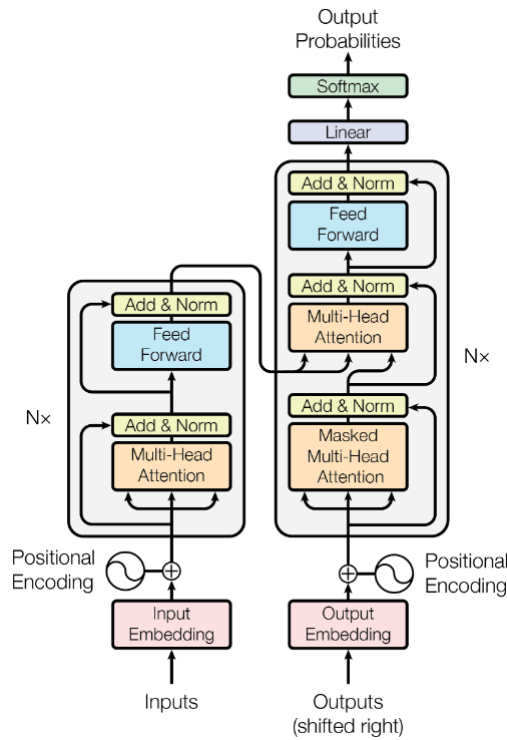


Figure 2.1: Transformer - model architecture. Image taken from Vaswani et al., 2017

The transformer has an encoder-decoder structure, with each consisting of stacks of layers having sub-layers of attention and feed-forward network layers, as shown in Figure 2.1. The encoder maps the entire input sequence (x_1, x_2, \dots, x_n) to a latent representation $z = (z_1, z_2, \dots, z_n)$ in a parallel manner. The decoder generates the output sequence (y_1, y_2, \dots, y_n) token by token based on z . At each step, the model refers to the encoder's output with cross-attention and also refers to tokens generated in previous time-steps

with masked self-attention in an autoregressive manner. Subsequent language models utilizing the transformer architecture have adapted its structure in diverse ways, such as using the architecture’s encoder or decoder independently (Wolf et al., 2020; Amatriain, 2023).

Encoder-based models Encoder models use only the encoder part of the transformer architecture. The self-attention mechanism of the encoder learns the relationships among all tokens in the input sequence simultaneously, capturing the context from both before and after each token’s position. In other words, encoder models consider the text’s context in a bidirectional³ manner. The training of the encoder model is based on masked language modeling, which involves predicting masked tokens based on the given context, a ‘fill-in-the-blanks’ task. During model training, tokens in the input sequence are randomly masked, and the model is trained to recover the original masked tokens. The training objective is to maximize the log likelihood given by Equation 2.2:

$$\sum_i m_i \log(P(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n); \theta_T) \quad (2.2)$$

where $m_i \in \{0, 1\}$ indicates whether x_i is masked and θ_T represents the model’s parameters (Min et al., 2023). Because the model learns the relationships among all tokens in the input sequence, it is effective at learning deep language representations. As a result, it shows excellent performance in natural language understanding (NLU) tasks such as text classification, natural language inference (NLI), and extractive question answering (Amatriain, 2023). When applying encoder models to downstream tasks, additional fine-tuning is performed specific to the task at hand. Representative models include BERT and a variety of model families (Rogers et al., 2020) derived from BERT, such as RoBERTa (Y. Liu et al., 2019) and ALBERT (Lan et al., 2020), fall under this category.

Multilingual language models with encoder-based architectures include mBERT, XLM, and XLM-R. mBERT (Devlin et al., 2019), one of the well-known multilingual pre-trained language models, is trained on Wikipedia dump data from the top 100 languages on Wikipedia. Due to the significant variance in the size of data across languages, the model uses exponentially smoothed weights to correct for this imbalance. This results in the under-sampling of high-resource languages and over-sampling of low-resource languages.⁴ The model also uses a shared vocabulary of 110k WordPiece tokens for tokenization. The pre-training task employed is masked language modeling (MLM), similar to BERT.

XLM (Conneau and Lample, 2019) employs not only masked language modeling (MLM) and causal language modeling (CLM) tasks that can be trained on monolingual corpora, but also incorporates translation language modeling (TLM) tasks combined with MLM tasks, leveraging parallel language data. This approach has led to improved performance compared to BERT. For training data, Wikipedia dumps are used for CLM and MLM,

³The term ‘bidirectional’ is used to emphasize contrast with traditional ‘unidirectional’ processing. In attention mechanisms, since the relationships between all tokens in the input sequence are learned simultaneously, it would be more accurate to describe the encoder in Transformer as ‘non-directional’.

⁴It is stated, for example, English would be sampled 1,000 times more than Icelandic in the original distribution, but after smoothing, it’s 100 times more.

while various sources of parallel data containing English (such as multilingual corpus from United Nation documents (MultiUN) and open parallel corpora(OPUS)) are used for TLM. Sampling adjustments based on language resource size are also employed, and a shared sub-word vocabulary is used through Byte Pair Encoding (BPE).

XLm-R (Conneau et al., 2020) takes inspiration from RoBERTa, which demonstrated that training BERT on more data for longer periods could significantly improve its performance. Utilizing the data CommonCrawl(Wenzek et al., 2020), XLm-R increased the amount of data for low-resource languages by an average of 100-fold, and is trained as a multilingual language model across 100 languages. Similar to mBERT, it is trained on multilingual MLM tasks using monolingual data corpora. With the substantial increase in training data for low-resource languages, the model’s performance in these languages improved considerably.

Encoder-decoder-based models Encoder-decoder models utilize both the encoder and decoder of the transformer architecture. This allows them to capture the entire text context using the bidirectional information from the encoder and also generate outputs using the decoder. Encoder-decoder models are sequence-to-sequence models that learn to generate a sequence (y_1, y_2, \dots, y_n) given an input sequence (x_1, x_2, \dots, x_m) . The models are pre-trained through mainly denoising tasks, where they generate the original sequence from a masked input sequence (Lewis et al., 2020; Raffel et al., 2020). Additionally, a causal language modeling task, which predicts the next token, can be combined for training (Soltan et al., 2022). Pre-trained models are subsequently applied to various NLP tasks through fine-tuning. The training objective of the encoder-decoder model is to maximize the log likelihood given by Equation 2.3:

$$\log(P(y_1, \dots, y_n | x_1, \dots, x_m); \theta_T) \quad (2.3)$$

where θ_T represents the parameters of the model (Min et al., 2023). Representative models in this category include the T5 family (Raffel et al., 2020; Xue et al., 2021, 2022), T0 (Sanh et al., 2022), BART (Lewis et al., 2020), and UL2 (Tay et al., 2022). Encoder-decoder models are suitable for tasks that involve generating new output sequences based on given input sequences, such as summarization, translation, or generative question answering (Raffel et al., 2020; Amatriain, 2023). Furthermore, the encoder-decoder model can be versatilely applied to various tasks that can be framed as sequence-input to sequence-output, such as image-captioning. However, the architecture of the encoder-decoder models is more complex compared to encoder-only or decoder-only models, making their training and scaling more challenging.

mBART (Y. Liu et al., 2020) is a multilingual encoder-decoder model based on the BART. It is trained on datasets extracted from 25 languages in the Common Crawl (CC) corpora. Similar to XLm, it employs up/down sampling based on the amount of data available for each language to achieve a balance between languages. mBART is trained on large-scale monolingual corpora from multiple languages with BART’s training objectives, which are to reconstruct the original text from the input text that has been noised by spanned masking and sentence permutation. As mBART trains both its encoder and decoder simultaneously, it can be utilized for a wide range of tasks that can be framed in a sequence-to-sequence (seq2seq) configuration, including translation

tasks.

mT5 (Xue et al., 2021) is a multilingual variant of T5 (Raffel et al., 2020). While inheriting the advantages of T5, it has been trained using a multilingual variant of the C4 dataset, known as mC4. mC4 comprises natural text in 101 languages sourced from public common crawl web scrapes. Similar to T5, mT5 employs span masking, training on input token sequences replaced with mask tokens to be reconstructed into output token sequences with masks removed. During training, like other multilingual models, it applies up/down sampling across languages. Given that one of the features of T5 was its scale, mT5 also tends to be on the larger scale, as of that time (300M, 580M, 1.2B, 3.7B, 13B). Especially, mT5 is an encoder-decoder model, meaning it has roughly twice as many parameters as encoder-only models of a similar scale. mT5 has shown robust performance across various benchmarks.

M2M-100 (Fan et al., 2021) is a transformer-based encoder-decoder translation model. It aims to overcome the English-centric bias found in existing translation models, which are typically trained on datasets where either the source or target language is English. A large-scale many-to-many dataset for 100 languages was constructed through extensive mining on CommonCrawl to cover thousands of language directions, and the M2M model trained on this data has shown improvements in translation quality between non-English language directions. The model is trained by feeding an input token sequence and source language information to the encoder, and supplying the decoder with target language information, allowing the decoder to produce an output token sequence.

Decoder-based models Decoder models utilize only the decoder part of the transformer architecture. The masked self-attention mechanism of the decoder is trained to predict the next token based on all the given preceding tokens. They are sometimes called “autoregressive language models” as they predict the current state from their past states, or “causal language models” as they generate the next token based on the causality with previous tokens. Their proficiency in text generation also earns them the name “generative language models.” The training objective of the decoder model is to maximize the log likelihood given by Equation 2.4:

$$\sum_i \log(P(x_i | x_1, x_2, \dots, x_{i-1}); \theta_T) \quad (2.4)$$

where θ_T represents the model’s parameters (Min et al., 2023). Representative models in this category include the GPT family (Radford et al., 2018, 2019; Brown et al., 2020), PaLM (Chowdhery et al., 2022), BLOOM (Scao et al., 2023) and LLaMA (Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023). Decoder models, with their simple structure yet high performance, have been widely adopted in subsequent language model research. Another advantage is their adaptability to various NLP tasks through methods like fine-tuning or prompt tuning. As demonstrated by the GPT 1, 2, and 3 models, increasing the model size and training data have led to significant performance improvements, contributing to the trend of upsizing language models. Large decoder models have performed well in few-shot or zero-shot learning (Brown et al., 2020; Kojima et al., 2022), and their emergent capabilities (Wei, Tay, et al., 2022) have recently garnered considerable attention from researchers.

XGLM (Lin et al., 2022) is a multilingual generative language model based on the GPT-3 (Brown et al., 2020) architecture. It was developed to test the few-shot and zero-shot learning capabilities shown in GPT-3 across multiple languages. For training, the pipeline used for mining the CC100 corpus (Conneau et al., 2020; Wenzek et al., 2020) is expanded to create a large-scale multilingual dataset CC100-XL, from which a pre-training dataset consisting of 500B tokens across 30 different languages is constructed. To achieve a more balanced language distribution, up-sampling is performed for middle and low-resource languages, and the model is trained at four sizes: 564M, 1.7B, 2.9B, 7.5B. Through tests on downstream tasks in various languages, XGLM demonstrates the ability to achieve competitive zero-shot and few-shot learning performances depending on the prompt, in a multilingual environment.

mGPT (Shliazhko et al., 2022) is also a multilingual generative language model developed using the GPT-3 architecture. It is trained on data from 60 languages across 25 language families, utilizing Wikipedia and the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020). Multiple tokenization strategies are tested and the optimal one is applied during training. The model comes in two sizes: 1.3B and 13B. Similar to XGLM, mGPT’s zero-shot and few-shot learning capabilities are evaluated across a broad range of multilingual tasks. The model’s perplexity is also assessed for each individual language. Despite supporting more languages than XGLM, mGPT demonstrates performance levels that are on par with it.

BLOOM (Scao et al., 2023) is an open-access multilingual language model with a decoder-only Transformer architecture. It is trained on the ROOTS corpus, a dataset composed of hundreds of sources in 46 natural languages and 13 programming languages, and is a large language model with 176B parameters. BLOOM is designed to perform a wide variety of tasks competitively in multilingual settings, and can achieve further improved performance through multitask fine-tuning. BLOOM is a model developed and released by the collaboration of hundreds of researchers, providing detailed process documentation, as an effort towards open research of large language models.

Instruction fine-tuned models Additionally, besides these pre-trained base models, models that have been further fine-tuned to facilitate the use of these large language models in downstream tasks have also been developed. Since fine-tuning large language models individually is challenging and inefficient, multitask prompt fine-tuning (MTF) (Sanh et al., 2022) or instruction tuning (Wei, Bosma, et al., 2022) was devised for zero-shot task generalization of large language models. T0 (Sanh et al., 2022) is a variant of T5 (Raffel et al., 2020) that has been through multitask prompt fine-tuning and showed strong zero-shot task generalization ability. FLAN (Wei, Bosma, et al., 2022), a model that applies instruction tuning to a 137B pre-trained base model, also showed significantly improved zero-shot performance than the base model. Building on this, FLAN-T5 (80M, 250M, 780M, 3B, 11B) (Chung et al., 2022) and FLAN-UL2 models, which applied the datasets and fine-tuning from FLAN to T5 and UL2 (Tay et al., 2022) models respectively, were also released.

Inspired by this, Muennighoff et al. (2023) applied multitask prompt fine-tuning to BLOOM and mT5, yielding fine-tuned variants named BLOOMZ and mT0. They constructed corpora P3, xP3, and xP3mt for fine-tuning, which are the English-only corpus used for training T0, the corpus added multilingual datasets in 46 languages,

and the corpus also included machine translation prompts, respectively. The models were then fined-tuned on each of these corpora and their performances were compared across various aspects. The results demonstrated improved zero-shot generalization, particularly showing the capability for zero-shot generalization to new tasks in unfamiliar languages. This suggests that the model has learned some abilities that are not constrained by specific languages or tasks.

2.3 Benefits of MLLMs

Multilingual language models have many advantages. Key benefits include:

- Cross-lingual transfer - Knowledge learned in one language can be transferred and utilized in tasks of another language. This allows leveraging knowledge from high-resource languages to improve the performance of tasks in low-resource languages.
- Capturing relationships between languages - Multilingual language models can better generalize tasks across various languages and be more resilient to noise, providing insights into linguistic universality.
- Resource efficiency - Multilingual language models are more efficient than training individual language models for each language, in terms of computational power, resources and storage, and easier to maintain.
- Inclusivity - By integrating and supporting low-resource languages, multilingual language models promote linguistic diversity and inclusivity.
- Broad applicability - Multilingual language models can be used in various NLP applications across multiple languages, and support multilingual collaboration.

In the following sections, focusing on cross-lingual transfer and relationships between languages, related research are examined in detail.

2.3.1 Cross-Lingual Transfer

One of the remarkable capabilities of multilingual language models is cross-lingual transfer, which refers to the multilingual model’s ability to apply knowledge learned in one language to tasks in another language. This is especially beneficial for tasks in low-resource languages with limited labeled or unlabeled data available. By leveraging knowledge learned from high-resource languages that have abundant training data, these models can enhance performance in downstream tasks for languages where such data is scarce. Given that labeled data in the target language may be limited, the model’s zero-shot and few-shot transfer abilities are usually explored. For the application of cross-lingual transfer, typically, a multilingual model pre-trained on multiple source languages serves as the base. This base model is fine-tuned on a dataset for a specific task in one language and then utilized to perform the same task in another language.

In relation to cross-lingual transfer in multilingual language models, various studies have explored different factors; for more detailed information, one can refer to Doddapaneni et al. (2021) and Philipppy et al. (2023). Here, we will provide a brief overview of the key

aspects related to cross-lingual transfer.

Linguistic Similarity Pires et al. (2019) and Conneau et al. (2020) demonstrate that multilingual models mBERT and XLM-R, respectively, possess competent zero-shot cross-lingual transfer abilities, despite being trained on monolingual corpora of multiple languages without explicit training for multilingual representation. Pires et al. (2019) carries out probing experiments on several hypotheses concerning this, finding that cross-lingual transfer works better between typologically similar languages. In line with this, K et al. (2019), Conneau and Lample (2019) and Dufter and Schütze (2020) observe that language similarity affects cross-lingual capabilities as well. Lauscher et al. (2020) also shows that language similarity plays an important role in cross-lingual transfer, especially for low-level tasks such as POS tagging and parsing. Additionally, Vries et al. (2022) illustrates that the similarity between source and target languages during fine-tuning, and further, the inclusion of the target language or languages similar to the target language in pre-training, can enhance cross-lingual transfer.

Shared Vocabulary Wu and Dredze (2019) observes a positive correlation between sharing subwords and zero-shot transfer performance. However, Pires et al. (2019) find that zero-shot transfer is largely independent of vocabulary overlap, showing considerable transfer capabilities even between languages with different scripts, i.e., no lexical overlap. Artetxe et al. (2020) also shows that sharing subwords is not necessary for zero-shot transfer, instead, the effective vocabulary size per language is important. Similarly, K et al. (2019) and Conneau and Lample (2019) derive findings through experiments that there is no significant correlation between shared vocabulary and cross-lingual transfer performance. In relation to this, Patil et al. (2022) underscore the importance of lexical overlap when the pre-training corpus for the source language is small. Similarly, Deshpande et al. (2022) highlight the significance of lexical overlap when the word order between the source and target languages differs. These studies contribute to a more nuanced understanding of the impact of lexical overlap. Philippy et al. (2023) concludes from these existing studies that lexical overlap is not a sufficiently independent factor for explaining cross-lingual transfer.

Model Architecture and Size K et al. (2019) examines the influence of several architectural factors on multilingual language models, finding that deeper network architectures lead to improved performance in both monolingual and cross-lingual tasks. They observe that the number of attention heads has little impact on cross-lingual capabilities, and that the total number of parameters also become less influential beyond a threshold.

In relation to the total number of parameters, i.e., model size, some studies (Goyal et al., 2021; Xue et al., 2021; Z. Chi et al., 2022) show that larger models improve multilingual performance. This aligns with findings from Conneau et al. (2020), the ‘curse of multilinguality’⁵ and its alleviation by increasing model size. However, Conneau and Lample (2019) demonstrates through experiments on shared

⁵The phenomenon where cross-lingual performance on languages improves only up to a point as the number of languages increases with a fixed model capacity, after which it declines as the model capacity gets diluted among the languages.

layer numbers that parameter sharing across languages is a crucial element for effective multilingual representation. Dufter and Schütze (2020) also shows that when the model is overparameterized, i.e., when the number of parameters is too large for the languages, cross-lingual transfer ability deteriorates. They suggest that when the number of model’s parameters is limited, the model identifies common structures across languages to efficiently use parameters, creating a multilingual space. Regarding these seemingly contradictory findings, Dufter and Schütze (2020) notes that XLM-R, for example, which is trained on 104 languages, would require a very large number of parameters to sufficiently train for all these languages, making it difficult to be overparameterized. It appears that further research is needed to ascertain whether recent large-scale multilingual models dealing with many languages are underparametrized or overparametrized for each language.

Pre-training Corpora Size Conneau et al. (2020) has demonstrated that using large-scale corpora for pre-training leads to an overall improvement in cross-lingual transfer performance, as shown through their research on XLM-R. Similarly, Lauscher et al. (2020) demonstrates that the amount of target language data used in pre-training has a crucial impact on zero-shot transfer performance, especially in high-level tasks such as NLI and QA. C.-L. Liu et al. (2020) also finds that using larger pre-training datasets improves cross-lingual capabilities. Particularly, they highlight that leveraging larger pre-training data along with longer context windows to capture longer dependencies during training is beneficial for cross-lingual performance.

Alignment between Languages Conneau and Lample (2019) has shown, along with the introduction of the XLM model, that explicit cross-lingual pre-training with the use of parallel corpora can yield enhanced zero-shot cross-lingual ability and machine translation performance in multilingual models. Inspired by this, Z. Chi et al. (2021) and Z. Chi et al. (2022) also leverage parallel corpora in their models’ pre-training and introduce the models InfoXLM and XLM-E respectively, demonstrating that explicit alignment between languages improves the performance of the models on cross-lingual tasks. Similarly, Dufter and Schütze (2020) demonstrates that as the training corpus becomes less parallel, meaning as the comparability between languages in the corpus decreases, the multilinguality diminishes.

On the other hand, research aimed at enhancing machine translation and cross-lingual transfer through the alignment of word embeddings between languages has also consistently continued (Ruder et al., 2019 since Mikolov et al. (2013) attempted to map bilingual word embeddings by linear transformation. Several studies (Cao et al., 2019; Q. Liu et al., 2019; Wang* et al., 2019) demonstrate improvements in zero-shot cross-lingual transfer performance through the alignment of contextualized word embeddings from pre-trained language models across languages. Deshpande et al. (2022) also analyzes the static token embeddings trained in a bilingual language model and finds that the alignment between embeddings is closely correlated with the zero-shot transfer performance.

In relation to these alignment methods between languages, Wang* et al. (2019) proposes a framework that combines word embedding alignment and joint training methods.

2.3.2 Capturing relationships between Languages

The cross lingual transfer ability of multilingual language models naturally raises the following questions: How is the cross lingual transfer in multilingual language models possible? Do multilingual language models learn some kind of common linguistic patterns from multiple languages? What does the sub-representation space in multilingual language models look like? How are the embeddings of languages arranged and what are their relationships? In this section, we will look at some studies that address these questions.

Linguistic Universals The effort to explore the universal features, structures, and principles of language is an old topic that many researchers, including Chomsky (1957), who proposed the theory of universal grammar, and Greenberg (1963), who tried to identify common features and patterns through comparative analysis of various languages, have been interested in. This stems from the idea that language originates from some unique and common human abilities. Can neural network models capture universal patterns of language through data-based statistical learning? Pires et al. (2019) demonstrates through nearest neighbor translation experiments using mBERT that accuracy is highest in the intermediate layers of the model. They suggests from this that hidden representations share a common subspace that represents linguistic information in a language-agnostic manner. The decrease in accuracy in initial or final layers is presumed to be due to the necessity of language-specific information at those stages. Libovický et al. (2019) evaluates mBERT representations on semantic cross-lingual tasks, and they observe that although mBERT representations are not very language-neutral, the use of certain strategies can enhance language neutrality, leading to improved performance in cross-lingual tasks. E. A. Chi et al. (2020) finds syntactic dependency clusters through sub-space exploration of mBERT using structural probing, suggesting that mBERT induces universal grammatical relations without explicit supervision. Muller et al. (2021) also demonstrates that mBERT has a multilingual encoder for aligning multilingual representations and a predictor for performing task-specific downstream tasks in a language-agnostic manner.

Multilingual Representation Space There have been various studies on what the subspace representations of languages in multilingual models might look like. As Pires et al. (2019) speculated the existence of a common space shared among languages through experiments on multilinguality with mBERT, Conneau and Lample (2019) also demonstrates, using BERT, that each language has similar alignments, and cross-lingual performance improves when there are layers shared among different languages, assuming the existence of a shared multilingual representation space across languages. On the other hand, Singh et al. (2019) presents an contrasting finding that the model does not use a common shared space across languages but partitions representations per language through canonical correlation analysis of mBERT representations. Regarding these conflicts, Del and Fishel (2022) analyzes that they are due to methodological discrepancies such as pooling strategies or similarity index selections, and shows that by altering these methods, it's generally found that most languages share a cross-lingual space. Chang et al. (2022) conducts a case study with XLM-R to demonstrate through visualization that language subspaces are similar to each other, having only

linear differences, and the shared representation space can be separated into axes coding language-sensitive information and axes coding language-neutral information. On the other hand, Ravishankar and Nivre (2022) discovers that the ability to construct multilingual spaces is more closely correlated with specific corpus-level characteristics rather than with languages or language families.

However, there may also be drawbacks to the alignment between languages and cross-lingual transfer in multilingual large language models. Languages contain the culture and mindset of the speakers, as well as unique usage that cannot be addressed by simple parallel translation. Additionally, issues like bias and toxicity found in large language models are also present in their multilingual large language models. In the next section, we will look at these issues and concerns related to the use of multilingual large language models.

2.4 Issues with MLLMs

Multilingual large language models come with their own issues and concerns, alongside their unique advantages. These include:

- Performance compared to monolingual models - Multilingual models might have lower performance on specific languages compared to monolingual models trained on those languages.
- Performance disparities across languages - Multilingual models can perform better on certain languages based on the language distribution of their training data,
- Linguistic and cultural biases - The language bias multilingual models can lead to linguistic and cultural biases in downstream tasks across different languages.
- Research gaps - While multilingual models operate in various languages, research from the perspective of each individual language is insufficient.

The following sections will provide a more detailed examination of these matters.

2.4.1 Performance gap

Monolingual Models vs. Multilingual Models

One question that arises regarding multilingual language models is whether they can perform as well for a specific language as a monolingual model trained only on that language data. It is easy to expect that a model trained entirely on one language would have a richer understanding of that language, compared to a model trained on multiple languages with the same resources. In relation to this, Conneau et al. (2020) mentions the curse of multilinguality, meaning that as the number of languages increases while the model size remains fixed, cross-language performance for low-resource languages tends to improve, but overall downstream performance suffers due to capacity dilution. However, Conneau et al. (2020) also demonstrates that increasing the model capacity alleviates

this issue to some extent, suggesting that it’s possible to have a large-scale model for all languages without performance degradation per language.

Similar questions have led to many studies comparing the performance of monolingual models and multilingual models over time. Some studies (Virtanen et al., 2019; Le et al., 2020; Samuel et al., 2023) show that pre-trained language models trained on a single language exhibit much better performance than their multilingual counterparts. However, Doddapaneni et al. (2021) that reviewed various research comparing the performance of monolingual and multilingual models concludes that there is no clear winner. This is because various factors, such as model capacity, the amount of pre-training data, tokenization, fine-tuning methods, and the amount of task-specific training data, all influence performance. Doddapaneni et al. (2021) suggests that systematic and comprehensive experiments controlling for these multiple variables are necessary to make a fair comparison between monolingual and multilingual models.

Difference between languages in multilingual models

On the other hand, another question that can arise is whether the performance would be similar across different languages when using the multilingual language models. Cross-lingual transfer performance in multilingual models is one of the main goals of multilingual language model development, so many studies and benchmarks have been conducted to test it. These include various multilingual benchmarks that extend existing NLP tasks into multiple languages, such as XNLI (Conneau et al., 2018), PAWS-X (Yang et al., 2019), XQuAD (Artetxe et al., 2020), and benchmark sets that integrate them, such as XGLUE (Liang et al., 2020), XTREME (Hu et al., 2020) and XTREME-R (Ruder et al., 2021).

Kassner et al. (2021) examines whether mBERT can be used as a multilingual knowledge base through multilingual knowledge probing. They find that mBERT can be used as a knowledge base but its performance varies depending on the language, which suggests that mBERT does not store entity knowledge in a language-independent manner. Cabello Piqueras and Søgaaard (2022) shows that pretrained multilingual language models (mBERT, XLM-R, mT5) exhibit differences in group fairness across languages. Huang et al. (2023) indicates, through several evaluation results of large language models (Bang et al., 2023; Hendy et al., 2023; Jiao et al., 2023; Zhu et al., 2023), that these models struggle to understand and generate non-English languages, especially in low-resource ones. Shi et al. (2022) investigates the multilingual reasoning abilities of large language models like GPT-3 and PALM, and Zhang et al. (2023) proposes a systematic method to analyze the performance disparities of LLM in multilingual environments.

There could be several reasons for the performance differences between languages in multilingual models. Nicholas and Bhatia (2023) points out that the data for training multilingual language models for low-resource languages are often machine-translated texts that contain words or errors that local users do not use, rather than natural languages of local users. Also, the usage and the context of local users for those languages may not have been sufficiently considered due to lack of understanding of the languages. Additionally, the benchmarks and tasks for testing the models are often scarce for low-resource languages, leading to the use of benchmarks translated from English, which further makes it difficult to properly evaluate and improve performance for low-resource

languages. Ranathunga and de Silva (2022) comprehensively reviews the imbalances existing among languages worldwide from various perspectives, including not only data resource availability, but also language family, speaker population, geographical location, and GDP, along with the possible reasons and attempts to address. Such disparities between languages can result in actual disadvantages for users of low-resource languages. Ahia et al. (2023) shows that OpenAI’s language model API policy doesn’t consider the token count difference needed to convey the same information in different languages, and as a result, speakers of many languages end up receiving lower results while paying more, as demonstrated through benchmarks in 22 languages.

2.4.2 Biases

General Biases in Language Models

Bias is not an issue exclusive to multilingual language models. Many machine learning models that are trained based on data, including language models, are not free from the consideration of bias. In some cases, the models learn bias due to the training data that contains the negative bias of the real world, and in other cases, the models learn bias due to the under-/over-representation of certain groups in the real world in the training data. Bender et al. (2021) indicates that when considering those who mainly contribute to the generation of the large internet-based data used for model training, young users, users from advanced countries, and men are excessively represented, and this can encode perspectives such as white supremacy, misogyny, and age discrimination, into the model and potentially harm marginalized people. Furthermore, Hovy and Prabhumoye (2021) suggests that bias can occur at various stages in NLP systems, including the data, the annotation process, the input representations, the models, and finally the research design.

The biases observed in language models are usually observed in multilingual language models as well. The biases discussed in language model studies include gender (González et al., 2020; J. Zhao et al., 2020; Kaneko et al., 2022; Steinborn et al., 2022), race & ethnicity (Ahn and Oh, 2021; Field et al., 2021; Nadeem et al., 2021; Ousidhoum et al., 2021), nationality (Venkit et al., 2023), religion (Abid et al., 2021), disability (Hutchinson et al., 2020; Venkit et al., 2022), occupation (Touileb et al., 2022), age (Diaz et al., 2018), intersectional bias (Guo and Caliskan, 2021), etc., across various minority identities.

Biases unique to multilingual models

On the other hand, there are biases that become highlighted in multilingual models. These biases can occur during the process of cross-lingual transfer in multilingual models. That is, due to cross-lingual transfer, characteristics that are more oriented towards high-resource languages can be discovered in the downstream tasks of low-resource languages.

Linguistic bias Firstly, there is the grammatical bias where the linguistic characteristics of high-resource languages influence those of low-resource languages. Papadimitriou et al.

(2023) compares the fluency of multilingual models to monolingual models in Spanish and Greek and finds that multilingual models are biased towards English-like grammatical structures. Naous et al. (2023) shows that when processing and generating Arabic text, language models often prefer and generate content suitable for Western culture, and finds that cultural bias is more exacerbated when using Arabic prompts with English-like structures than when using Arabic prompts with Arabic-like structures. This implies that sentence structures that are more grammatically aligned with English contribute to increasing Western-oriented cultural bias.

Cultural bias Secondly, there is the cultural bias where the cultural characteristics of high-resource languages influence those of low-resource languages. Hämmerl et al. (2022, 2023) investigates whether the cultural values including moral norms of high-resource languages are applied to low-resource languages in multilingual models. The experiment results show that pre-trained multilingual language models encode varying moral biases depending on the language, but these do not necessarily align with cultural differences or common human opinions. Arora et al. (2023) attempts to measure cultural values embedded in multilingual pre-trained language models by asking questions based on the World Values Survey⁶ in 13 languages and analyzing the correlation with existing survey results. The results demonstrate that multilingual pre-trained language models capture cultural value differences, but the correlation with existing value survey results is weak. Ramezani and Xu (2023) similarly uses the World Values Survey to investigate differences in moral norms encoded in English pre-trained language models. The authors also find that these norms do not align with existing survey results but discover that fine-tuning the model with survey data improves inference across countries at the expense of a slight decreased accuracy in English moral norm estimation. Naous et al. (2023) investigates whether language models are culturally biased using multilingual models and Arabic monolingual models. The experiment results show that both multilingual and Arabic monolingual models exhibit biases towards Western culture across eight cultural aspects including person names, food, clothing, location, literature, beverages, religion, and sports. Masoud et al. (2023) proposes an approach to evaluate whether LLMs align with specific cultural values and norms using Hofstede’s Cultural Alignment Test (CAT) framework. The research results reveal that GPT-3.5 and GPT-4 align relatively well with the cultural values of the United States, but the cultural values of other countries like China, Saudi Arabia, and Slovakia are not fairly reflected in GPT-3.5, GPT-4, BARD.

On the other hand, Yin et al. (2022) analyzes geographically-diverse common sense in multilingual language models. Using GEOMLAMA, a benchmark set that covers concepts shared by the cultures of the United States, China, India, Iran, and Kenya, they test 11 multilingual pre-trained language models and find that larger models do not necessarily store more geo-diverse knowledge than smaller models, that multilingual language models are not inherently biased towards Western countries’ knowledge, that the native language of a country may not be the most suitable language for investigating knowledge of that country, and that a language may be more competent in investigating knowledge of its non-native countries. Meanwhile, B. Li and Callison-Burch (2023) introduces a concept of geopolitical bias, which is the tendency to report different geographic knowledge depending on the linguistic context. Taking

⁶<https://www.worldvaluessurvey.org>

territorial disputes between countries as a case study, they discover that linguistic context significantly impacts GPT-3’s responses and, unlike multilingual humans, language model’s knowledge lacks consistency across languages.

These research findings indicate that measuring cultural values or moral norms using multilingual language models is not straightforward and leaves room for additional experiments and interpretation.

On another note, Ventura et al. (2023) explores the multilingual Text-to-Image (TTI) models’ cultural awareness, cultural differences, and cultural characteristics. Through comprehensive evaluations of TTI models such as StableDiffusion, AltDiffusion, DeepFloyd, and DALL-E, including intrinsic assessments, extrinsic assessments, and human evaluations, the authors observe that the models possesses unique cultural awareness, identify cultural similarities in the generated images, raise questions about the fair representation of diverse cultures, and reveal how language encodes cultural nuances and influences cultural perception and interpretation. The study shows that exploring the Text-to-Image model’s cultural biases is also a promising approach.

The language bias stemming from the English-centric training data of multilingual large language models, as examined above, could become increasingly problematic as the use of these models spreads globally. The language bias of multilingual models can influence downstream tasks across different languages in the form of linguistic or cultural biases, potentially leading to the loss or distortion of unique linguistic or cultural characteristics and nuances of individual languages. Particularly, identifying and quantifying such phenomena can be challenging due to their ambiguity and subjectivity. In this study, we aim to investigate the linguistic biases in multilingual large language models by using the text generation task in Korean, a non-English language. In the following chapters, experiments will be designed and conducted to explore this issue.

Chapter 3

Methods

In this chapter, we describe the experimental design and evaluation methods used in the study. First, we present the details of the experiments conducted in the research. This includes the language models employed in the experiments, the experimental tasks, the datasets used, the hyperparameter settings for text generation, and the text generation procedure are explained. Next, the evaluation methods, procedures, and metrics used to assess the experimental results are introduced. The texts generated by the language models and the Korean sentences filtered from them are evaluated and analyzed at various dimensions, including basic statistics, surface-level, lexical, syntactic, semantic, and English translationese aspects.

3.1 Models

This study uses open-source models in the experiments to ensure research accessibility, reproducibility, transparency, and cost-efficiency. Also, this study employs models with parameter sizes of 13B or less in the experiments due to practical considerations in experimentation and computational resource constraints. To investigate the multilingual capabilities of the pre-trained base models themselves, without the influence of the datasets and tasks used for fine-tuning, the study focuses on pre-trained language models. The language models to be tested in this study encompass a diverse range of sizes and types, including Korean monolingual models, multilingual models trained on various languages including Korean, language models with additional pre-training on Korean, and language models not trained on Korean. These models can be organized as follows.

3.1.1 Korean Monolingual Models

Korean monolingual models refer to language models pre-trained on Korean text data. These models are used to evaluate the performance of language models trained on a single language, Korean, in the task of Korean text generation. This allows for the assessment of the capabilities of language models specialized in Korean.

KoGPT2 base (v2) (SKT-AI, 2020) Released in 2020, KoGPT2 is a Korean monolingual model utilizing the GPT2 (Radford et al., 2019) architecture with a parameter size of 125M. It was trained on 40GB of Korean data composed of various sources, including Korean Wikipedia and news articles. The vocabulary size is 51,200.

Ko-GPT-Trinity 1.2B (v0.5) (SKT-AI, 2021) Released in 2021, Ko-GPT-Trinity is a Korean monolingual model replicating the GPT3 (Brown et al., 2020) architecture with 1.2B parameters. It was trained on ko-DAT, a large scale curated dataset predominantly consisting of Korean data, with a total of 35 billion tokens. The vocabulary size is 51,200.

KoGPT (I. Kim et al., 2021) Released in 2021, KoGPT is a Korean monolingual model utilizing the GPT3 architecture with 6B parameters. It was trained on 200B tokens of Korean data. The vocabulary size is 64,512, and the context window size is 2048.

Polyglot-Ko (Ko et al., 2023) Initially released in 2022, and additionally released with a model of 12.8B parameters in 2023, Polyglot-Ko is a Korean monolingual model trained using the GPT-NeoX codebase (Andonian et al., 2023). It was released in four model sizes: 1.3B, 3.8B, 5.8B, and 12.8B parameters. The models were trained on 1.2TB (filtered down to 863GB after preprocessing) of Korean data. Specifically, the 1.3B, 3.8B, 5.8B, and 12.8B models were trained on 213B, 219B, 172B, and 167B tokens, respectively. The vocabulary size is 30,003, and the context window size is 2,048.

3.1.2 Multilingual Models

Multilingual models refer to language models pre-trained on text data from multiple languages, including Korean. These models are used to examine the performance of language models that have learned Korean as one of the various languages. Additionally, it can be observed how the cross-lingual transfer abilities, which are expected to be acquired during the learning process of multiple languages, influence Korean text generation.

XGLM (Lin et al., 2022) Initially released in 2021, and additionally released with a model of 4.5B parameters in 2022, XGLM is a multilingual language model replicating the GPT3 architecture and is trained on a multilingual corpus. Five model sizes were released: 564M, 1.7B, 2.9B, 4.5B, and 7.5B parameters. This study uses the 564M, 1.7B, 4.5B, and 7.5B models. The multilingual corpus used for training consists of 500B tokens in 30 diverse languages (134 for the 4.5B model¹), with under-resourced languages being upsampled to attempt a more balanced language representation rendering. The Korean data is reported to have a size of 79.08GB data and 20,002M tokens, with an estimated² proportion of 1% before upsampling and about 9% after upsampling in the entire corpus. The vocabulary size is 256,008, and the context window size is 2,048.

mGPT (Shliazhko et al., 2022) Released in 2022, mGPT is a multilingual language

¹<https://github.com/facebookresearch/fairseq/tree/main/examples/xglm>

²Estimated from Figure 1 in Lin et al. (2022)

model based on the GPT3 architecture and trained on a multilingual corpus. Two model sizes were released: 1.3B and 13B parameters. The multilingual corpus used for training consists of MC4 and Wikipedia, with a size of 600GB (488B UTF characters, 400B tokens), supporting 61 languages across 25 language families. The pre-training corpus was balanced by language family to improve language modeling capabilities. The Korean data size is estimated³ to have between 3B to 4B tokens. The vocabulary size is 100,000, and the context window size is 512.⁴

Bloom (Scao et al., 2023), another well-known multilingual model supporting 46 natural languages, was excluded from the experiments as it was not trained on Korean.

3.1.3 Korean Continually Pre-trained Models

Korean continuously pre-trained models refer to models that have been further pre-trained on Korean text data, starting from existing pre-trained models. This study includes models that have been continuously pre-trained on Korean data, based on models pre-trained on English, or English and Chinese. These models can be interpreted as bilingual or trilingual models that include Korean. This allows for the evaluation of the performance of language models that are not solely specialized on Korean but include Korean as one of the main training languages.

(1) Llama 2 series

Llama-2-Ko 7B (L. Junbum, 2023a) Released in 2023, Llama-2-Ko 7B is based on the Llama 2 7B model and has been continually pre-trained (Son et al., 2022; Ke and Liu, 2023) using Korean data. It is based on the Llama 2 architecture and has 7B parameters. The continual training data consists of a mix of Korean online data, with over 40B tokens.⁵ To improve the efficiency of Korean tokenization, Korean tokens obtained from the continual training data using SentencePiece were added to the existing Llama 2 tokens,⁶ expanding the vocabulary size to 46,336. The context window size is 4,096.

Open-Llama-2-Ko 7B (L. Junbum, 2023b) Released in 2023, Open-Llama-2-Ko 7B is based on the Llama 2 7B model and has been continually pre-trained using publicly accessible Korean data. The difference from the Llama-2-ko model is that the training data consists only of publicly accessible data, allowing use by everyone without restriction. It is based on the Llama 2 architecture and has 7B parameters. The training data consists of a curated mix of publicly accessible Korean corpora, with a size of 61GB data and over 15B tokens.⁷ It also uses an expanded vocabulary with a size of 46,336, by adding Korean tokens. The context window size is 4,096.

³Estimated from Figure 1 in Shliazhko et al. (2022)

⁴Although the Hugging Face page for the mGPT-13B model states that the context window size is 2048, considering the description in Shliazhko et al. (2022) and the nonsensical generated results when providing input tokens exceeding 512, it can be estimated that the context window size of the mGPT13B model is also 512.

⁵over 40B tokens with the expanded tokenizer

⁶Junbum Lee, 모두를 위한 한국어 Open Access LLM, ModuPop seminar 2024, 2024.01.30 <https://www.youtube.com/watch?v=ZoMWvu4RsGc>

⁷over 15B tokens with the expanded tokenizer, which is over 60B tokens with the original llama tokenizer

Llama-2-KoEN 13B (L. Junbum and Choi, 2023) Released in 2023, Llama-2-KoEn 13B is based on the Llama 2 13B model and has been continually pre-trained using Korean and English data. To prevent catastrophic forgetting (Koloski et al., 2023) in English, the model is trained on mixed data consisting of Korean and English in a 1:1 ratio, with over 60B tokens.⁸ It is based on the Llama 2 architecture and has 13B parameters. Like other Llama-2-Ko models, it uses an expanded vocabulary with a size of 46,336, by adding Korean tokens. The context window size is 4,096.

(2) SOLAR series

OPEN-SOLAR-10.7B (L. Junbum, 2024a) Released in 2024, OPEN-SOLAR-10.7B is based on the SOLAR-10.7B model and has been continually pre-trained using publicly accessible Korean data. It is based on the llama2 architecture, following the base model SOLAR-10.7B, and has 10.7B parameters. The training data consists of a curated mix of publicly accessible Korean corpora, with a size of 61GB data and over 15B tokens. The vocabulary was expanded by adding Korean tokens to the existing Llama-2 tokens. The vocabulary size is 46,592, and the context window size is 4,096.

SOLAR-KOEN-10.8B (L. Junbum and Choi, 2024) Released in 2024, SOLAR-KOEN-10.8B is based on the SOLAR-10.7B model and has been continually pre-trained using Korean and English data. Similar to the Llama-2-KoEn 13b model, to prevent Catastrophic Forgetting in English, mixed data consisting of Korean and English in a 1:1 ratio, with over 60B tokens, was used for continual pre-training. It is based on the llama2 architecture and has 10.8B parameters. The extended vocabulary is used for efficient Korean tokenization. The vocabulary size is 46,336, and the context window size is 4,096.

(3) Yi series

Yi-Ko-6B (Yi-KoEn-6B) (L. Junbum, 2024b) Released in 2024, Yi-Ko-6B is based on the Yi-6B model and has been continually pre-trained using Korean and English data. To prevent Catastrophic Forgetting in English, mixed data consisting of Korean and English in a 1:1 ratio, with over 60B tokens, was used for continual pre-training. It is based on the llama2 architecture, following the base model Yi-6B, and has 16B parameters. It uses an expanded vocabulary by adding Korean tokens. The vocabulary size is 78,464, and the context window size is 4,096.

3.1.4 Non-Korean-trained Models

Non-Korean-trained models, i.e., models not trained on Korean refer to models whose pre-training data does not explicitly target the Korean language. While a minimal amount of Korean data may be included in the actual pre-training data, it is expected

⁸Junbum Lee, op.cit.; Taekyoon Ted Choi and Junbum Lee, Efficient Training for Korean-English Cross-Language Models: Unlocking Solutions under \$10K, LangChain KR meet-up 2024 Q1, 2024.02.26 <https://aifactory.space/task/2719/discussion/836>

The model is continually trained on a dataset of over 60B tokens with the expanded tokenizer, and since the dataset consists of Korean and English mixed in a 1:1 ratio, the Korean tokens can be estimated to be approximately 30B.

to be limited. This allows us to examine the performance of language models that are not explicitly trained on Korean in the task of Korean text generation. For comparison, this study includes the base models of the Korean continually pre-trained models as Non-Korean-trained Models.

Llama 2 7B and Llama 2 13B (Touvron, Martin, et al., 2023) Released in 2023, Llama 2 is a large language model has been used as a base model for various open-source models so far. It is based on the transformer architecture and incorporates several improvements from subsequent research (GPT3, PaLM, GPTNeo). It was released in four sizes: 7B, 13B, 34B, and 70B. This study uses the 7B and 13B models. The training data consists of a new mix of publicly available online data, covering various language data and programming code data, with a total of 2.0T tokens. Approximately 90% of the data is in English, while Korean data accounts for 0.06%.⁹ The total vocabulary size is 32,000 tokens, and the context window size is 4,096.

SOLAR-10.7B (D. Kim et al., 2024) Released in 2023, SOLAR-10.7B a large language model trained based on the LLaMA 2 architecture. It takes the llama2 architecture as the base model and initializes it with the pre-trained weights of Mistral 7B (Jiang et al., 2023), then applies depthwise scaling and continual pre-training to achieve depth up-scaling (DUS). It has 10.7B parameters. Although the datasets for continual pre-training is not mentioned, considering the use of Mistral 7B’s pre-trained weights (although the Mistral 7B also does not disclose its pre-training data) and the datasets used for instruction and alignment tuning mentioned in the paper, it is speculated to be predominantly English data. The vocabulary size is 32,000, and the context window size is 4,096.

Yi 6B (AI et al., 2024) Released in 2023, Yi 6B is an English-Chinese bilingual model based on the llama 2 architecture with some modifications applied, pre-trained on English and Chinese data. Yi models were released in two sizes: 6B and 34B. The pre-training data consists of 3.1T tokens from various sources in English and Chinese. The ratio is approximately 75% English, 20% Chinese, and 5% code, etc.¹⁰ To produce high-quality pretraining data, a series of data cleaning pipelines were designed and applied. The vocabulary size is 64,000, and the models with context window sizes of 4k and 200k are provided. This study uses the model with a context window size of 4,096.

In summary, the language models to be used in this study and their features are as shown in Table 3.1.

3.2 Tasks

As the tasks to be performed in this study, we designed two Korean text generation tasks by referring to the methods from Muñoz-Ortiz et al. (2023) and Wendler et al. (2024).

⁹Thus, the number of Korean tokens can be indirectly estimated as 1.2B, which is 0.06% of 2T.

¹⁰Estimated from Figure 2 in AI et al. (2024)

Model	Base architecture**	Model size	Korean data size in pre-training	Vocabulary size	Context size	Language type
<i>Korean Monolingual Models</i>						
KoGPT2 base	GPT2	125M	40GB data	51,200	-	Mono.(Ko)
Ko-GPT-Trinity	GPT3	1.2B	35B tokens	51,200	-	Mono.(Ko)
Polyglot-Ko	GPT-NeoX	1.3B	213B tokens	30,003	2,048	Mono.(Ko)
		3.8B	219B tokens			
		5.8B	172B tokens			
		12.8B	167B tokens			
KoGPT	GPT3	6B	200B tokens	64,512	2,048	Mono.(Ko)
<i>Multilingual Models</i>						
XGLM	GPT3	564M	79GB data, 20B tkns	256,008	2,048	Multi.
		1.7B				
		4.5B				
		7.5B				
mGPT	GPT3	1.3B	3B-4B tokens*	100,000	512	Multi.
		13B				
<i>Korean Continually Pre-trained Models</i>						
Yi-Ko-6B	Yi	6B	30B tokens	78,464	4,096	Tri.(Ko,En,Zh)
Llama-2-Ko 7b	Llama 2	7B	40B tokens	46,336	4,096	Bi.(Ko,En)
Open-Llama-2-Ko 7b	Llama 2	7B	15B tokens	46,336	4,096	Bi.(Ko,En)
Llama-2-KoEN 13b	Llama 2	13B	30B tokens*	46,336	4,096	Bi.(Ko,En)
OPEN-SOLAR-10.7B	SOLAR	10.7B	15B tokens	46,592	4,096	Bi.(Ko,En)
SOLAR-KOEN-10.8B	SOLAR	10.8B	30B tokens*	46,336	4,096	Bi.(Ko,En)
<i>Non-Korean-trained Models</i>						
Yi 6B	Yi	6B	-	64,000	4,096	Bi.(En,Zh)
Llama 2 7b	Llama 2	7B	1.2B tokens*	32,000	4,096	Weak Multi.
Llama 2 13b	Llama 2	13B	1.2B tokens*	32,000	4,096	Weak Multi.
SOLAR-10.7B	SOLAR	10.7B	-	32,000	4,096	Mono.(En)*

*Estimated

**For models that utilize the architecture of other models like GPT3 and Llama2 but with additional modifications applied, we classified them as having their own architectures.

Table 3.1: Summary of language models used in this study.

3.2.1 Task 1: Text generation without reasoning (“zero-shot”)

To evaluate the basic Korean text generation ability of the model, we designed a text generation task that does not require reasoning. In this task, the first three words¹¹ of a Korean news article or column are provided as an input prompt, and the model is asked to generate the following text, from the beginning of a sentence. It requires the ability to create grammatically correct and semantically natural sentences based on the syntactic and semantic understanding of Korean sentences.

A specific example of input prompts for Task 1 is shown in Table 3.2.

¹¹The term ‘word’ here refers to ‘eojel(어절)’, which generally corresponds to a unit separated by spaces in Korean. As Korean is an agglutinative language, an ‘eojel’ often consists of a content word (e.g., noun, verb, adjective) and one or more functional elements (e.g., particles, endings). However, using terms like ‘word phrase’ or other similar expressions to refer to ‘eojel’ might unnecessarily complicate readers’ understanding, and it is not an essential distinction for this study. Therefore, in this study, we use the term ‘word’ to denote the Korean ‘eojel’ for simplicity.

3.2.2 Task 2: Text generation with reasoning (“few-shot”)

In order to involve not only the basic text generation ability but also the reasoning ability when generating Korean text, we designed a text generation task that requires reasoning. In this task, three examples¹² of headlines and the first part¹³ of Korean news articles or columns are provided, and a new headline and the first three words of the article are given as input prompts to generate text. The language model is asked to understand the relationship between the headlines and texts in the examples and complete the content that follows based on this understanding, given a new headline and the beginning of a sentence. It requires the ability to understand the relationship of the provided examples, infer appropriate content from the understanding, and generate grammatically correct and semantically natural sentences.

A specific example of input prompts for Task 2 is shown in Table 3.2. The original texts¹⁴ are in Korean, and English translations are provided here.

The relationship between Task 1 and Task 2 is not exactly a zero-shot and few-shot relationship. Task 1 is closer to a kind of language modeling task concerning the basic text generation ability of language models, as it provides only the first three words of the body text without a headline. On the other hand, Task 2 is more akin to a type of conditional text generation task, where the model is required to infer the relationship from the given headline-body text pair examples and generate appropriate output based on the new input of a headline and the first three words of the body text. However, since the format of the two tasks is similar to the general format of zero-shot and few-shot

¹²The decision to provide three examples was made to ensure that the input prompts fit within the context window size of the models to be tested. For the mGPT models, which has the context window size of 512, when input prompts exceeding the window size were provided (four examples are given), they generated nonsensical Korean text.

¹³In terms of character count, it averages around 170-180 characters, and in terms of token count, it varies depending on the tokenizer, but with tokenizers trained on Korean, it averages around 70-110 tokens.

¹⁴The original text and the sources for each article are as follows:

<北형제국 쿠바와 65년 만에 수교> 정부가 북한의 형제국인 쿠바와 외교관계를 수립했다. 1959년 교류가 단절된 지 65년 만이다. 외교부는 한국과 쿠바가 14일(현지시간) 미국 뉴욕에서 양국 유엔 대표부가 외교 공한을 교환하는 방식으로 공식 외교관계를 수립했다고 밝혔다. 우리나라의 193번째 수교국으로, 유엔 회원국 가운데 이제 시리아만 미수교국으로 남았다.

(허백운 기자, 서울신문, 2024.02.15, <https://n.news.naver.com/article/081/0003430425>)

<신선식품까지 판다...中 알리, 전방위 韓 공습> 초저가 공산품을 무기로 국내 시장을 빠르게 잠식하고 있는 중국 온라인 쇼핑 플랫폼 알리익스프레스가 신선식품 사업 진출을 준비 중인 것으로 확인됐다. 온라인 그로서리 전문가 영입을 진행하는 가운데 한국을 본격 공략하기 위해서는 시장 규모가 크고 반복 구매가 잦은 신선식품까지 영역을 확대해야 한다고 판단한 것으로 분석된다.

(이경운 기자, 서울경제, 2024.02.16, <https://n.news.naver.com/article/011/0004300661>)

<들리나요, 어린 누이의 귓속말> 이제 갓 걸음마를 떤 어린 동생이 울며 투정을 부리자, 누이가 무어라 말하며 어깨를 토닥인다. 누이라고는 하지만, 세상의 언어들을 얼마나 익혔을까 싶은 어린아이다. 그래도 누이는, 그 빈약한 언어 속에 동생을 달랠 수 있는 말 몇 마디를 품고 있었던가 보다. 엿들을 수 없는 누이의 말을, 사진이 들려준다.

(조병준 시인 원문, 박미경 류가현 관장, 중앙SUNDAY, 2024.02.17, <https://n.news.naver.com/mnews/article/353/0000047215?sid=110>)

<달에서 광고하는 시대> 어두운 밤하늘을 비추는

(정제혁 논설위원, 경향신문, 2024.02.25, <https://n.news.naver.com/mnews/article/032/0003280897?sid=110>)

Task	Input Prompt
1	“Illuminating the dark night sky”
2	<p><Establishing Diplomatic Ties with North Korea’s Brotherly Nation Cuba After 65 Years> The government has established diplomatic relations with Cuba, a brotherly nation of North Korea. It has been 65 years since exchanges were cut off in 1959. The Ministry of Foreign Affairs announced that South Korea and Cuba officially established diplomatic relations on the 14th (local time) in New York, USA, by exchanging diplomatic notes between the two countries’ UN missions. Cuba is the 193rd country to establish diplomatic ties with South Korea, leaving only Syria as the remaining country among UN member states without diplomatic relations.</p> <p><Even Selling Fresh Produce...China’s AliExpress, All-Out Invasion of Korea> It has been confirmed that AliExpress, the Chinese online shopping platform rapidly eroding the domestic market by leveraging ultra-cheap industrial products, is preparing to enter the fresh produce business. While recruiting online grocery experts, it is analyzed that AliExpress has determined that it needs to expand its business to fresh produce, which has a large market size and frequent repeat purchases, in order to fully target the Korean market.</p> <p><Can You Hear the Little Older Sister’s Whisper?> When her little sibling, who had just started toddling, cried and threw a tantrum, the older sister whispered something while patting the younger one’s shoulder. Although called an older sister, she is just a little child herself, making one wonder how much of the world’s languages she has learned. Nevertheless, it seems that the older sister had a few words in her limited vocabulary to comfort her younger sibling. The photo conveys the sister’s words that cannot be overheard.</p> <p><The Era of Advertising on the Moon> Illuminating the dark night sky”</p>

Table 3.2: Examples of the input prompts for text generation tasks.

prompts, and using familiar terms would be helpful for readers’ understanding, we will refer to them as zero-shot and few-shot tasks. Nevertheless, it should be noted that they do not perfectly align with the typical zero-shot and few-shot task relationship.

3.3 Dataset

We constructed a dataset to provide inputs for Korean text generation tasks. To handle both generation tasks, news articles and columns with headlines (or titles) were selected as the dataset.

To prevent the dataset from being included in the training data used for the language models to be tested, news articles and columns between February and March 2024, which is after the training of all the language models, were collected through web crawling from the news section of the Korean portal site Naver.¹⁵ The news articles were from 6 categories (Politics, Economy, Society, Life/Culture, IT/Science, World), and columns were included to provide a relatively diverse style of input prompts other than news articles. This was done to evaluate the model’s overall capability across a variety of

¹⁵<https://news.naver.com/>

topics and styles. The collected data features are as follows:

- Category
- Headline
- Body text (first approx.120 chars)¹⁶
- Link
- Press
- Date

From the collected data, a total of 1000 cases were sampled in a 6:4 ratio of news to columns to construct a dataset for experiments, as shown in Table 3.3.

	Category	Number of cases
News	Politics	100
	Economy	100
	Society	100
	Life/Culture	100
	IT/Science	100
	World	100
	Columns	400
	Total	1,000

Table 3.3: *Composition of the dataset for text generation.*

3.4 Hyperparameter search

When a preliminary investigation of text generation was conducted for each model, it was observed that there was a variance in the quality of the generated texts depending on the hyperparameter configuration of the models. We considered that it is fair to allow each model to generate the best quality text possible and to analyze those results, rather than setting the hyperparameters of all models uniformly. To achieve this, a hyperparameter search was conducted for each model.

3.4.1 Hyperparameter configurations

Regarding sampling, we decided to test both sampling and beam search approaches. For beam search, we set the beam size (*num_beams*) to 5, as preliminary tests showed no significant difference in results when testing beam sizes of 5 and 10. For sampling, among the parameters related to adjusting the sampling probability distribution, such as *top_k*, *top_p*, and *temperature*, we chose to test *temperature* as a representative parameter. Considering the ranges that showed better results for models in preliminary

¹⁶This is the amount of body text that can be scraped from the news section page displaying the news and columns.

tests, we decided to perform a grid search for temperature from 0.7 to 1.1, with an interval of 0.1. Next, as some results in the preliminary tests showed a tendency to repeat parts of the text, we decided to adjust *repetition_penalty*, one of the parameters related to repetition avoidance, along with *no_repeat_ngram_size*. Again, considering the ranges that showed better results for each model in preliminary tests, we decided to perform a grid search for *repetition_penalty* from 1.2 to 2.0, with an interval of 0.2. Accordingly, as shown in Table 3.4, we planned to test a total of 30 hyperparameter configurations for each model.

Configuration	do_sample	num_beams	top_p	temperature	repetition_penalty
1-25	True	1	0.95	0.7, 0.8, 0.9, 1.0, 1.1	1.2, 1.4, 1.6, 1.8, 2.0
26-30	False	5	None	None	1.2, 1.4, 1.6, 1.8, 2.0

Table 3.4: Hyperparameter configurations for Hyperparameter search.

3.4.2 Text generation for hyperparameter search

For the text generation, we provided the models with the first 10 words of the article body from the dataset as input and had the models generate the following text. By generating text for 10 cases from the dataset per model, consisting of 6 news articles and 4 columns, we aimed to evaluate the models’ performance more comprehensively, considering the diversity of topics and styles.

3.4.3 Evaluation of the generated text quality

To evaluate the quality of the generated texts, we attempted semantic similarity with the original text, use of LLM as a judge, and manual assessment.

(1) Semantic similarity

Since the first 10 words of the article body were provided as the input prompt, we considered that the model could understand the context to some extent, and the following text would have some degree of similarity with the content of the original article. Therefore, we decided that the semantic similarity between the original article and the generated text could be used as an indirect indicator for evaluating text quality.

To calculate the semantic similarity, we used a pre-trained multilingual sentence transformer model **paraphrase-multilingual-mpnet-base-v2** (Reimers and Gurevych, 2020¹⁷) that supports Korean for semantic similarity evaluation.

To verify whether semantic similarity is an appropriate metric for quality evaluation, we calculated the average semantic similarity scores of the texts generated by each model, as shown in Table 3.5. The results generally aligned with our expectation that models trained on Korean are more likely to generate texts with content similar to the original

¹⁷<https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2>

articles, as they can better understand the meaning and context of Korean texts, and models not trained on Korean may have difficulty properly understanding the meaning and context of Korean texts, resulting in relatively lower quality of the generated texts. This suggests that the semantic similarity measure can be used as a valid indicator for evaluating the quality of texts generated in this task.

Model	Type	Mean	Std
OPEN-SOLAR-KO-10.7B	Ko. Continually Pre-trained	0.765	0.117
open-llama-2-ko-7b	Ko. Continually Pre-trained	0.764	0.098
llama-2-koen-13b	Ko. Continually Pre-trained	0.762	0.089
polyglot-ko-12.8b	Ko. Monolingual	0.750	0.099
Yi-Ko-6B	Ko. Continually Pre-trained	0.749	0.113
polyglot-ko-3.8b	Ko. Monolingual	0.742	0.106
SOLAR-KOEN-10.8B	Ko. Continually Pre-trained	0.741	0.103
polyglot-ko-5.8b	Ko. Monolingual	0.740	0.104
polyglot-ko-1.3b	Ko. Monolingual	0.740	0.101
kogpt-6B	Ko. Monolingual	0.736	0.090
ko-gpt-trinity-1.2B-v0.5	Ko. Monolingual	0.733	0.102
llama-2-ko-7b	Ko. Continually Pre-trained	0.730	0.104
xglm-7.5B	Multilingual	0.715	0.101
xglm-4.5B	Multilingual	0.701	0.106
xglm-1.7B	Multilingual	0.690	0.097
xglm-564M	Multilingual	0.685	0.102
mGPT-13B	Multilingual	0.684	0.102
mGPT-1.3B	Multilingual	0.681	0.100
skt/kogpt2-base-v2-125M	Ko. Monolingual	0.666	0.112
Llama-2-7b	Non-Korean-trained	0.567	0.202
Yi-6B	Non-Korean-trained	0.566	0.191
Llama-2-13b	Non-Korean-trained	0.554	0.216
SOLAR-10.7B-v1.0	Non-Korean-trained	0.513	0.193

Table 3.5: Average semantic similarity scores of models for texts generated in hyperparameter search.

For each model, we calculated the mean and standard deviation of the semantic similarity scores for 10 cases across 30 hyperparameter configurations. From this, we extracted the top 5 hyperparameter configurations that obtained the highest average scores for each model, as shown in Figure 3.1.¹⁸

(2) LLM as a judge

We also attempted to use LLM as a judge (Zheng et al., 2023) to evaluate the quality of the generated texts. Although we finally decided not to use the LLM’s evaluation for quality assessment due to the lack of consistency and persuasiveness in the results, we briefly describe the process and results here.

¹⁸Among the top 5 configurations for each model, there were no configurations that employed beam search. Therefore, only the temperature (T) and repetition_penalty (R) parameter values are shown in the figure.

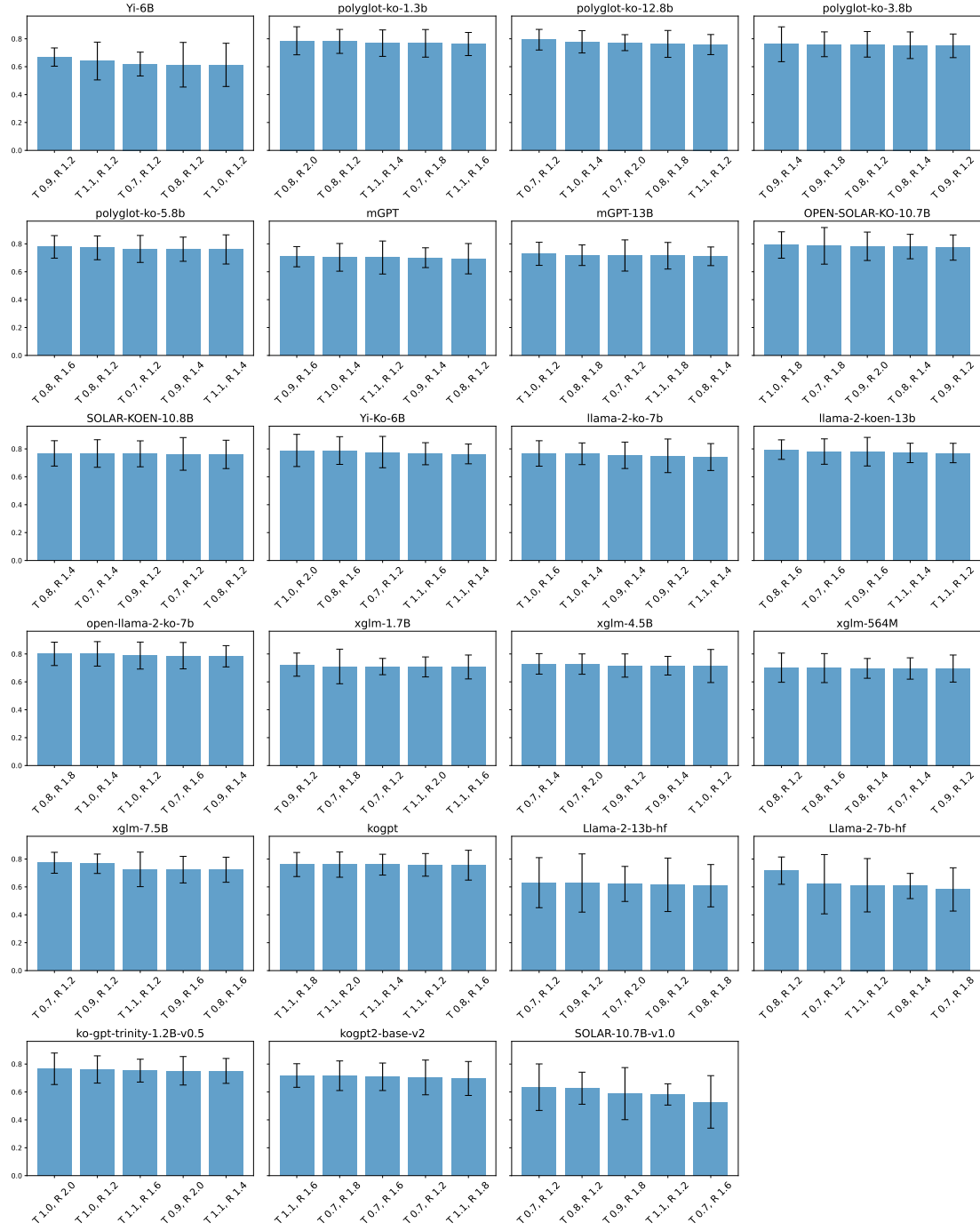


Figure 3.1: Top 5 hyperparameter configurations for each model based on semantic similarity scores of the texts generated in hyperparameter search.

We designed the input prompt for evaluation as follows,¹⁹ considering various aspects related to text quality evaluation. Refer to Figure A.1 for the original Korean prompt.

“As a linguist fluent in Korean and with expert knowledge of the language, you will evaluate the quality of the given Korean text. The first ten words are the given input prompt, and

¹⁹The original text was provided in Korean, and an English translation is provided here.

the following part is the text generated by the LLM. Focus your evaluation on the part generated by the LLM. Read the text carefully and review it thoroughly, taking sufficient time according to the following ‘Evaluation Criteria’.

Grammatical Accuracy: Evaluate whether the text is grammatically accurate.

Fluency: Evaluate whether the text is natural and fluent.

Clarity of Meaning: Evaluate whether the meaning of the text is clear and easy to understand.

Relevance: Evaluate whether the text is relevant to the context provided by the input text.

Creativity: Evaluate whether the text exhibits creativity and originality.

Coherence: Evaluate whether the connections within the text are natural and maintain a consistent flow.

Overall Evaluation: Evaluate the overall quality of the entire text.

Be sure to evaluate using the same consistent scale as in previous cases, and double-check that all evaluations are valid before answering. Provide the score for each item in the ‘Evaluation Criteria’ between 1 (low) and 5 (high), and output the answer only in the following format:

<Answer Example>

Grammatical Accuracy: 5

Fluency: 4

Clarity of Meaning: 4

Relevance: 4

Creativity: 3

Coherence: 3

Overall Evaluation: 4”

When testing several recent high-performance open-source LLMs (Mixtral-8x7B Instruct-v0.1²⁰, Gemma 1.1 7B (IT)²¹, Llama-2-Chat 70B²², OpenChat 3.5 (0106)²³, and Nous Hermes 2 Mixtral 8x7B DPO²⁴) with this prompt, the model that yielded the most desired output format and convincing results was Nous Hermes 2 Mixtral 8x7B DPO. We decided to use this model as the judge and had it assess the quality of the texts generated during the hyperparameter search. Due to the non-deterministic nature of LLM generation, we conducted two rounds of evaluations to ensure the consistency of the results.

When we listed the average scores obtained from each evaluation in descending order, it showed the models specifically trained on Korean were followed by multilingual models and then models not trained on Korean, which was generally similar to the results of semantic similarity. Like the semantic similarity, LLM as a judge also seems to discern the overall quality of the texts, whether good or bad.

However, the consistency of the evaluation results for detailed hyperparameter configurations was lacking, and upon manually reviewing the evaluation results, there were often outcomes that were not convincing. Therefore, we finally decided to exclude the use of LLM as a judge for the text quality assessment.

²⁰<https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1>

²¹<https://huggingface.co/google/gemma-1.1-7b-it>

²²<https://huggingface.co/meta-llama/Llama-2-70b-chat-hf>

²³<https://huggingface.co/openchat/openchat-3.5-0106>

²⁴<https://huggingface.co/NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO>

Nevertheless, we suppose that if we had evaluated using a language model fluent in Korean or employed different prompts, we might have obtained different results. We leave this as a topic for future research.

(3) Manual assessment

Semantic similarity scores cannot be said to necessarily correlate with quality of the generated texts, because they are an indicator of content similarity with the original text and do not fully reflect various quality evaluation criteria for the text generation, such as grammatical accuracy, fluency, and creativity. Additionally, considering the standard deviations in Figure 3.1, the rankings of hyperparameter configurations based on semantic similarity scores are not definitive. Therefore, in addition to semantic similarity, we decided to conduct manual evaluations as a supplementary measure for quality assessment.

Due to time constraints in reviewing the entire generated texts from 30 hyperparameter configurations for each model, we manually reviewed the texts generated from the top 5 hyperparameter configurations with the highest average scores based on semantic similarity, shown in Figure 3.1. Among the 5 hyperparameter configurations, the one that was deemed to generate the best overall quality texts was selected as the final hyperparameter configuration for that model. The results are shown in 3.6.

While we cannot definitively claim that these configurations are the best hyperparameter setups for each model, we consider them to be hyperparameter configurations that generate relatively better results among the tested hyperparameter configurations for each model. Therefore, we proceeded with the subsequent process using these configurations.

3.5 Text Generation

Using the dataset constructed in Section 3.3, we had the language models listed in Table 3.1 in Section 3.1 generate text for the two tasks described in Section 3.2. The hyperparameter configurations shown in Table 3.6 in Section 3.4 were used for each model. Additionally, for the hyperparameter “*new_max_tokens*” related to the length of the text to be generated, we set it to 256 for models trained on Korean (i.e., models with Korean vocabulary) and 512 for models not trained on Korean (i.e., models without Korean vocabulary). This is because models without Korean vocabulary process Korean text as byte-level tokens, requiring more tokens to represent the same length of Korean content.

For text generation, we used NVIDIA A100 on Saga,²⁵ the high performance computing (HPC) service of NRIS (Norwegian Research Infrastructure Services).

²⁵https://documentation.sigma2.no/hpc_machines/saga.html

Model	<i>do_ sample</i>	<i>top_p</i>	<i>temperature</i>	<i>repetition_ penalty</i>
kogpt2-base-v2			0.7	1.2
ko-gpt-trinity-1.2B-v0.5			1.0	2.0
polyglot-ko-1.3b			0.8	1.2
polyglot-ko-3.8b			0.8	1.2
polyglot-ko-5.8b			0.8	1.2
polyglot-ko-12.8b			0.7	1.2
kogpt			0.8	1.6
xglm-564M			0.8	1.2
xglm-1.7B			0.9	1.2
xglm-4.5B			1.0	1.2
xglm-7.5B			0.7	1.2
mGPT	True	0.95	0.8	1.2
mGPT-13B			0.7	1.2
Yi-Ko-6B			0.7	1.2
llama-2-ko-7b			0.9	1.2
open-llama-2-ko-7b			0.8	1.8
llama-2-koen-13b			0.7	1.6
OPEN-SOLAR-KO-10.7B			0.8	1.4
SOLAR-KOEN-10.8B			0.9	1.2
Yi-6B			0.9	1.2
Llama-2-7b			0.8	1.2
Llama-2-13b			0.7	1.2
SOLAR-10.7B-v1.0			0.7	1.2

Table 3.6: Hyperparameter configurations for each model derived from hyperparameter search.



3.6 Analysis: Generated texts

The text generation results were analyzed in two dimensions related to the research questions. Specifically, when various types of language models were asked to generate Korean text in Korean, we examined how the responses from the models differed, and how the Korean sentences generated by the models differed. This can be distinguished as an analysis of the generated texts and an analysis of the Korean sentences within the generated texts. This section describes the methods used to analyze the texts generated by various types of language models in response to Korean input prompts requesting Korean text generation.

Preprocessing: Removing input prompt In the texts generated from the few-shot task, the input prompt was included at the beginning of the generated texts. Therefore, the examples and a new headline provided in the input prompt were removed from the generated texts. The first three words following the headline were not removed because they are part of the first sentence. In the texts generated from the zero-shot task, the first three words given as the input prompt were also not removed because they are part of the first sentence.

3.6.1 Basic Statistics


The basic statistical characteristics of texts generated by each model for each task were investigated. These include the length of the generated text, the number of tokens generated, the types of characters and tokens generated, the number of sentences generated, and the length of sentences generated.

- **Length of the generated text:** Calculated as the length of the string.
- **Number of tokens generated:** Calculated as the number of tokens when the generated text was encoded by the tokenizer of the model.
- **Types of characters generated:** The characters of the generated texts were classified by type. They were classified into Hangul (Korean script), Latin (Latin alphabet), CJK (Chinese characters²⁶), DIGIT (numbers), Symbol, Whitespace, and Undecodable (represented as ²⁷). Python’s unicodedata library²⁸ was used for character script classification.
- **Types of tokens generated:** The tokens of the generated texts were classified by type. After encoding the generated text with the tokenizer of the model, each token was decoded to classify the token type. They were classified into Hangul (Korean script), Latin (Latin alphabet), CJK (Chinese characters), DIGIT (numbers), Symbol, Mixed (Different types of script are mixed within a token), and Undecodable (represented as ). Python’s unicodedata library was used for character script classification.
- **Number of sentences generated:** To examine the generated texts at the sentence level, the generated texts were divided into Korean sentence. The Korean morphological analysis library Kiwi (M. Lee, 2022) was used to split Korean sentences.²⁹ It is not easy to accurately split Korean sentences due to the frequent omission of sentence-ending punctuation, missing spaces, and the handling of quotations; thus, it should be noted that the results of the Korean sentence segmentation are not completely accurate.
- **Length of sentences generated:** Calculated as the length of the sentence string and the number of words (split by spaces) per a sentence in the separated Korean sentences.

3.6.2 Surface-level Evaluation

Next, we examined the sentences by considering the surface-level characteristics of the texts. This is also to filter out the Korean sentences that will be analyzed in Section 3.7. As formal characteristics of general Korean sentences, we investigated the ratio of Korean characters in the generated sentences, whether the sentence contains unusual patterns

²⁶CJK Unified Ideographs or CJK Compatibility Ideographs

²⁷Unicode replacement character ()

²⁸<https://docs.python.org/3/library/unicodedata.html>

²⁹In sentences containing a quotation ending with sentence-ending ‘da(다)’ followed by a conjunctive ‘myeo(며)’, the library separates them into two sentences. To maintain consistency with other sentences where a sentence containing a quotation is treated as a single sentence, an additional modification was made to connect the preceding and following sentences.

such as URLs or emails, and whether the sentence has a typical Korean sentence-ending format. These characteristics were selected through a preliminary investigation by reviewing sentences generated by various language models.

- **Korean character ratio per sentence:** The ratio of Korean characters (syllables) within each sentence was calculated. Excluding newline characters, the ratio of the number of Korean characters to the total length of the sentence was obtained.
- **Inclusion of unusual patterns in sentences:** Each sentence was examined whether it contained patterns that are not expected to be in typical Korean sentences. These patterns include URLs, email addresses, date formats, time formats, phone number formats, HTML tags, ellipsis notations, news bylines (e.g., [= Reporter xxx]), consecutive special characters (e.g., “*****”), various symbols, and emojis. Although the inclusion of these patterns does not necessarily disqualify a sentence as Korean sentence, sentences containing these patterns may not be suitable for understanding the characteristics of natural Korean sentences. Moreover, these patterns would be noise in the subsequent Korean sentence analysis, so they were intended to be filtered out. These patterns were collected through manual review of the actually generated texts. The results are presented as the ratio of the number of sentences containing unusual patterns to the total number of sentences.
- **Inclusion of the headline pattern provided in examples:** Among unusual patterns, we separately investigated the inclusion of the pattern “<>” used as a headline marker in the examples. The tasks did not request headline writing, so headlines or headline patterns were not expected in the generated text. However, the occurrence of headline patterns in the generated texts may reveal the influence of examples on text generation. Both cases where the headline pattern is fully included (“<...>”) and where only part of the headline pattern is included (“<” or “>”) were detected and analyzed. The results are presented as the ratio of the number of sentences containing the headline pattern to the total number of sentences.
- **Korean sentence-ending formats:** For each sentence, it was examined whether it contains sentence-ending punctuation marks (., ?, !) or ends with common Korean sentence-endings verb endings, such as “da(다)”, “yo(요)”, “jyo(쥬)”, etc. In casual language use, sentence-ending punctuation marks are often omitted in sentences, and there are Korean sentences that do not end with verb endings (e.g., sentences ending with noun-form endings). To accommodate this, if a sentence satisfies one of the two requirements—ending with sentence-ending punctuation marks or common Korean sentence-ending verb endings—it was considered to have a complete Korean sentence-ending format. Otherwise, it was regarded as an incomplete Korean sentence. Before detecting sentence-ending formats, to ensure accurate detection, we removed parentheses, symbols, Korean consonants or vowels that do not form complete characters, or consecutive periods that appear to be present due to the influence of sentence separation, colloquial influence, or characteristics of some generated texts. The results are presented as the ratio of the number of sentences that do *not* have typical Korean sentence-ending formats to the total number of generated sentences.

By applying these characteristics step by step, we extracted sentences that formally

conformed to the complete Korean sentence form. These extracted sentences are used in the analysis of the generated Korean sentences in Section 3.7. The steps taken are as follows:

- Extracted only sentences with a Korean character ratio of 0.4 or higher from the generated sentences.
- Extracted only sentences that did not contain unusual patterns within the sentence.
- Extracted only sentences that did not include headline patterns (< or >) within the sentence.
- After cleaning non-textual elements at the beginning and end of the sentence, extracted only sentences that had a typical Korean sentence-ending format.

The Korean sentences obtained through this filtering process were stored for each model and task. Additionally, we calculated the ratio of the number of these formally complete Korean sentences to the total number of generated sentences for each task and model to evaluate the models' ability to generate formally complete Korean sentences.

3.6.3 Semantic Evaluation

In addition to the surface-level evaluation, we aimed to perform a semantic evaluation of the texts generated by each model for each task.

- **Semantic embedding similarity:** For the semantic evaluation, we utilized the semantic embedding similarity comparison with the original article headlines and body texts in the dataset.

For text generation tasks, evaluating the quality of generated texts based on semantic similarity with the original text may have limitations. However, when the context is provided to some extent as a prompt, high semantic similarity with the original text can indicate that the generated text captures the topic and context of the provided prompt well. Therefore, we expected to indirectly assess the semantic quality of the generated texts through semantic similarity analysis.

For 1,000 cases in the dataset, we created semantic embeddings of the combined text of the headline and the first part of the body text (approximately 120 characters) as a reference for comparison. We also tried creating semantic embeddings of only the first part of the body text without the headline, but there was no significant difference between the two, so we used the combined text. For semantic embedding transformation, we used the Korean sentence embedding model KR-SBERT-V40K-klueNLI-augSTS (S. Park and Shin, 2021).³⁰

Next, for the generated texts of 1,000 cases for each tmodel and task, we created semantic embeddings as well using the same semantic embedding model. Considering the semantic embedding model is a sentence-based model, to ensure accurate transformation, we separated the generated text into sentences, embedded each sentence, and calculated the average embedding vector of these sentence embeddings.

The cosine similarity between the embedding vector of the original text and the

³⁰<https://huggingface.co/snunlp/KR-SBERT-V40K-klueNLI-augSTS>

embedding vector of the generated text was computed for each case. From this, the average similarity and standard deviation for the 1,000 cases were calculated.

- **Semantic visualization:** Furthermore, we visually examined the relationships between these semantic embedding vectors by visualizing them in the same low-dimensional embedding space using dimensionality reduction algorithms such as PCA (Jolliffe, 2002), t-SNE (Maaten and Hinton, 2008), and UMAP (McInnes et al., 2020). To compare the embeddings in the same embedding space, we created an embedding space using data that stacked embeddings from all models, including the original text. However, to distinguish the distribution of each embedding, we represented the embeddings of each model as separate subplots.

3.7 Analysis: Generated Korean sentences

This section describes the methods used to analyze Korean sentences generated by various language models, which were filtered at the surface level to conform to the typical Korean sentence format, as outlined in Section 3.6.2. For the analysis of sentences, various linguistic patterns used in Muñoz-Ortiz et al. (2023) for comparing human and LLM-generated texts were employed.

Preprocessing: Correcting Spacing Before conducting an analysis on Korean sentences, we corrected the spacing in the texts generated by the SOLAR-KOEN-10.8B model, which had issues with missing necessary spaces between words. This was made to evaluate the model’s generated texts without the influence of spacing issues, as we believed that the spacing problem was not inherent to the model itself. For spacing correction, the Korean morphological analysis library Kiwi’s spacing correction feature was used. Although not perfect, it automatically supplements spacing based on morphological analysis.

Additional Data for Comparison: Korean Corpus When analyzing Korean sentences generated by language models, Naturally occurring Korean sentences can be used as baselines for comparison. For this purpose, spoken and written Korean corpora were downloaded from a publicly available Korean corpus platform “Modu Corpus (모두의 말뭉치)” (National Institute of Korean Language, 2021).³¹ Since the original corpora were not separated into sentences, we applied the same sentence segmentation method used for separating the texts generated by the language models. 10,000 sentences from each corpus were extracted and used in the study.

3.7.1 Basic Statistics

As a basic statistic of the Korean sentences generated by each model for each task, the number of sentences and the length of sentences was examined.

³¹The written and spoken corpora from the NIKL Dependency-Parsed Corpus (v.2.0) were used. <https://kli.korean.go.kr/corpus/>

- **Number of sentences:** The number of sentences filtered for each model, including the Korean corpora, is presented. It can serve as a useful information when interpreting the evaluations of the sentences.
- **Length of sentences:** Calculated as the length of the sentence string and the number of words (split by spaces) per a sentence in the separated Korean sentences.

3.7.2 Lexical Evaluation

- **Lexical Diversity:** From a lexical perspective, we aimed to investigate the lexical diversity of Korean sentences generated by language models and in Korean corpora. To achieve this, the sentences were tokenized based on Korean morphemes, and the lexical diversity was calculated for the Korean tokens. The basic approach for examining lexical diversity would be the type-token ratio (TTR; the number of unique tokens/the number of total tokens). However, in this study, there was a large variation in the number of filtered sentences, so a simple comparison of type-token ratios could not serve as an appropriate measure of lexical diversity. This is because the type-token ratio tends to decrease as the text length increases. To mitigate the impact of text length, referring to Torruella and Capsada (2013), we calculated various lexical diversity measures, such as mean segmental type-token ratio (MSTTR; Johnson, 1944), moving average type-token ratio (MATTR; Covington and McFall, 2010), Root type-token ratio (RTTR; Guiraud, 1960), corrected type-token ratio (CTTR; Carroll, 1964), and measure of lexical diversity (MTLD; McCarthy, 2005). The methods for calculating the measures were based on the Python module Lexical Richness (Shen, 2021).³²

3.7.3 Syntactic Evaluation

For the syntactic evaluation of Korean sentences, the Universal Dependencies (UD) framework was utilized. UD is “a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages”.³³ It facilitates natural language processing research across various languages by annotating grammatical information and syntactic structures in a standardized format. To investigate the syntactic relations in Korean sentences generated by various language models, universal part-of-speech tags, language-specific part-of-speech tags, and dependency relation information provided by UD annotations were used in this study.

UD annotation data is usually stored in the CoNLL-U format, which includes word forms, lemmas, universal part-of-speech tags, language-specific part-of-speech tags, morphological features, and dependency relations with the head. The conversion of Korean sentences to UD annotations in CoNLL-U format was done using the Korean UD annotation training models provided by UDPipe.³⁴ UDPipe provides two Korean

³²<https://pypi.org/project/lexicalrichness/>

³³<https://universaldependencies.org/>

³⁴<https://lindat.mff.cuni.cz/services/udpipe/>

UD annotation models: one based on the GSD Treebank³⁵ and another based on the KAIST Treebank³⁶ (Chun et al., 2018). In this study, we tested both models, but as there were no substantial differences in the results and the Korean-specific POS tagsets differ slightly between the two, we only include the results from the model based on the GSD treebank here.

Additionally, when analyzing the converted UD annotation files, we also included the Korean UD treebanks used for training each UD annotation model to compare the differences with natural language.

It is worth noting that upon manual review of the UD annotation results, they were found to be not entirely accurate. Consequently, errors in the UD tagging may introduce some noise in the analysis of the results.

- **Universal part-of-speech tags (UPOS):** The distribution of UPOS in the Korean sentences generated by each model for each task was represented using cumulative bar plots. This was done to investigate whether there are differences in the general POS structure of Korean sentences generated by various language models.
- **Language-specific part-of-speech tags (XPOS):** The distribution of XPOS in the Korean sentences generated by each model for each task was represented using cumulative bar plots. This was also done to examine whether there were differences in the general part-of-speech tags of Korean sentences generated by various language models. Since XPOS has a more fine-grained composition compared to UPOS, it can be possible to observe the part-of-speech structure of the generated Korean sentences from multiple angles.
- **Universal dependency relations (deprel):** The distribution of dependency relation types in the Korean sentences generated by each model for each task was represented using cumulative bar plots. This was done to investigate whether there were differences in the dependency relations between components of Korean sentences generated by various language models.
- **Dependency arc:** The direction of the dependency arcs in the Korean sentences generated by each model for each task, the length in each direction, and the average length were represented using bar plots. This was done to examine whether there were differences in the direction and length of dependency relations in the sentences generated by various language models. Korean is a language with a subject-object-verb word order, and the word being modified generally comes after the modifier, resulting in a tendency for dependency arcs to point towards the left.

3.7.4 English Translationese

‘Translationese’ is a term coined by Gellerstam (1986) to compare the differences between texts originally written in a language and those translated from another language. It refers to the nuances of a text that exhibit excessive literal translation, reflecting the characteristics of the source language, such as vocabulary, syntax, and word order. To

³⁵https://universaldependencies.org/treebanks/ko_gsd/index.html

³⁶https://universaldependencies.org/treebanks/ko_kaist/index.html

detect the influence of English translationese in the generated Korean sentences, we examined several grammatical features related to the differences between Korean and English.

- **Articles:** In English, articles are one of the essential components of a sentence, but they do not exist in Korean (T. Park, 2004). We examined the frequency of the words ‘han(한)’ and ‘geu(그)’, which are commonly used when translating the English articles ‘a/an’ and ‘the’ into Korean.
 - ‘han(한)’: The frequency of the word with the form ‘한’ as a numeric modifier (dependency relation type: nummod) is counted and presented as a ratio to the total number of nouns.
 - ‘geu(그)’: The frequency of the word with the form ‘그’ as a determiner (UPOS tag: DET) is counted and presented as a ratio to the total number of nouns.
- **Singular/Plural:** While singular/plural distinctions are naturally made in English, they are not essential in Korean (MacDonald and Carroll, 2018). We investigated the frequency of the word ‘deul(들)’, which is commonly used when translating English plural forms into Korean.
 - ‘deul(들)’: The frequency of words with the form ‘들’ as a numeric modifier (XPOS tag: XSN (noun-deriving suffix)) is counted and presented as a ratio to the total number of nouns.
- **Passive Voice:** Compared to English sentences, passive voice is relatively less frequently used in Korean sentences (Y. Lee, 2000). The frequency of passive subjects (dependency relation type: nsubj:pass) is counted and presented as a ratio to the total number of sentences.
- **Pro-drop:** Korean is one of the pro-drop languages where subjects or objects can be omitted. Unlike English, where essential sentence components are generally not omitted, Korean often allows for the omission of subjects or objects when they can be understood from the context (Xia et al., 2000; C. Park, 2012).
 - Sentences without Subjects: The frequency of sentences without a subject is counted and presented as a ratio to the total number of sentences.
 - Sentences without Objects: The frequency of sentences without an object is counted and presented as a ratio to the total number of sentences.
 - Sentences without Both Subjects and Objects: The frequency of sentences without both subject and object is counted and expressed as a ratio to the total number of sentences.
- **Word Order:** One of the fundamental differences between Korean and English is that English follows a Subject-Verb-Object word order, while Korean follows a Subject-Object-Verb word order (Xia et al., 2000).
 - Sentences with Object-Verb Word Order: The number of sentences with an object-verb order is counted and presented as a ratio to the number of sentences that contain both an object and a verb.
 - Sentences with Verb-Object Word Order: The number of sentences with a verb-object order is counted and presented as a ratio to the number of sentences that contain both an object and a verb.

3.7.5 Semantic Evaluation

Evaluations at the semantic level are crucial in assessing the quality of the generated Korean sentences. However, due to time and resource constraints, we were unable to perform extensive semantic evaluations. In this study, as one of the semantic-level evaluations, we attempted sentiment classification on the generated Korean sentences.

- **Sentiment Classification:** Using a model³⁷ trained for Korean sentiment classification, we classified the Korean sentences generated by the language models and those from the Korean corpora into positive, negative, and neutral categories, and presented the distribution of their proportions.

³⁷<https://huggingface.co/dudcjs2779/sentiment-analysis-with-klue-bert-base>

This model is based on a pre-trained BERT model for Korean (S. Park et al., 2021) and has been fine-tuned for sentiment classification using a review dataset.

Chapter 4

Results

This chapter presents and analyzes the experimental results. First, the texts generated by various language models are evaluated and analyzed in terms of the basic statistics, surface-level, and semantic aspects. Next, the Korean sentences filtered from the generated texts are evaluated and analyzed in terms of basic statistics, lexical, syntactic, semantic, and English translationese aspects. Through this analysis, we aim to develop a comprehensive understanding of the Korean text generation capabilities of different language model types, considering various dimensions.

4.1 Analysis for generated texts

4.1.1 Basic Statistics

Text length

The length of the generated text for each model and task is shown in Figure 4.1. For the table of generated text lengths on which Figure 4.1 is based, refer to Table A.1 in the appendix.¹ Note that non-Korean-trained models, which do not have Korean vocabulary, were generated with the *max_new_tokens* parameter set to 512, while the other models that are trained on Korean and have Korean vocabulary were generated with the *max_new_tokens* parameter set to 256.

Token count

The number of tokens generated by task and model is shown in Figure 4.2. For the underlying table, refer to Table A.2 in the appendix.

Looking at Figure 4.1 and Figure 4.2, some models (KoGPT2-base-v2, XGLM-564M,

¹For ease of comparison, size information was appended to the names of models that did not have size information in their original names.

4.1. Analysis for generated texts

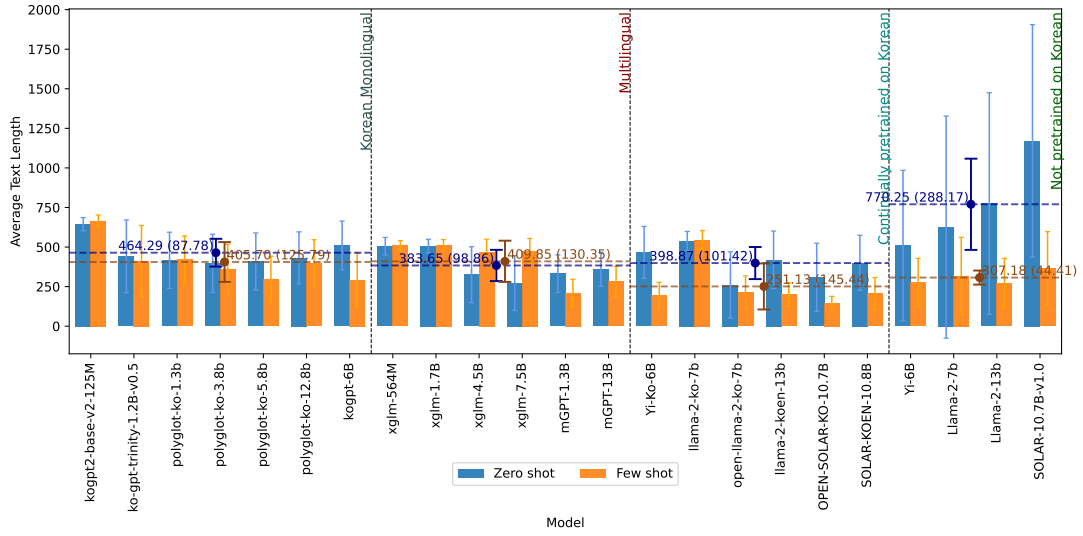


Figure 4.1: Average length of texts generated by language models. The four segments in the plot represent Korean monolingual models, multilingual models, Korean continually pre-trained models, and non-Korean-trained models, respectively. Blue indicates results in the zero-shot setting, while orange indicates results in the few-shot setting. The height of the bars represents the mean value, and the vertical lines denote the standard deviation. The horizontal lines in each segment represent the mean of each language model group, with the numeric value on the line indicating the mean and standard deviation (in parentheses).

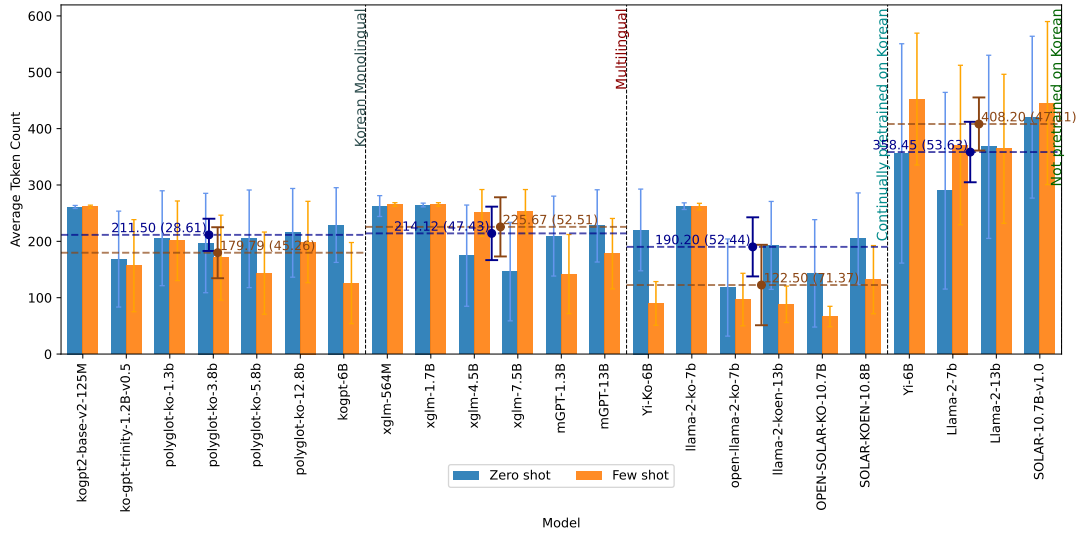


Figure 4.2: Average number of tokens generated by language models.

XGLM-1.7B, and Llama-2-ko-7b) tend to generate the maximum number of tokens specified by `max_new_tokens`, while other models show some variation within the `max_new_tokens` range.

Korean monolingual models, mGPT models, and Korean continually pre-trained models tend to generate fewer tokens in the few-shot task compared to the zero-shot task.

On the other hand, XGLM-4.5B, XGLM-7.5B and most language models not trained on Korean, except for Llama-2-13b, tend to generate more tokens in the few-shot task compared to the zero-shot task.

Furthermore, Figure 4.1 and Figure 4.2 show that the number of generated tokens and the text length are not necessarily proportional. This may be due to differences in generated tokens across models, such as generating more subword-level tokens than character-level tokens or byte-level tokens. For example, KoGPT2-base-v2, XGLM-564M and XGLM-1.7B generally produce tokens up to the given *max_new_tokens* number of tokens, but the average length of the generated text is longer for KoGPT2-base-v2. This implies that the average token length generated by the KoGPT2-base-v2 model is longer, which means it contains more subword-level tokens. Also, the models that were not trained on Korean, such as Yi-6B, Llama-2-7b, Llama-2-13b, and SOLAR-10.7B, generate around 400 tokens in the few-shot task, but the generated text length is around 250. This implies that the tokens generated by these models include many byte-level tokens smaller than characters.

Character types

The distribution of character types generated by language models for each task is shown in Figure 4.3 and Figure 4.4. For the tables of generated character types on which Figure 4.3 and Figure 4.4 are based, refer to Table A.3 and Table A.4 in the appendix.

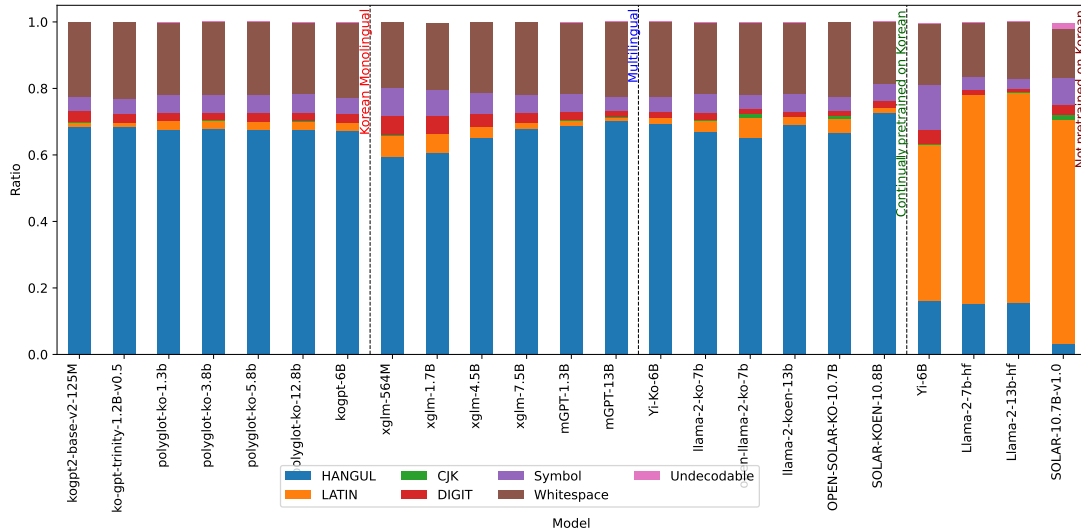


Figure 4.3: Character type distribution of the texts generated language models in the zero-shot task. The proportion of each character type is represented by a different color in the stacked bar graph. Note that the total ratio for each model does not equal, 1 because only the major types are shown.

According to Figure 4.3 and Figure 4.4, models trained on Korean (Korean monolingual models, multilingual models, and Korean continually pre-trained models) generally generate text in Korean when given Korean input prompts. The XGLM-564M, XGLM-1.7B models, and open-llama-2-ko-7b show a slightly higher tendency to generate Latin alphabets compared to other models. Although the difference between zero-shot task

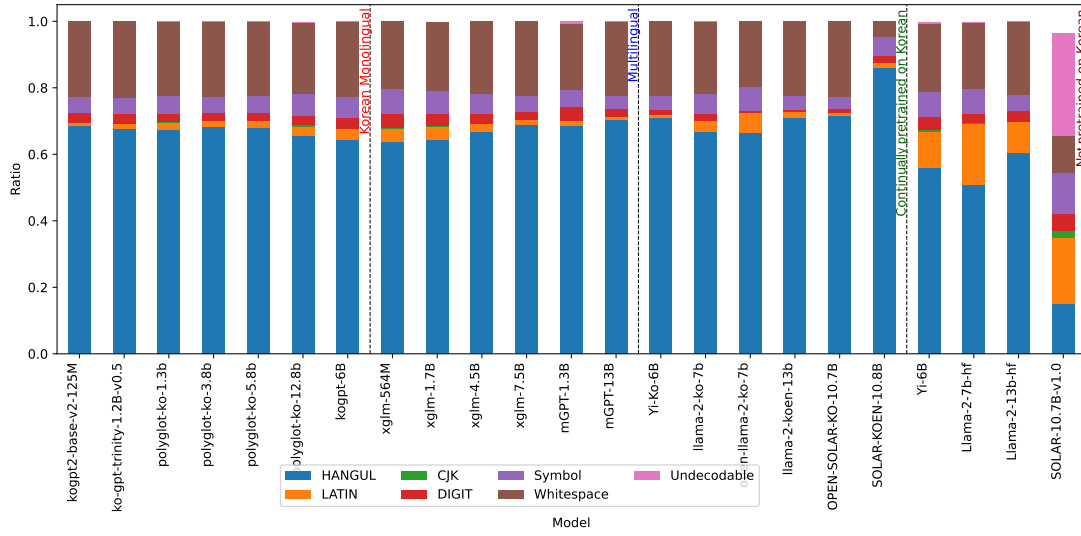


Figure 4.4: Character type distribution of the texts generated language models in the few-shot task.

and few-shot task is not significant, the proportion of Latin alphabet generation slightly decreases, and the proportion of Korean generation slightly increases in the few-shot task for multilingual models and Korean continually pre-trained models. Interestingly, in the few-shot task, KoGPT shows a slightly higher proportion of symbol and Latin alphabet generation, and polyglot-ko-12.8b shows a slightly higher proportion of symbol and Chinese character generation. The high proportion of Korean character generation in the SOLAR-KOEN-10.8B model in the few-shot task is due to the generated text not including most of the spaces that should have been there. The reason why this tendency is less pronounced in the zero-shot task is unclear.

Models that are not trained on Korean show distinct patterns from other models. In the zero-shot task, the generated text has a high proportion of Latin alphabet characters. In the few-shot task, however, the proportion of Latin alphabet generation significantly decreases, and the proportion of Korean character generation increases considerably. Although the SOLAR-10.7B model did not show a substantial increase in the generation ratio of Korean characters even in the few-shot task, it is speculated that the increase in the ratio of the undecodable characters may be related to the attempt to understand and generate Korean characters.

Token types

The distribution of token types generated by each model for each task is shown in Figure 4.5 and Figure 4.6. For the tables of generated token types on which Figure 4.5 and Figure 4.6 are based, refer to Table A.5 and Table A.6 in the appendix.

According to Figure 4.5 and Figure 4.6, models trained on Korean generally generate tokens in Korean when given Korean input prompts. However, there are slight differences between the zero-shot and few-shot tasks across different models.

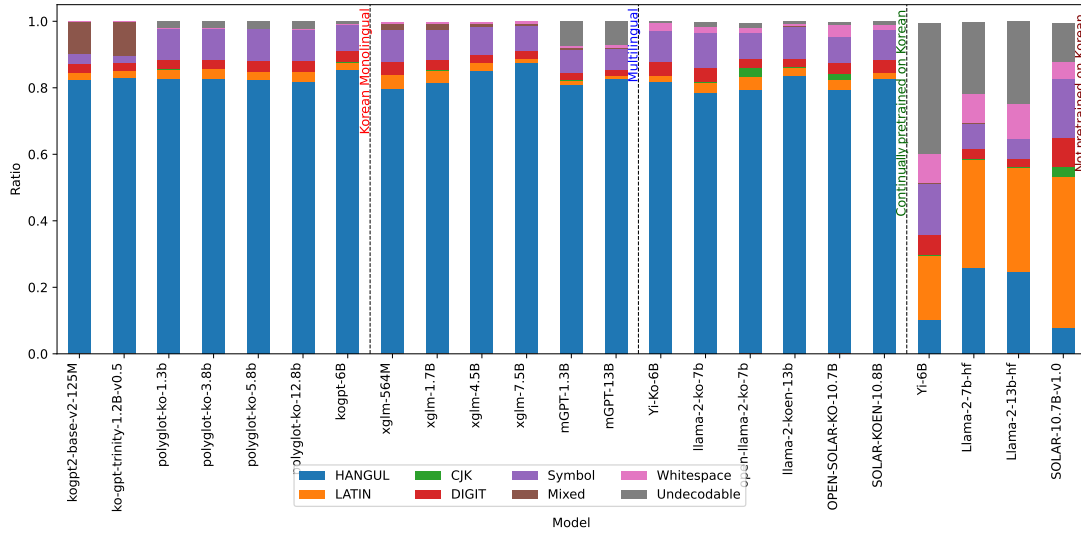


Figure 4.5: Token type distribution of the texts generated by each model in the zero-shot task.

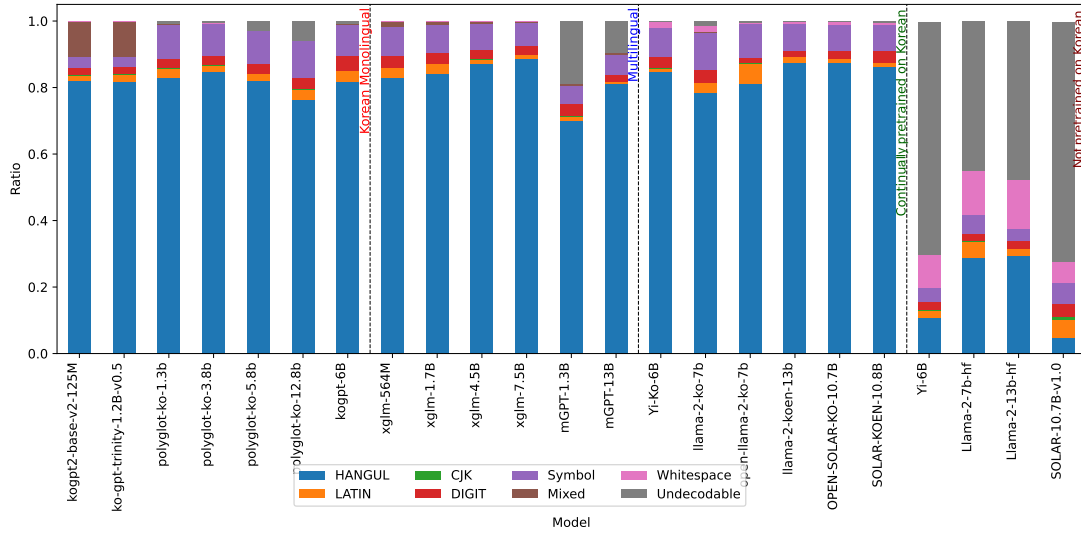


Figure 4.6: Token type distribution of the texts generated by each model in the few-shot task.

For the Korean monolingual models, KoGPT2-base and Ko-GPT-Trinity show almost no difference between the zero-shot and few-shot tasks. Ko-gpt-Trinity has a very slight increase in the proportion of symbol tokens in the few-shot task, which may be due to the ‘<>’ symbols included in the headlines of the few-shot task examples. Both models also have a characteristic of having a relatively high proportion of mixed tokens, compared to other models. The Polyglot-Ko models, although from the same model family, do not show consistent differences between the two tasks. In the zero-shot task, these models exhibit almost similar token generation distributions, but in the few-shot task, there are some differences in their responses. The 3.8B model shows an increase in the proportion of Korean tokens, while the largest model, the 12.8B model, shows a slight decrease in Korean tokens and an increase in the proportion of undecodable and symbol tokens. The KoGPT model also tends to have a slight decrease in Korean tokens in the few-shot task.

For the multilingual models, the XGLM models generally show a slight increase in the proportion of Korean tokens in the few-shot task. In the mGPT models, on the other hand, the 1.3B model shows a decreased proportion of Korean tokens and an increased proportion of undecodable tokens in the few-shot task. Actually, reviewing the text generated by mGPT-1.3B in the few-shot task reveals that many of the combined Korean characters often do not make sense.

For the Korean continually pre-trained models, the proportion of Korean token generation generally increases slightly in the few-shot task, except for the Llama-2-ko-7b model. The models trained on open access Korean data, such as the open-llama-2-ko-7b model and the OPEN-SOLAR-Ko-10.7B model, show decreases in the proportion of Chinese characters.

For the models not trained on Korean, they show noticeable differences between the zero-shot task and the few-shot tasks, just as in character type distribution. In the zero-shot task, all four models have a considerable proportion of Latin alphabet tokens, but in the few-shot task, the proportion of Latin alphabet tokens significantly decreases, and the proportion of tokens classified as undecodable greatly increases. These undecodable tokens are byte-level tokens, which are often combined into Korean characters in the actual generated text, as seen in Figure 4.4. It can be speculated that since these models have very few Korean tokens, they tried to generate Korean characters by combining byte-level tokens.

Sentence count

The number of sentences generated by each model for each task is shown in Table 4.1. As seen in Figure 4.1, since the total length of the generated text varies by model, it is fairer and more informative to examine the ratio of the number of sentences to the total text length, as shown in Figure 4.7, rather than simply comparing the number of sentences.

Excluding the models not trained on Korean, the ratio of the number of sentences does not differ significantly.

Among the Korean monolingual models, KoGPT2-base and Ko-GPT-Trinity-1.2B are relatively higher, while in the Korean continual pre-trained models, the open-llama-2-ko-7b model is a bit lower, and the SOLAR-KOEN-10.8B model is a bit higher.

In the Korean monolingual models, the ratio of the number of sentences is slightly higher in the few-shot task compared to the zero-shot task.

Conversely, in the multilingual models, the ratio of the number of sentences is slightly higher in the zero-shot task compared to the few-shot task.

For models not trained on Korean, especially in the zero-shot task, as the ratio of Korean text generation itself is not high and is incomplete, it would not be appropriate to meaningfully consider the number of Korean sentences. Nevertheless, it is noteworthy that in the few-shot task, the ratio of the number of sentences for the YI-6B, Llama-2-7b, and Llama-2-13b models, which have increased Korean text generation ratios, has

Model	Zero-shot			Few-shot		
	Sent. Count	Total Text Len.	Prop. (%)	Sent. Count	Total Text Len.	Prop. (%)
<i>Korean Monolingual Models</i>						
kogpt2-base-v2	10,681	644,873	1.66	11,956	664,539	1.80
ko-gpt-trinity-1.2B-v0.5	7,320	441,053	1.66	7,830	412,148	1.90
polyglot-ko-1.3b	5,925	416,102	1.42	6,074	420,803	1.44
polyglot-ko-3.8b	5,909	396,876	1.49	5,721	358,839	1.59
polyglot-ko-5.8b	6,221	409,077	1.52	4,731	293,732	1.61
polyglot-ko-12.8b	6,242	431,676	1.45	5,911	397,772	1.49
kogpt	7,623	510,168	1.49	4,599	292,196	1.57
<i>Multilingual Models</i>						
xglm-564M	7,516	505,381	1.49	7,381	509,155	1.45
xglm-1.7B	7,575	506,556	1.50	7,711	513,346	1.50
xglm-4.5B	5,236	325,602	1.61	6,961	469,243	1.48
xglm-7.5B	4,802	272,752	1.76	7,129	475,935	1.50
mGPT	5,252	331,550	1.58	3,257	209,382	1.56
mGPT-13B	6,047	359,860	1.68	4,597	282,140	1.63
<i>Korean Continually Pre-trained Models</i>						
Yi-Ko-6B	8,208	466,169	1.76	3,407	195,189	1.75
llama-2-ko-7b	9,089	537,698	1.69	9,260	544,250	1.70
open-llama-2-ko-7b	3,707	260,659	1.42	2,975	212,986	1.40
llama-2-koen-13b	6,993	419,342	1.67	3,150	200,676	1.57
OPEN-SOLAR-KO-10.7B	5,174	308,816	1.68	2,495	147,603	1.69
SOLAR-KOEN-10.8B	7,609	400,490	1.90	4,302	206,126	2.09
<i>Non-Korean-trained Models</i>						
Yi-6B	7,846	509,516	1.54	4,423	274,528	1.61
Llama-2-7b	4,716	625,940	0.75	4,252	317,377	1.34
Llama-2-13b	5,595	775,026	0.72	4,220	270,985	1.56
SOLAR-10.7B-v1.0	10,265	1,170,566	0.88	1,963	365,755	0.54

Table 4.1: Number of sentences from text generated by language models after sentence segmentation. Since the total text length generated differs across models, the proportion relative to the total text length (%) is also shown.

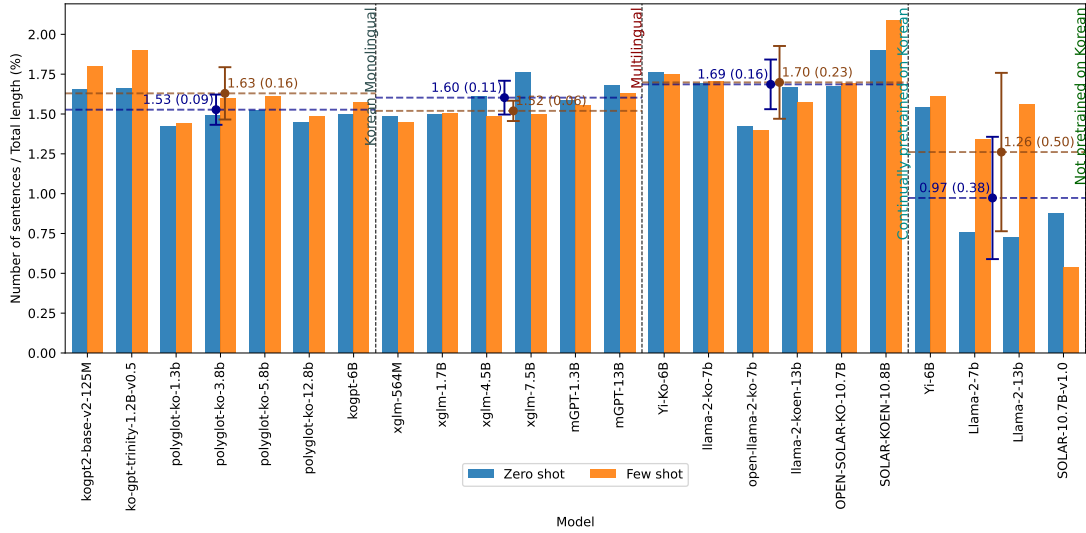


Figure 4.7: Proportion of the number of sentences to the total length of the texts generated by language models

become similarly high compared to other Korean-trained models.

The ratio of the number of sentences is also related to the sentence length, which will be examined next.

Sentence length

In terms of the length of sentences generated, Figure 4.8 and Figure 4.9 show the average text length and the average number of words per sentence generated by each model for each task. For the table of generated sentence lengths on which Figure 4.8 and Figure 4.9 are based, refer to Table A.7 in the appendix.

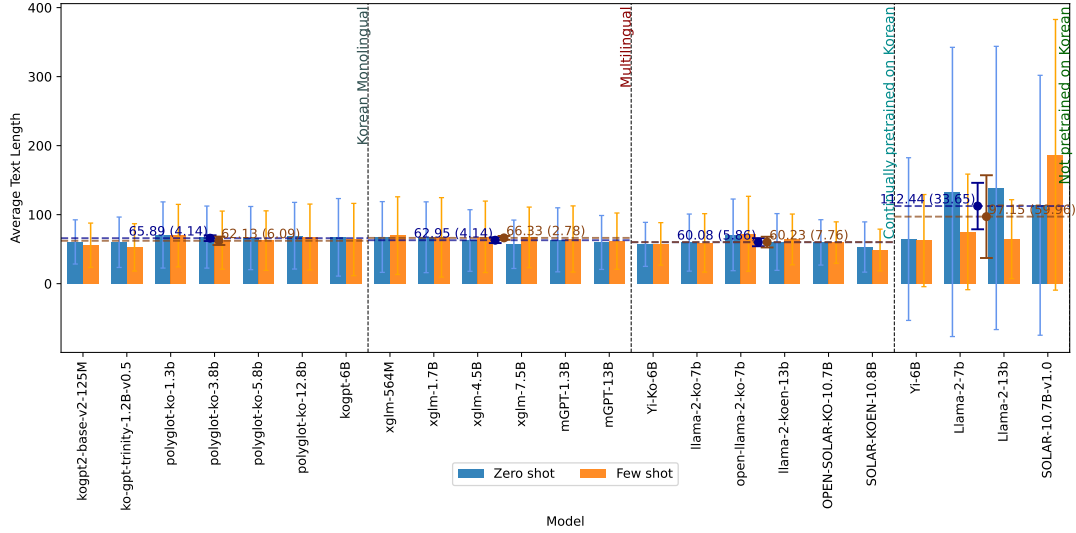


Figure 4.8: Average text length per sentences generated by language models.

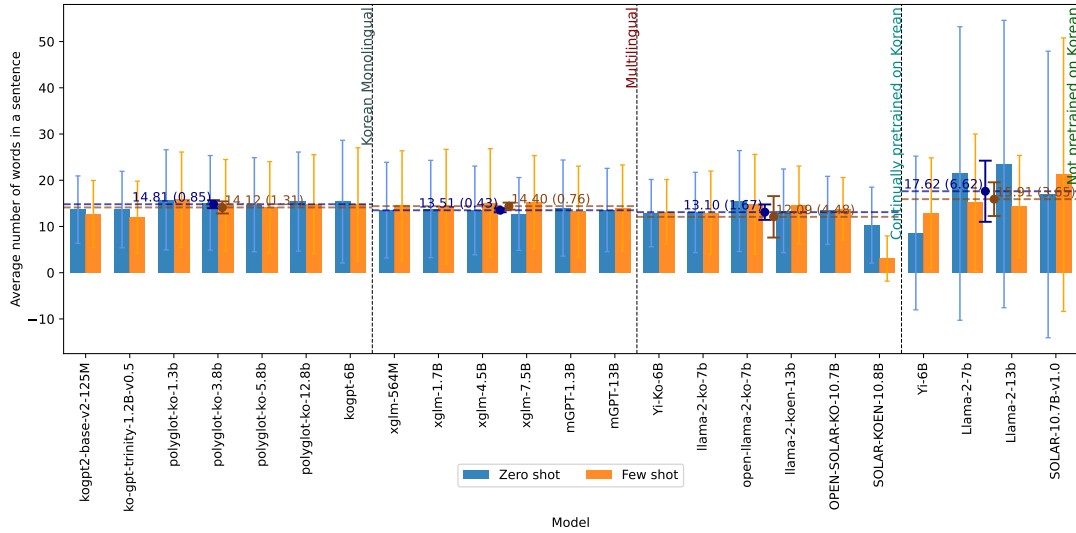


Figure 4.9: Average number of words per sentence generated by language models.

As shown in Figure 4.8 and Figure 4.9, the average text length and the average number of words per sentence are generally proportional. The relatively low number of words for the SOLAR-KOEN-10.8B model, especially in the few-shot task, is due to the issue of generated text by the model not including spaces, as mentioned earlier. Additionally, the high decrease in the number of words to the text length for the SOLAR-10.7B model in the few-shot task suggests that the number of characters within a word is excessively high, indicating unusual sentences.

Within the total text length, sentence length is inversely proportional to the number of sentences. Therefore, contrary to the number of sentences, Korean monolingual models tend to have shorter sentence lengths in the few-shot task, while multilingual models tend to have shorter sentence lengths in the zero-shot task, except for the mGPT-1.3B model. The KoGPT2-base and Ko-GPT-Trinity-1.2B models tend to be slightly shorter, and the open-llama-2-ko-7b model tends to be slightly longer, but overall, the models trained on Korean generate sentences with an average of around 10 to 15 words.

For models not trained on Korean, it would not be appropriate to compare these models together with other models, especially in the zero-shot task where the Korean text generation was not high. However, it is noteworthy that in the few-shot task, the average number of words for the YI-6B, Llama-2-7b, and Llama-2-13b models, which have increased Korean text generation, is similar to other Korean-trained models.

4.1.2 Surface-level Evaluation

Korean character ratio per sentence

Figure 4.10 shows the box plot of the Korean character ratio in the generated sentences for each model and task.

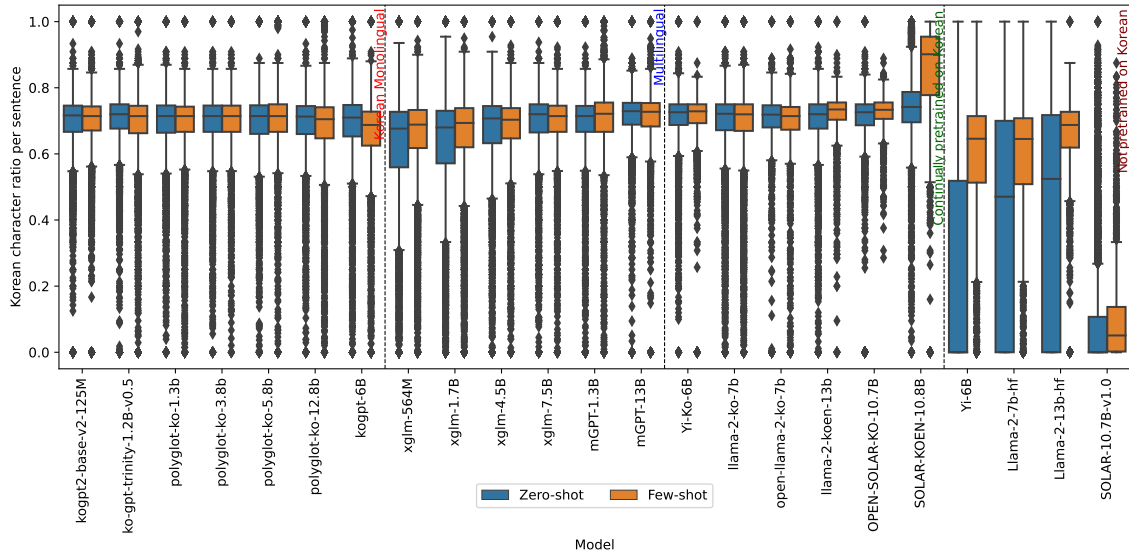


Figure 4.10: Distribution of Korean character ratio per sentence in the texts generated by the language models. It is represented using box plots. The box shows the range of the middle 50% of the data (from $Q1$ (25th percentile) to $Q3$ (75th percentile)), and the horizontal line inside the box represents the median of the data. The whiskers extending from the top and bottom of the box indicate the typical range of the data (from $Q1 - 1.5 \times (Q3 - Q1)$ to $Q3 + 1.5 \times (Q3 - Q1)$). Points outside the whiskers are marked as outliers.

For most models trained on Korean, it can be observed that the generated sentences have an average Korean character ratio of around 0.7. Considering that the total sentence length includes spaces, this ratio is thought to be typical for Korean sentences.

The high Korean character ratio of the SOLAR-KOEN-10.8B model reveals the issue of the absence of spaces again.

Given that the box represents the distribution of the middle 50% of the data, it can be seen that among the multilingual models, the XGLM-564M and XGLM-1.7B models have a slightly lower Korean character ratio in their generated sentences, especially in the zero-shot task. Among the Korean monolingual models, the KoGPT model shows a slightly lower Korean character ratio in its generated sentences, in the few-shot task.

Non-Korean-trained models show a larger variance in the Korean character ratio in their generated sentences compared to the models trained on Korean, but they exhibit substantial differences between the zero-shot and few-shot tasks. The Yi-6B, Llama-2-7b, and Llama-2-13b models have a Korean character ratio ranging from 0 to 0.7 for the middle 50% of the data in the zero-shot task, but it considerably increases to around 0.5 to 0.7 in the few-shot task. Especially, the Llama-2-13b model has a higher Korean character ratio and a narrower variance than other two models. The SOLAR-10.7B model has a Korean character ratio of less than 0.1 in both the zero-shot and few-shot tasks.

Unusual patterns in sentences

Figure 4.11 shows the ratio of sentences containing unusual patterns (such as URLs, email addresses, date formats, time formats, phone number formats, HTML tags, ellipsis notations, news bylines, consecutive special characters, various symbols, and emojis) in sentences generated by each model for each task. For the underlying table, refer to Table A.8 in the appendix.

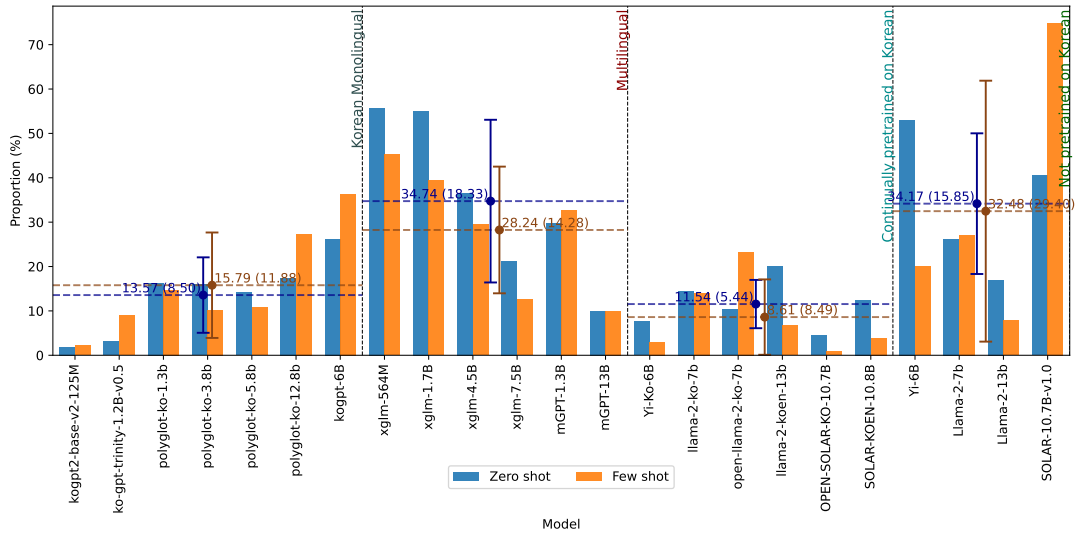


Figure 4.11: Proportion of sentences containing unusual patterns to the total sentences generated by language models.

Since the input prompts given to the tasks do not contain these patterns (except for one example headline in the few-shot task that contains an ellipsis notation "..."), these patterns are presumed to be influenced by the pre-training Korean data of the models.

The ratio of sentences containing these patterns is generally higher in multilingual models. The XGLM and mGPT model family, estimated to have been trained on the same data respectively, show a decrease in the ratio of sentences with these patterns as the model size increases. Among the Korean monolingual models, the Polyglot-Ko-12.8b and Ko-GPT models show relatively high ratios, especially in the few-shot task, despite their large size. The composition of their pre-training data may have influenced this. For the Korean continually pre-trained models, the ratios are generally low, with the open-llama-2-ko-7b model being a bit higher in the few-shot task and the llama-2-koen-13b model in the zero-shot task. For non-Korean-trained models, the Yi-6B model has a high ratio in the zero-shot task, and the SOLAR-10.7B model also shows high ratios, particularly in the few-shot task.

Headline pattern provided in examples

Figure 4.12 and Figure 4.13 show the ratio of sentences containing the pattern “<...>”, which was used as the headline marker in the example, among the sentences generated by each model for each task. For the underlying tables, refer to Table A.9 and Table A.10 in the appendix. Figure 4.12 shows the ratio of sentences that include the headline marker pattern as it is, in the form of “<...>”, while Figure 4.13 represents the ratio of sentences that contain at least one of “<” and “>”, taking into account the possibility of incomplete sentence separation or text generation.

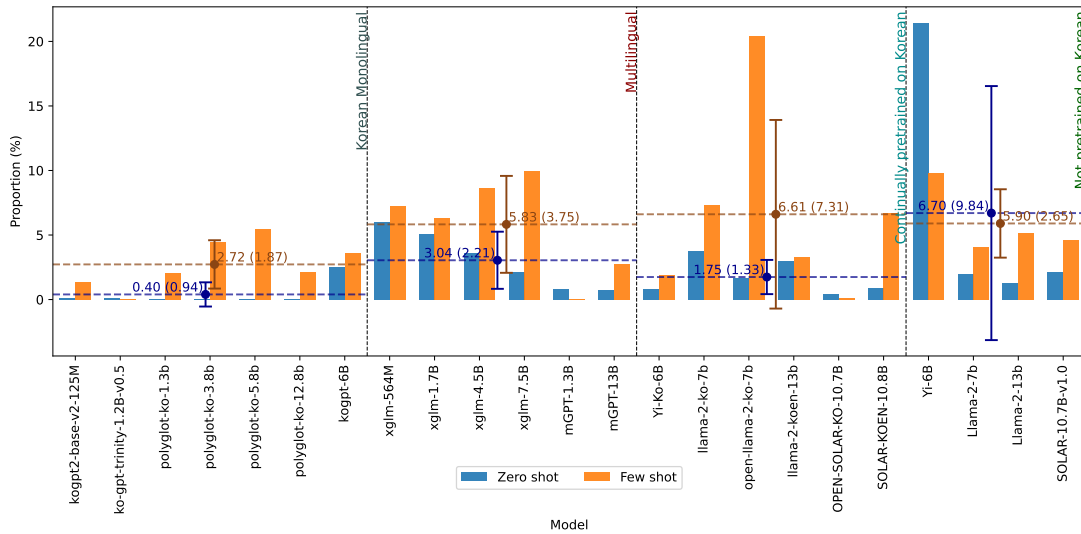


Figure 4.12: Proportion of sentences containing the headline pattern “<...>” provided in examples to the total sentences generated by language models.

In the zero-shot task, examples including headlines were not provided, so the headline pattern would not have had an influence. Therefore, the ratio in the zero-shot task can be considered as a reference for the model, indicating how much the model generates such patterns naturally without external influence.² In the few-shot task, where headlines

²The high percentage of the “<...>” pattern in the zero-shot setting for the Yi-6B model is due to the HTML tags in the texts generated by this model. The headline patterns were detected as “<...>”, so the format of HTML tags is also included. For detecting HTML tags in the unusual patterns, the

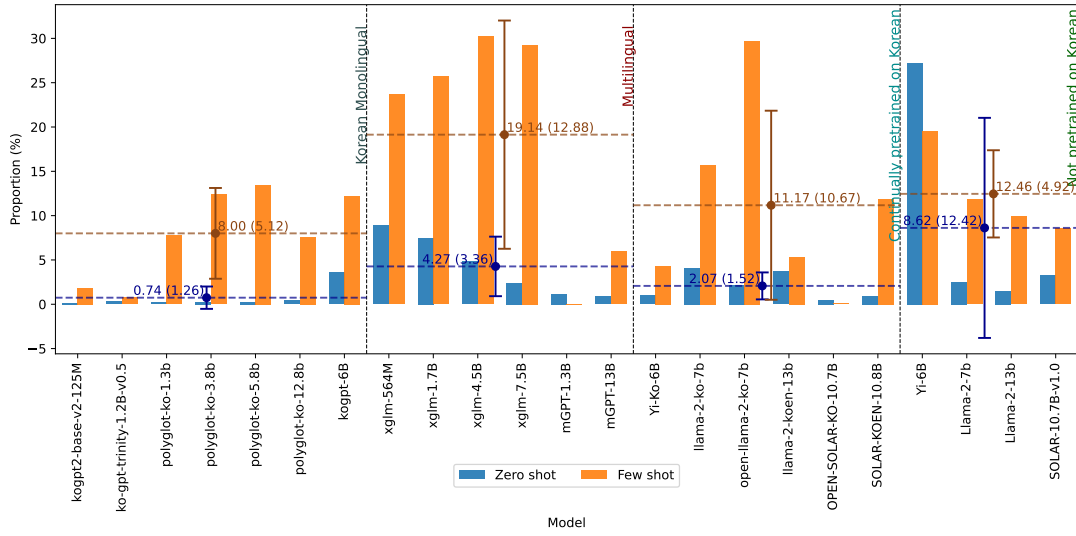


Figure 4.13: Proportion of sentences containing the headline pattern “<” or “>” provided in examples to the total sentences generated by language models.

were provided, the proportion of sentences containing the pattern is clearly higher, which suggests that the models referred to the headlines in text generation. Figure 4.13 shows the proportion of sentences containing a part of the headline pattern. Compared to Figure 4.12, it is noticeable that the ratios of XGLM models have increased significantly, suggesting that these models might have referred to the headline pattern but learned or imitated it incompletely.

Korean sentence-ending formats

Figure 4.14 shows the ratio of the number of sentences that do not have typical Korean sentence-ending formats to the total number of generated sentences. For the table of sentences with typical Korean sentence-ending formats on which Figure 4.14 is based, refer to Table A.11 in the appendix.

Korean monolingual models generally exhibit similar performance, with slightly lower ratios for KoGPT2-base and Ko-GPT-Trinity-1.2B models. Ko-GPT-Trinity has a higher ratio of incomplete sentence endings in the few-shot task.

Multilingual models generally have slightly higher ratios of sentences with incomplete endings compared to other models trained on Korean. Within the same model family, the ratio tends to decrease as the model size increases.

Korean continually pre-trained models show some variation. The Yi-Ko-6B and OPEN-SOLAR-KO-10.7B models have lower ratios of incomplete endings, while the Llama-2-ko-7b and Open-llama-2-ko-7b models have higher ratios. Results are not consistent even within the same model family or among models trained on the same pre-training data (e.g., models trained on Korean open access data or Korean-English mixed data). Further investigation may be needed.

pattern of “<...>” that do not include Korean characters were used for detection.

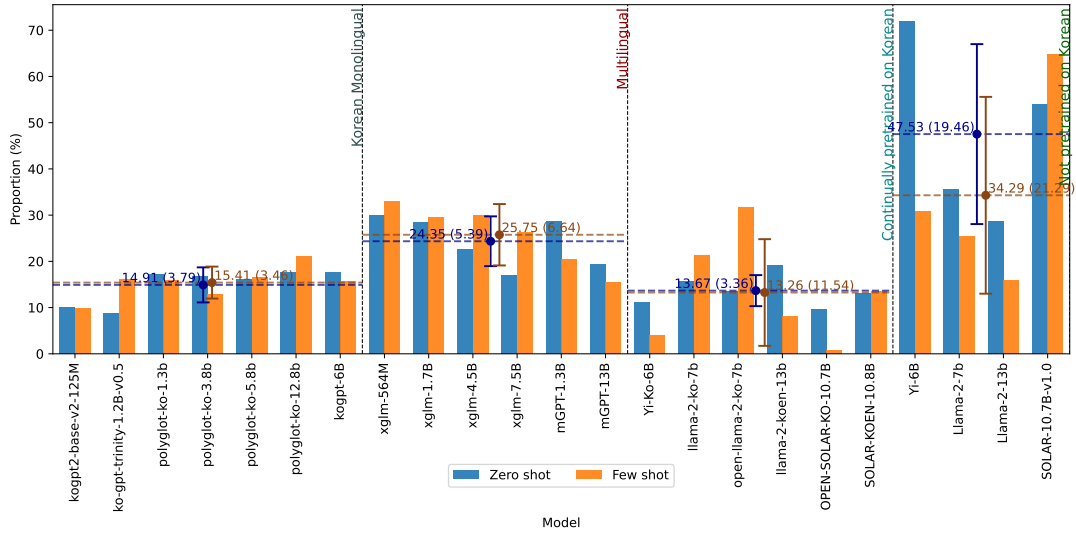


Figure 4.14: Proportion of the number of sentences that do not have typical Korean sentence-ending formats to the total number of generated sentences.

As might be expected, non-Korean-trained models have higher ratios of sentences with incomplete endings. However, similar to previous patterns, these ratios decrease in the few-shot task for all models except SOLAR-10.7B, which has a higher ratio of incomplete endings in the few-shot setting. A manual review of the generated data reveals that this is because in the zero-shot task, the model generated many sentences that are not in Korean but have complete sentence ending punctuation.

Filtering for formally complete Korean sentences

Table 4.2 shows the number of sentences filtered at each step in the process of applying the general Korean sentence formats above (i.e., Step 1: Korean character ratio of 0.4 or higher, Step 2: not containing unusual patterns, Step 3: not containing headline patterns, and Step 4: having a typical Korean sentence-ending format). Figure 4.15 presents the ratio of the number of formally complete Korean sentences to the total number of generated sentences for each model and task, obtained by the filtering process mentioned above.

Figure 4.15 can be seen as demonstrating the ability of each model in generating formally complete or typical Korean sentences for the given tasks, which were Korean generation tasks provided with Korean input prompts.

Korean monolingual models show generally high rates of formally complete Korean sentence generation, ranging from an average of 50% to 90%. KoGPT2-base and KoGPT-Trinity1.2B show higher rates than other models, with an average of over 80%. Polyglot models have an average rate of around 70%, while the KoGPT model has an average rate of around 50% to 60%. It is interesting to note that KoGPT2-base has a high rate despite its small size of 125M, while relatively larger models such as KoGPT (6B) and Polyglot-12.8b have lower rates. Further text review may be necessary. Korean monolingual models show slightly lower rates of formally complete Korean sentence

Model	Zero-shot						Few-shot					
	Total	Step1	Step2	Step3	Step 4	Pr.(%)	Total	Step1	Step2	Step3	Step 4	Pr.(%)
<i>Korean Monolingual Models</i>												
kogpt2-base-v2-125M	10,681	10,542	10,363	10,351	9,396	88.0	11,956	11,743	11,494	11,325	10,361	86.7
ko-gpt-trinity-1.2B-v0.5	7,320	7,128	6,952	6,937	6,481	88.5	7,830	7,388	6,837	6,832	6,315	80.7
polyglot-ko-1.3b	5,925	5,744	4,913	4,910	4,270	72.1	6,074	5,908	5,149	4,829	4,297	70.7
polyglot-ko-3.8b	5,909	5,756	4,921	4,911	4,334	73.3	5,721	5,629	5,114	4,627	4,200	73.4
polyglot-ko-5.8b	6,221	6,057	5,295	5,289	4,651	74.8	4,731	4,615	4,181	3,726	3,299	69.7
polyglot-ko-12.8b	6,242	6,075	5,101	5,085	4,396	70.4	5,911	5,669	4,256	4,017	3,487	59.0
kogpt-6B	7,623	7,299	5,536	5,540	4,627	60.7	4,599	4,303	2,868	2,539	2,238	48.7
<i>Multilingual Models</i>												
xglm-564M	7,516	6,598	3,241	3,108	2,477	33.0	7,381	6,867	3,939	2,887	2,181	29.5
xglm-1.7B	7,575	6,814	3,332	3,230	2,529	33.4	7,711	7,183	4,550	3,296	2,553	33.1
xglm-4.5B	5,236	4,861	3,277	3,203	2,812	53.7	6,961	6,709	4,851	3,327	2,650	38.1
xglm-7.5B	4,802	4,524	3,734	3,673	3,358	69.9	7,129	7,001	6,203	4,482	3,726	52.3
mGPT-1.3B	5,252	5,000	3,648	3,624	2,964	56.4	3,257	3,095	2,162	2,162	2,003	61.5
mGPT-13B	6,047	5,902	5,379	5,338	4,566	75.5	4,597	4,440	4,031	3,871	3,620	78.7
<i>Korean Continually Pre-trained Models</i>												
Yi-Ko-6B	8,208	8,084	7,550	7,502	6,910	84.2	3,407	3,356	3,261	3,173	3,142	92.2
llama-2-ko-7b	9,089	8,792	7,688	7,625	6,754	74.3	9,260	8,954	7,842	6,644	5,779	62.4
open-llama-2-ko-7b	3,707	3,476	3,041	2,988	2,765	74.6	2,975	2,837	2,241	1,678	1,414	47.5
llama-2-koen-13b	6,993	6,805	5,562	5,394	4,821	68.9	3,150	3,121	2,931	2,812	2,727	86.6
OPEN-SOLAR-KO-10.7B	5,174	4,919	4,602	4,582	4,279	82.7	2,495	2,492	2,458	2,456	2,437	97.7
SOLAR-KOEN-10.8B	7,609	7,492	6,628	6,578	5,911	77.7	4,302	4,181	4,024	3,645	3,487	81.1
<i>Non-Korean-trained Models</i>												
Yi-6B	7,846	2,481	1,852	1,816	1,249	15.9	4,423	3,717	3,162	2,818	2,497	56.5
Llama-2-7b	4,716	2,530	2,025	1,998	1,354	28.7	4,252	3,551	2,679	2,542	2,289	53.8
Llama-2-13b	5,595	3,013	2,717	2,675	2,134	38.1	4,220	3,942	3,666	3,401	3,213	76.1
SOLAR-10.7B-v1.0	10,265	1,187	990	971	412	4.0	1,963	1,319	909	907	72	3.7

Table 4.2: Number of sentences filtered at each step in general Korean sentence format filtering process. The proportion of the final filtered sentences relative to the initial total number of sentences is also shown.

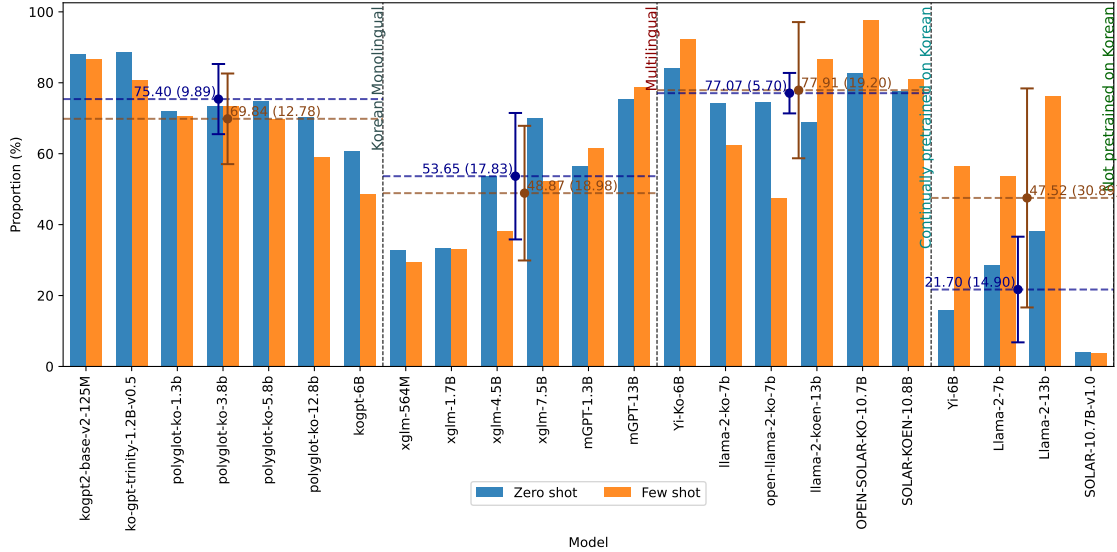


Figure 4.15: Proportion of the number of formally complete Korean sentences to the total number of sentences generated by language models.

generation in the few-shot task. Considering Figure 4.13, this may be partly due to the influence of the headline pattern.

Multilingual models show rates of formally complete Korean sentence generation ranging from an average of 30% to 70%. mGPT models have higher rates compared to XGLM models. Within the same model family, there is a tendency for the rate of formally

complete Korean sentence generation to increase as the model size grows. XGLM models have higher rates in the zero-shot task, while mGPT models have higher rates in the few-shot task. Considering Figure 4.13, the lower rates of XGLM models in the few-shot task also seems to be partly influenced by the headline pattern.

Korean continually pre-trained models show overall high rates of formally complete Korean sentence generation, ranging from an average of 50% to 90%. However, there is some variation among models, and the difference between zero-shot and few-shot tasks is larger compared to Korean monolingual models. It is interesting to note that they do not show consistent patterns according to the same model family or the same Korean training dataset for continual pre-training. The model architecture and total pre-training data may have interacted in complex ways. The relatively low rate of open-llama-2-ko-7b in the few-shot task seems to be partly influenced by the headline pattern, considering Figure 4.13.

Non-Korean-trained models show a rate of formally complete Korean sentence generation ranging from an average of 20% to 70%. The SOLAR-10.7B model shows a low rate of around 5% in both zero-shot and few-shot tasks. Based on a review of the actual generated text, it is difficult to say that this model has the ability to generate formally complete Korean sentences. Other models, such as Yi-6B, Llama-2-7b, and Llama-2-13b, show low rates in the zero-shot task but a substantial increase in the few-shot task, becoming similar to other models trained on Korean. In particular, the Llama-2-13b model shows a high rate of over 70% in the few-shot task, similar to Korean monolingual models or Korean continually pre-trained models. The Yi-6B model also shows a greatly increased ratio of around 60% in the few-shot task compared to its rate of less than 20% in the zero-shot task.

One thing to keep in mind is that the complete sentences filtered here indicate that they have a formally complete or typical form of Korean sentences, but it does not necessarily mean that these sentences are semantically meaningful. For example, manually reviewing the actual sentences, complete sentences generated by non-Korean-trained models, such as Llama-2-7b, Yi-6B, and Llama-2-13b, are formally complete Korean sentences but often do not make sense. This leads to the speculation that these models might have learned the ability to imitate formally complete Korean sentences through the examples provided in the few-shot task, regardless of their understanding of the meaning. It is also noteworthy that even among these models, the complete sentences from the Llama-2-13b model are slightly more often meaningful, suggesting the influence of model size.

4.1.3 Semantic Evaluation

Semantic embedding similarity

Table 4.3 shows the semantic embedding similarity between the original text in the dataset and the text generated by language models, sorted in descending order. Figure 4.16 illustrates the cosine similarity scores for each task and model based on Table 4.3.

In the zero-shot task, the models were given only the first three words of the body text as

4.1. Analysis for generated texts

Model	Zero-shot Type	Similarity	Model	Few-shot Type	Similarity
Yi-Ko-6B	Ko. Continually Pre-trained	0.521	Yi-Ko-6B	Ko. Continually Pre-trained	0.663
Polyglot-Ko-5.8b	Ko. Monolingual	0.515	OPEN-SOLAR-KO-10.7B	Ko. Continually Pre-trained	0.655
Llama-2-koen-13b	Ko. Continually Pre-trained	0.513	Llama-2-koen-13b	Ko. Continually Pre-trained	0.646
Polyglot-Ko-12.8b	Ko. Monolingual	0.511	Polyglot-Ko-12.8b	Ko. Monolingual	0.599
OPEN-SOLAR-KO-10.7B	Ko. Continually Pre-trained	0.506	Polyglot-Ko-1.3b	Ko. Monolingual	0.598
KoGPT-6B	Ko. Monolingual	0.505	Polyglot-Ko-5.8b	Ko. Monolingual	0.595
Polyglot-Ko-3.8b	Ko. Monolingual	0.501	Polyglot-Ko-3.8b	Ko. Monolingual	0.592
Polyglot-Ko-1.3b	Ko. Monolingual	0.491	Llama-2-13b	Non-Korean-Trained	0.581
SOLAR-KOEN-10.8B	Ko. Continually Pre-trained	0.488	Ko-GPT-Trinity-1.2B-v0.5	Ko. Monolingual	0.572
Open-llama-2-ko-7b	Ko. Continually Pre-trained	0.488	mGPT-13B	Multilingual	0.569
Ko-GPT-Trinity-1.2B-v0.5	Ko. Monolingual	0.486	Open-llama-2-ko-7b	Ko. Continually Pre-trained	0.563
XGLM-7.5B	Multilingual	0.478	KoGPT-6B	Ko. Monolingual	0.559
Llama-2-ko-7b	Ko. Continually Pre-trained	0.469	SOLAR-KOEN-10.8B	Ko. Continually Pre-trained	0.521
mGPT-13B	Multilingual	0.463	mGPT-1.3B	Multilingual	0.501
XGLM-4.5B	Multilingual	0.459	Llama-2-7b	Non-Korean-Trained	0.497
XGLM-1.7B	Multilingual	0.452	XGLM-7.5B	Multilingual	0.484
mGPT-1.3B	Multilingual	0.450	Llama-2-ko-7b	Ko. Continually Pre-trained	0.481
XGLM-564M	Multilingual	0.439	Yi-6B	Non-Korean-Trained	0.481
KoGPT2-base-v2-125M	Ko. Monolingual	0.427	XGLM-4.5B	Multilingual	0.459
Llama-2-13b	Non-Korean-Trained	0.381	KoGPT2-base-v2-125M	Ko. Monolingual	0.456
Llama-2-7b	Non-Korean-Trained	0.378	XGLM-1.7B	Multilingual	0.426
Yi-6B	Non-Korean-Trained	0.361	XGLM-564M	Multilingual	0.420
SOLAR-10.7B-v1.0	Non-Korean-Trained	0.300	SOLAR-10.7B-v1.0	Non-Korean-Trained	0.347

Table 4.3: Cosine similarity semantic embeddings between the original texts and the generated texts by language models, presented in descending order of similarity scores.

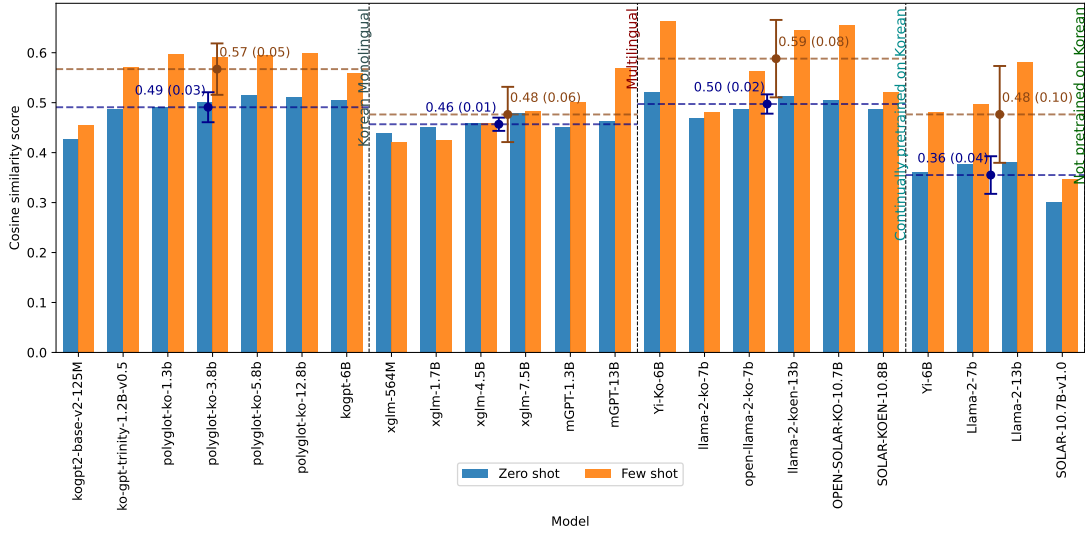


Figure 4.16: Cosine similarity scores of semantic embeddings between the original texts and the generated texts by language models.

the input prompt, which may lead to generated content that diverges from the original. Thus, the semantic similarity scores between the generated text and the original text may not be highly meaningful. However, these scores can still serve as a reference for comparison with the few-shot task results. Additionally, the scores of the SOLAR-10.7B model, which barely generated meaningful Korean text, can be considered as a lower bound for semantic similarity scores.

For Korean monolingual models in the few-shot task, scores are generally consistent between 0.5 and 0.6. Interestingly, the KoGPT2-base model, which demonstrated high performance in the surface-level evaluation (i.e., the ability to generate formally complete

Korean sentences in Figure 4.15), has a lower score in semantic embedding similarity compared to other Korean monolingual models. This may suggest that it generated sentences that are not directly related to the content of the dataset or the input prompts. Another possible explanation is, as seen in Table 4.1, the KoGPT2-base model generates longer texts and more sentences compared to other models. As the text gets longer, it may have deviated into content that is not directly related to the given input prompts. It is also noteworthy that the Polyglot-Ko-12.8B model, which showed relatively lower performance in the surface-level evaluation, achieves the highest score among Korean monolingual models.

For multilingual models in the few-shot task, scores generally range from 0.4 to 0.5, lower than those for Korean monolingual models and Korean continually pre-trained models. The scores of mGPT models are a bit higher than those of XGLM models. Within the same model family, the scores tend to increase with model size. For example, XGLM-7.5B model and mGPT-13B model show the highest scores among their model families. In contrast, it is noteworthy that the XGLM-564M and XGLM-1.7B models show lower scores in the few-shot task. As observed in the surface-level evaluation, these models may have been influenced by the input prompts and attempted to utilize them in text generation, but their limited ability to generate Korean sentences might have hindered their performance.

For Korean continually pre-trained models in the few-shot task, scores range between 0.5 and 0.7, with some variations among the models. The Yi-Ko-6B, Llama-2-koen-13b, and OPEN-SOLAR-KO-10.7B models all show high scores of 0.65 or above. The score of the Llama-2-ko-7b model is relatively low among Korean continually pre-trained models. As seen in Table 4.1, similar to the KoGPT2-base model, the Llama-2-ko-7b model has longer text lengths and larger number of sentences. This may have caused the text to deviate into content that is not directly related to the given input prompts.

Among the non-Korean-trained models in the few-shot task, the Yi-6B and Llama-2-7b models show scores of 0.48 and 0.50, and the Llama-2-13b model demonstrates a relatively high score of 0.58. It is noteworthy that the semantic similarity scores of these models are higher than some of the Korean pre-trained models, including Korean monolingual models, multilingual models, and Korean continually pre-trained models. It is especially remarkable that these models' semantic similarity scores are higher than the KoGPT2-base model, which excelled at generating formally complete Korean sentences. Although the texts generated by non-Korean-trained models contained many formally incomplete sentences, their attempts to capture the context given in the input prompts might be reflected in their scores. It can be speculated that these models have the ability to capture the relationships provided in the input prompts through the few-shot examples and incorporate them into text generation.

Semantic embedding visualization

Figure 4.17 and Figure 4.18 visualize the semantic embeddings created from the original texts and the texts generated by the language models in a 2-dimensional embedding space using t-SNE. For comparison, the embedding space was generated using all the semantic embeddings. In other words, the embeddings of each model are displayed in separate subplots for distinction, but they share the same embedding space. We experimented

with visualizations using PCA, t-SNE, and UMAP, but here we only show the t-SNE visualization, as it provides the best distinction between embeddings. For visualizations using PCA and UMAP, refer to Figure A.2 and Figure A.3 in the appendix. Additionally, since the visualization produces non-deterministic results that vary slightly depending on the random seed, we tested with multiple random seeds. However, despite slight variations in shape, the overall patterns generally remain consistent.

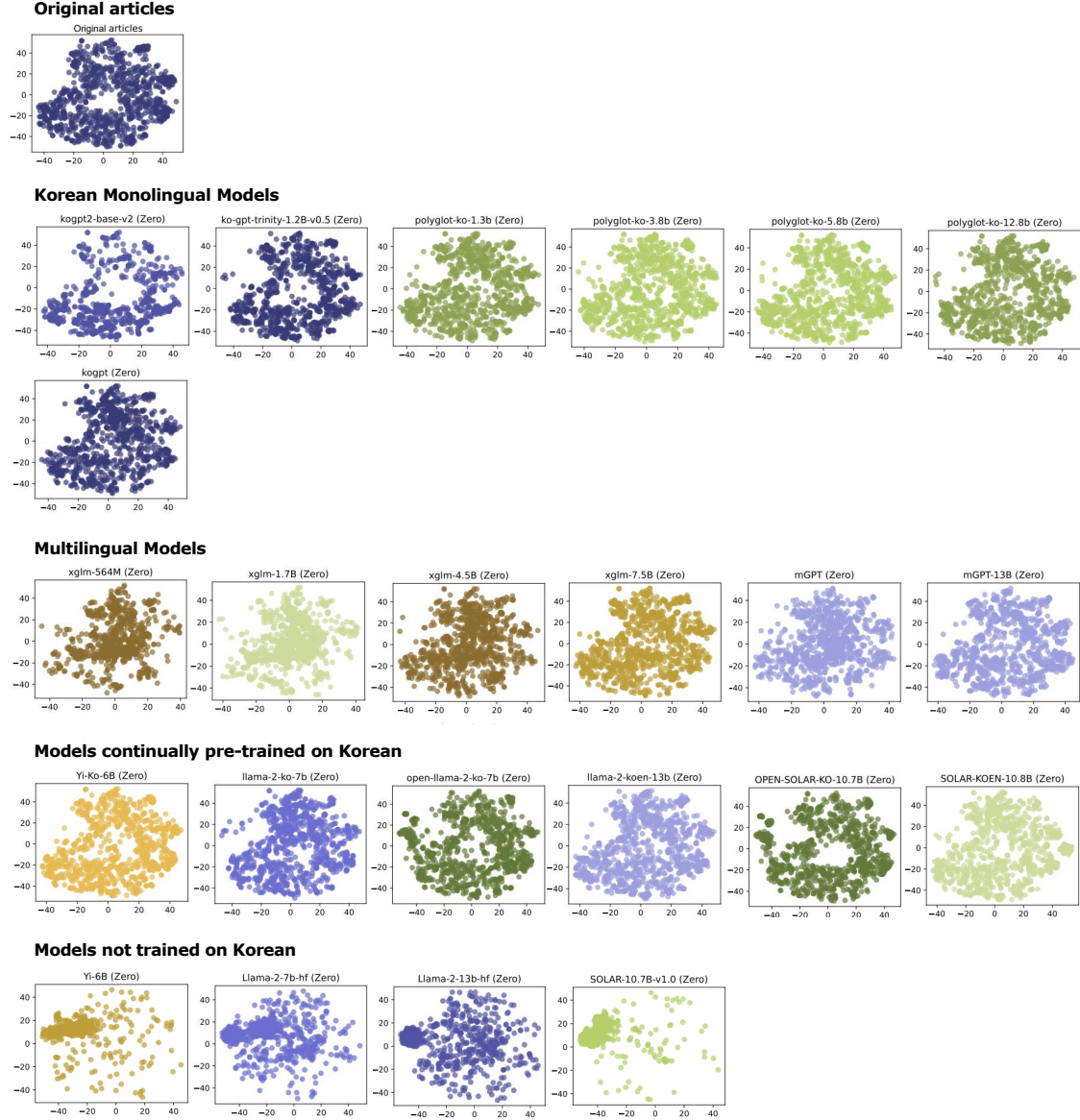


Figure 4.17: Visualization using t-SNE of semantic embedding distribution for the original texts and the texts generated by language models in the zero-shot task.

Figure 4.17 shows a clear distinction between the distribution of embeddings from models trained on Korean and models not trained on Korean. The embeddings of models not trained on Korean are concentrated in the upper left region of the plot, which is presumed to be associated with non-Korean textual representations, considering the actual generated texts. Additionally, multilingual models generally have the space around the origin filled, while Korean monolingual models and Korean continually pre-

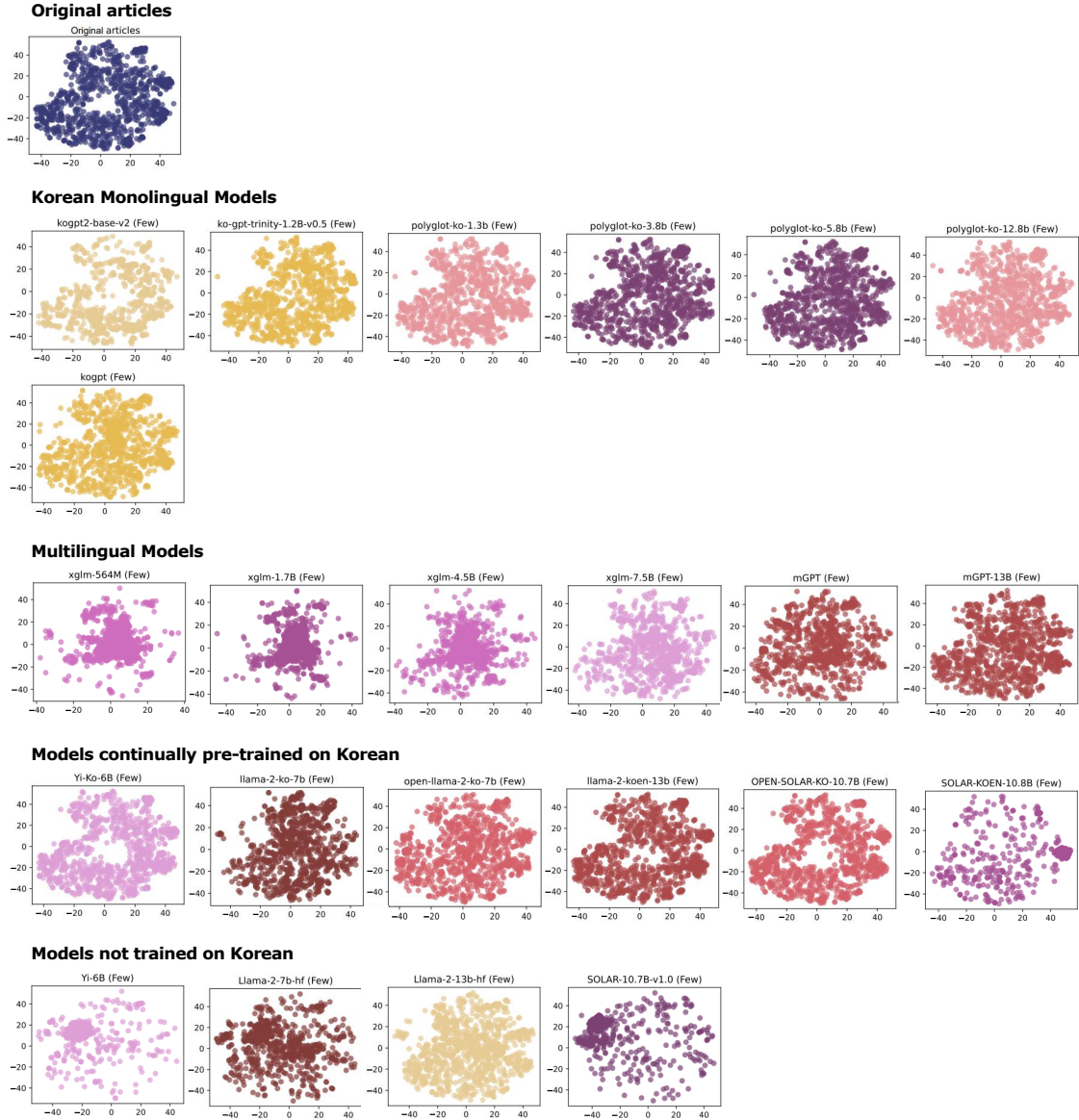


Figure 4.18: Visualization using t -SNE of semantic embedding distribution for the original texts and the texts generated by language models in the few-shot task.

trained models have the space around the origin mostly empty. Examining the actual texts corresponding to the embeddings distributed around the origin reveals that they are in Korean but often contain disjointed or incoherent content. It can be speculated that the embeddings of the original text have unique and diverse topics or meanings, thus positioned widely distributed away from the origin, while the embeddings with mixed or ambiguous meanings are located near the origin.

Figure 4.18 shows that the embedding distributions of models not trained on Korean have changed significantly compared to Figure 4.17. For the non-Korean-trained models, Except for the SOLAR-10.7B model, the tendency to skew towards the upper left has highly reduced, and the Llama-2-13b model’s embedding distribution has become quite similar to that of the original text, except for the space around the origin. Korean monolingual models show a slight increase in density around the origin, compared to

Figure 4.17. Among the Korean continually pre-trained models, the Yi-Ko-6B, llama-2-koen-13b, and OPEN-SOLAR-10.7B models have become more similar to the original text distribution compared to Figure 4.17, while the llama-2-ko-7b, open-llam-2-ko-7b, and SOLAR-KOEN-10.8B models have become less similar. Examining the generated texts from the llama-2-ko-7b and open-llama-2-ko-7b models in the few-shot setting reveals a tendency to generate multiple headlines and body texts for each case. This might be interpreted as incoherent semantic content overall. For the multilingual models, XGLM-564M, XGLM-1.7B, and XGLM-4.5B models show a tendency to converge towards the origin in the few-shot task compared to Figure 4.17. Examining the generated texts from these models also reveals a tendency to include multiple headlines and symbols or phrases unrelated to the body text, which might explain their placement near the origin.

The similarity in the distribution patterns does not seem to fully correspond to the cosine similarity rankings provided in Table 4.3. This may be because the embedding visualization shows the overall distribution and relative positions of embeddings in the embedding space, which may differ from the perspective of pairwise comparisons with the original text. For example, if the text generated for original text A is closer to original text B , and the text generated for original text B is closer to original text A , the similarity would be low from the perspective of pairwise comparisons for each case, but the distributions might still look similar in the embedding space. In other words, even if the generated texts have semantic differences when compared with the original text for each case, if they are generally similar to the overall semantic distribution of the original texts, they will appear similar in the distributional visualization. Additionally, limitations due to 2D visualization and the variability caused by non-deterministic visualization may also influence the interpretation to some extent.

4.2 Analysis for generated Korean sentences

4.2.1 Basic Statistics

Sentence count

The number of Korean sentences in the Korean corpora and filtered from the texts generated by language models is shown in Figure 4.19. For the underlying table, refer to Table A.12 in the appendix.

The KoGPT2-base model has a high number of sentences, around 10,000 on average, while other models show sentence counts ranging from 2,000 to 6,000.

Sentence length

Regarding the length of Korean sentences, Figure 4.20 and Figure 4.21 show the average text length and the average number of words per sentence in the Korean corpora and filtered from the texts generated by language models. For the underlying table, refer to

4.2. Analysis for generated Korean sentences

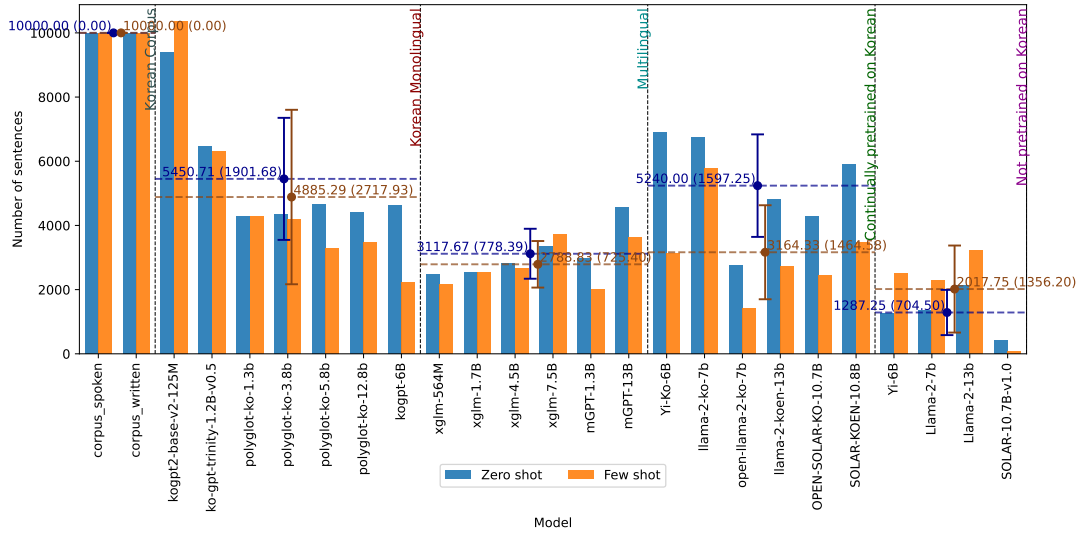


Figure 4.19: Number of Korean sentences generated by language models.

Table A.13 in the appendix.

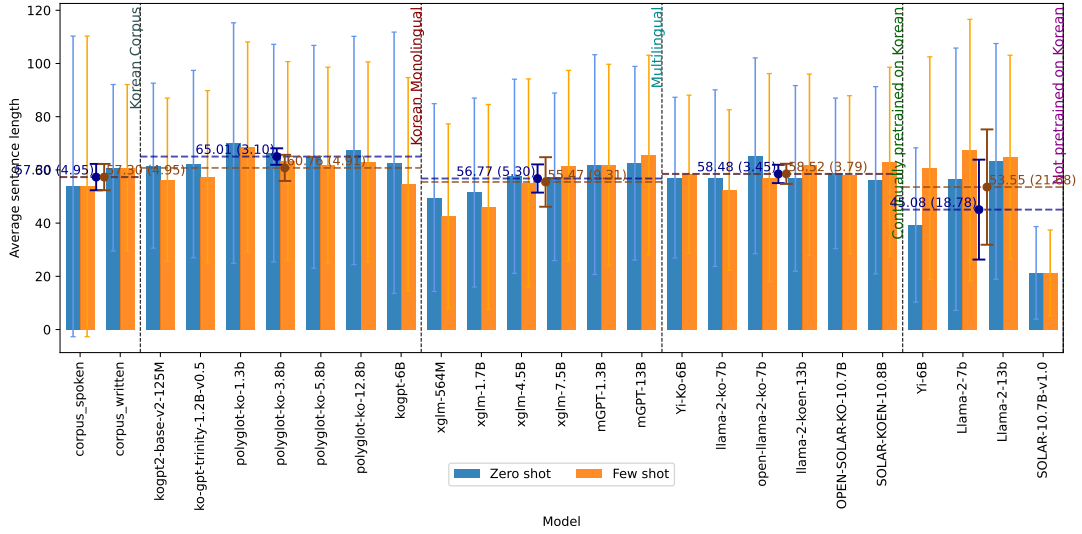


Figure 4.20: Average text length of Korean sentences generated by language models.

In terms of sentence length, there are no significant differences across the models overall. For non-Korean-trained models, the abnormally short length of the SOLAR-10.7B model suggests that its sentences are not typical, and the short length of the Yi-6B model in the zero-shot setting also indicate some differences from the sentences of other models. Additionally, the sentences of the XGLM-564M and XGLM-1.7B models are relatively shorter, while those of the Polyglot-Ko-1.3B model are relatively longer.

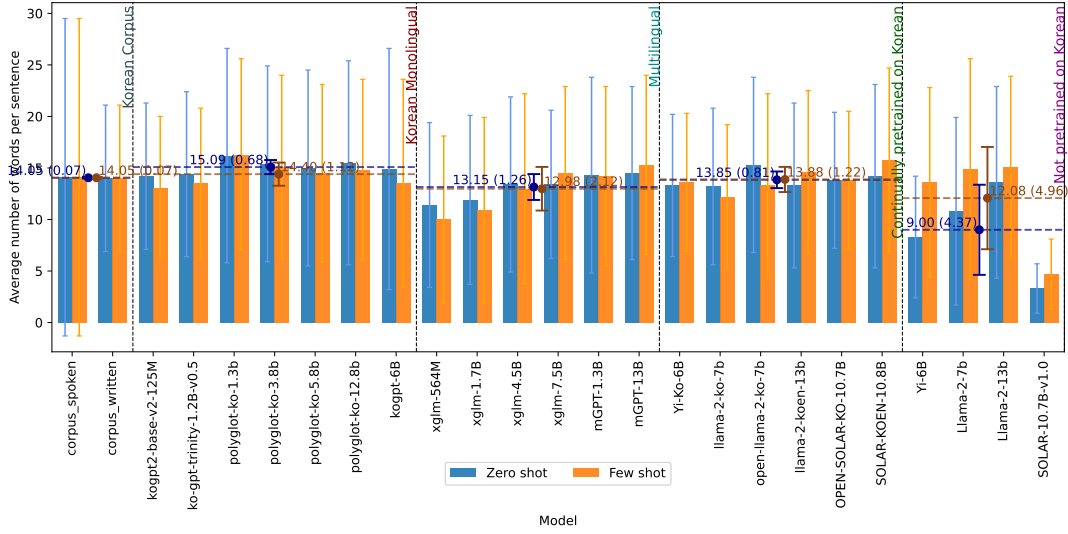


Figure 4.21: Average number of words per sentence generated by language models.

4.2.2 Lexical Evaluation

Lexical Diversity

From a lexical perspective, to examine the lexical diversity of sentences in Korean corpora and sentences generated by language models, the distribution of the root type-token ratios (RTTR) ³ is shown in Figure 4.22. For the underlying table, refer to Table A.14 in the appendix.

According to Figure 4.22, the lexical diversity of Korean monolingual models, Korean continually pre-trained models, and non-Korean-trained models is slightly higher than that of multilingual models, and there is some variation within the Korean monolingual model and non-Korean-trained model groups. Among individual models, KoGPT, Yi-6B, and Llama-2-7b models show high lexical diversity in the zero-shot setting. However, the high lexical diversity in non-Korean-trained models may be due to the presence of non-existent Korean vocabularies resulting from attempts of these models to generate Korean at the surface-level.

To address this, we downloaded a Korean Dictionary⁴ from the National Institute of the

³As described in Section 3.7.2, we also calculated several other indicators such as MSTTR, MATTR, and MTLD as measures of lexical diversity. However, possibly due to the large variation in text length or the distribution of tokens within the texts, we thought that the results were less appropriate than RTTR and CTTR and did not include them in the main text. For example, as seen in Table A.15, the spoken corpus, written corpus, and KoGPT2-base model, which have similar total token counts, should show much higher lexical diversity for the written corpus based on the size of the unique token count. However, these indicators did not reflect this. Similarly, for XGLM-7.5B and SOLAR-KOEN-10.8B, which have similar total token counts, the lexical diversity of SOLAR-KOEN-10.8B should be higher based on the size of the unique token count, but these indicators did not show this. We believe that further analysis is needed to identify the reasons for this. RTTR and CTTR yield very similar results, so we include only RTTR in the main text.

⁴“Urimalsaem (우리말샘)”, an open Korean dictionary with more than 1.1 million entries including

4.2. Analysis for generated Korean sentences

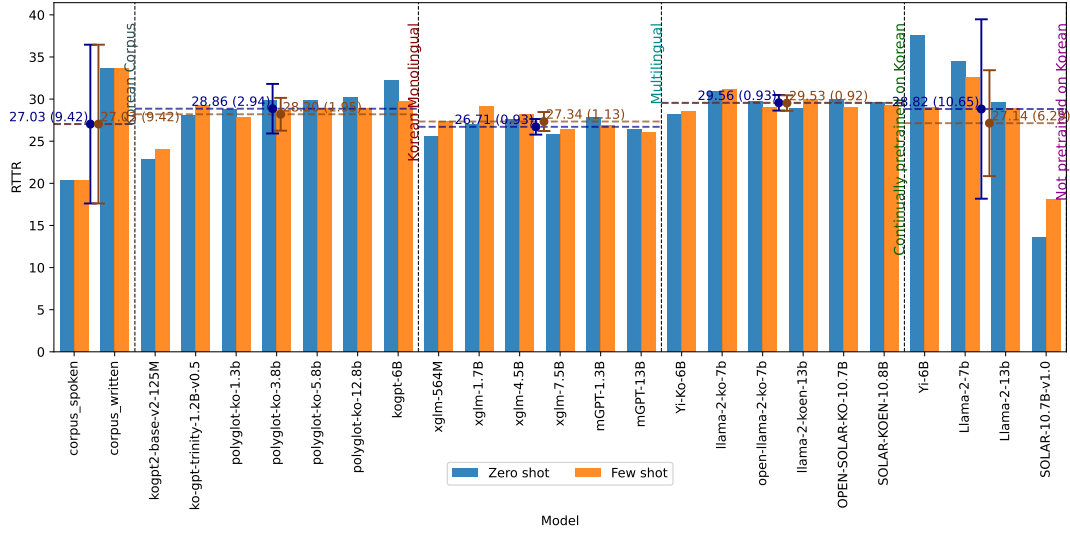


Figure 4.22: Root Type-Token Ratio (RTTR) of Korean corpora and Korean sentences generated by language models.

Korean Language, extracted the headwords from the dictionary, and created a set of Korean lexical morphemes by morphologically segmenting these headwords.⁵ We then recalculated the lexical diversity measures using only the Korean tokens included in this Korean lexical morpheme set. The results are shown in Figure 4.23. For the underlying table, refer to Table A.15 in the appendix.

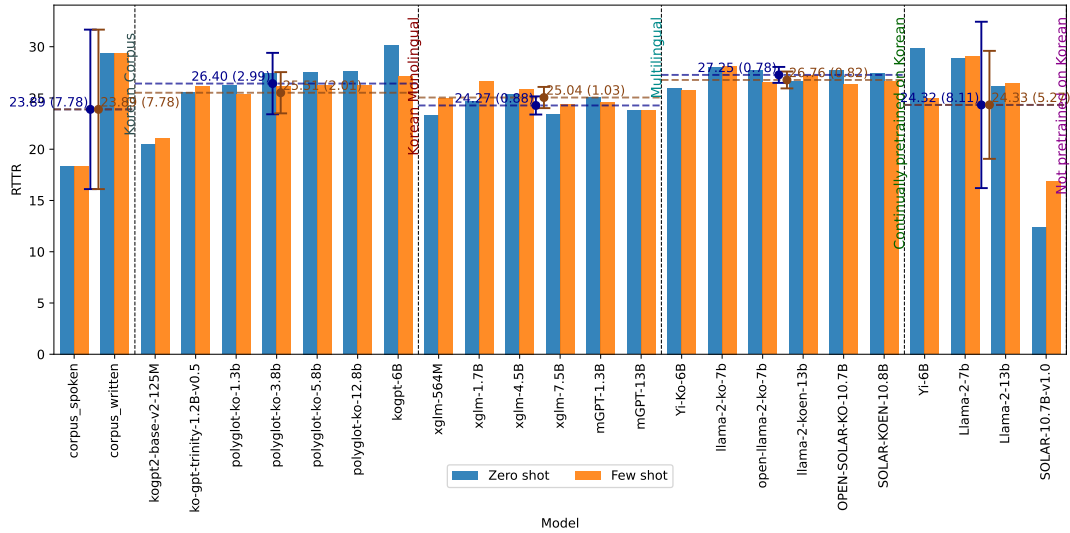


Figure 4.23: Root Type-Token Ratio (RTTR) of Korean corpora and Korean sentences generated by language models, considering only tokens included in the Korean dictionary entry morpheme set.

recent vocabularies. (<https://opendict.korean.go.kr/>)

⁵The number of dictionary headword forms, excluding duplicates, is 957,123. The number of morpheme forms extracted from these headwords, excluding duplicates, is 200,977. For morphological segmentation, the Kiwi library (M. Lee, 2022) was used.

According to Figure 4.23, after removing morphemes that do not exist in the Korean dictionary, the RTTR values for non-Korean-trained models have somewhat decreased. This suggests that the texts generated by these models may have included some vocabulary that does not actually exist in Korean. However, the RTTR values of the Yi-6B and Llama-2-7b models in the non-Korean-trained models remain relatively high, indicating that additional filtering for incorrect vocabulary at the word level may be necessary. A manual review of the actual texts reveals that the sentences generated by these models often contain Korean vocabulary that either does not exist or does not make sense, resulting from attempts to generate Korean at the surface level. For a more accurate analysis of lexical diversity, it seems that an examination of lexical errors should also be conducted. Among the Korean monolingual models, the RTTR score of KoGPT is higher than average, while the RTTR score of KoGPT2-base-v2 is quite lower than average. Considering the high performance of KoGPT2-base-v2 in generating Korean sentences at a surface level (Figure 4.15), it is interesting, and it is speculated that the small size of this model may have led to a relatively limited vocabulary selection in sentence generation.

4.2.3 Syntactic Evaluation

Universal part-of-speech tags (UPOS)

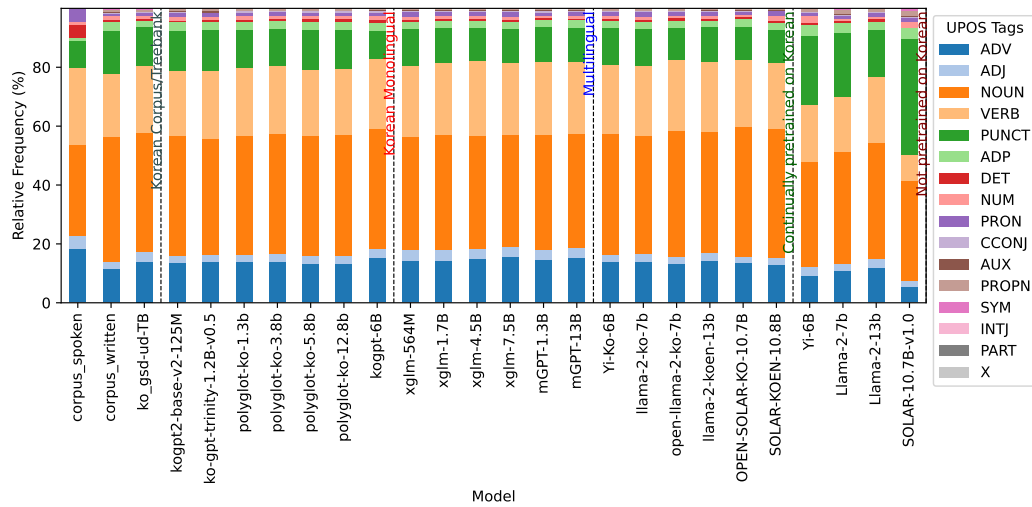


Figure 4.24: UPOS distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.

Figure 4.24 and Figure 4.25 show the distribution of UPOS in the Korean corpora and generated texts. For the underlying tables and the description of UPOS tags, refer to Table A.16, Table A.17, and Table A.18 in the appendix, respectively.

Except for the non-Korean-trained models, there are no substantial differences among the generated texts. Rather, the spoken language corpus shows a relatively higher proportion of adverbs, verbs, determiners, and pronouns, and a relatively lower proportion of nouns, punctuation, and adpositions, indicating a difference in part-of-speech tag distribution compared to the written languages. In the non-Korean-trained models,

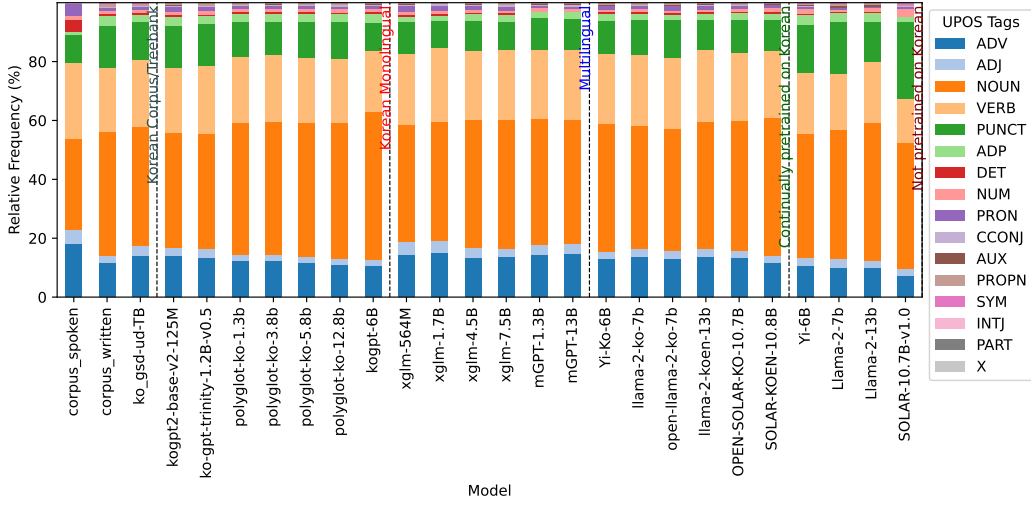


Figure 4.25: UPOS distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.

the high proportion of parts of speech classified as PUNCT suggests the presence of unusual symbols in the sentences. In the few-shot task, the difference in part-of-speech distribution compared to the models trained on Korean is much reduced. In the few-shot task, the Polyglot-ko and KoGPT models show a decrease in the proportion of adverbs and an increase in the proportion of nouns compared to the zero-shot task.

Language-specific part-of-speech tags (XPOS)

Figure 4.26 and Figure 4.27 show the distribution of language-specific tags in Korean corpora and generated texts. For the underlying tables and the description of XPOS tags,⁶ refer to Table A.19, Table A.20, and Table A.21 in the appendix, respectively.

The Korean language-specific tags have a more fine-grained classification than the universal tag classification, so the Figures show the distribution of up to the 20 most frequently occurring tags. The average ratios and their rank are displayed in the Figures. The presence of “+” in the tags indicates that in Korean, an agglutinative language, words often combine with particles or endings to form an “eojeol” (a unit split by spaces), representing the combination of these components.

As with the universal tag distribution, in the zero-shot task, non-Korean-trained models

⁶It is stated that the XPOS in the GSD Treebank was assigned by the KOMA morphological analyzer (D.-G. Lee and Rim, 2005) program, but no reference could be found for the tagset. However, the form of the tags suggests that they are based on the commonly used 21C Sejong Project-based tag set (https://www.korean.go.kr/front/reportData/reportDataView.do?mn_id=45&report_seq=197&pageIndex=1). Therefore, we have included the tag set table from the “Part-of-Speech Tag Set for Morphological Annotation of Written Texts”(https://committee.tta.or.kr/summary/standard_view.jsp?section=1&pk_num=TTAK.KO-11.0010/R1&nowSu=4), which is based on the Sejong tag set, in the appendix. Note that the original text is in Korean and the English translation was provided by the author. As such, the terms used in the table may not be the precise terminology commonly used in the field.

4.2. Analysis for generated Korean sentences

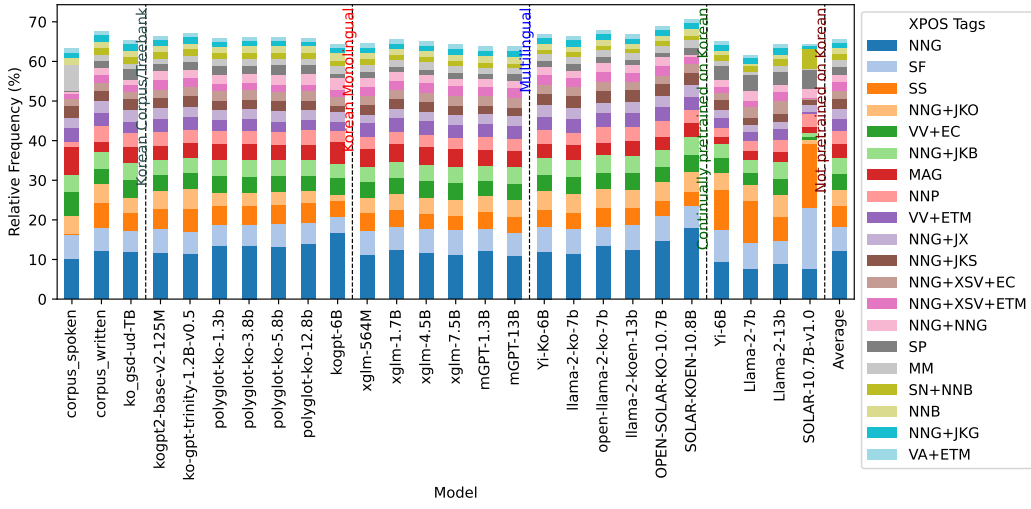


Figure 4.26: XPOS distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.

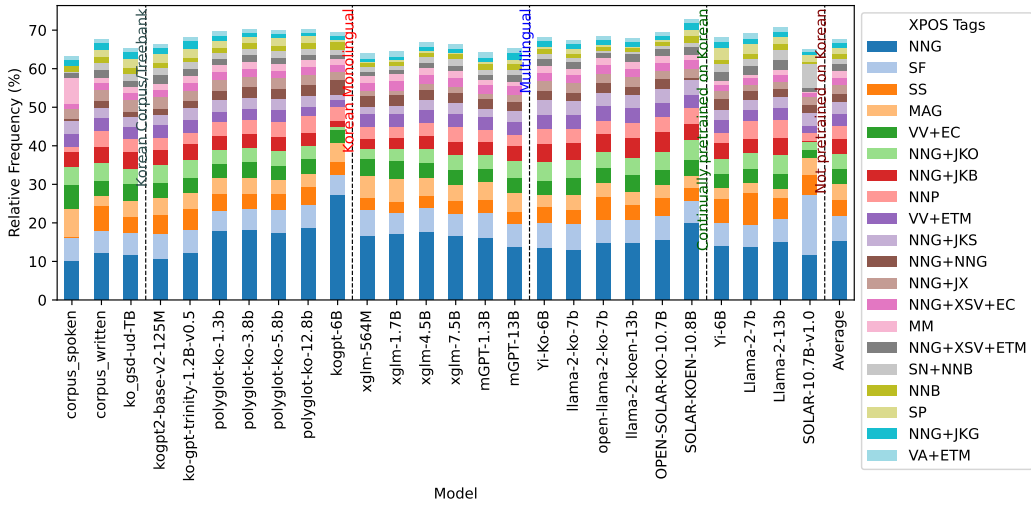


Figure 4.27: XPOS distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.

and the spoken corpus are differentiated from other groups. Both have lower ratios of NNG (general nouns), and non-Korean-trained models have higher ratios of symbols such as SS (quotation marks, parentheses, dashes), while the spoken corpus has a higher ratio of MAG (general adverbs).

Apart from that, there are some differences as individual models rather than differences between model groups (Korean monolingual models, multilingual models, Korean continually pre-trained models). For example, the KoGPT and SOLAR-KOEN-10.8B models have higher ‘general noun’ (NNG) ratios compared to other models.

In the few-shot task, the differences with models not trained on Korean are reduced. Overall, the ratio of general nouns increases in all models, with a particularly noticeable increase in KoGPT.

4.2. Analysis for generated Korean sentences

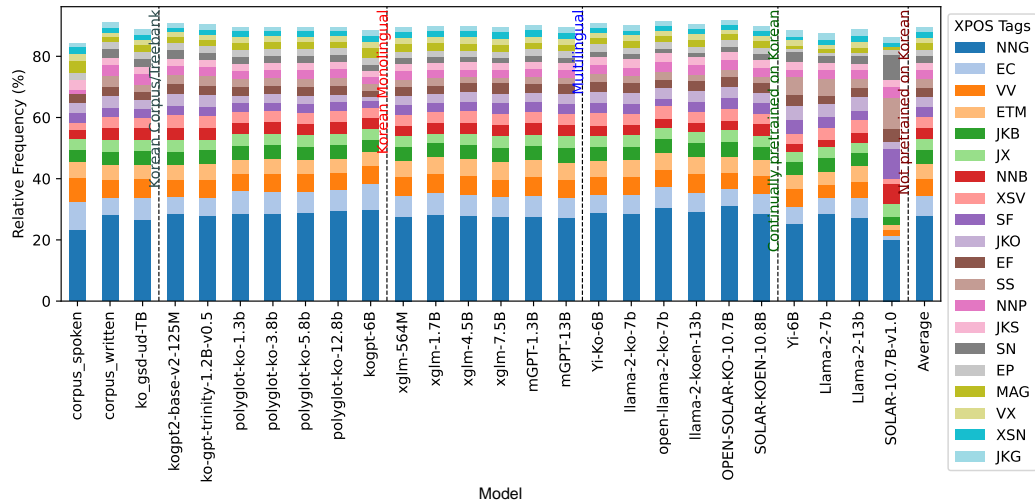


Figure 4.28: XPOS distribution (separated tag) in Korean corpora and Korean sentences generated by language models in the zero-shot task.

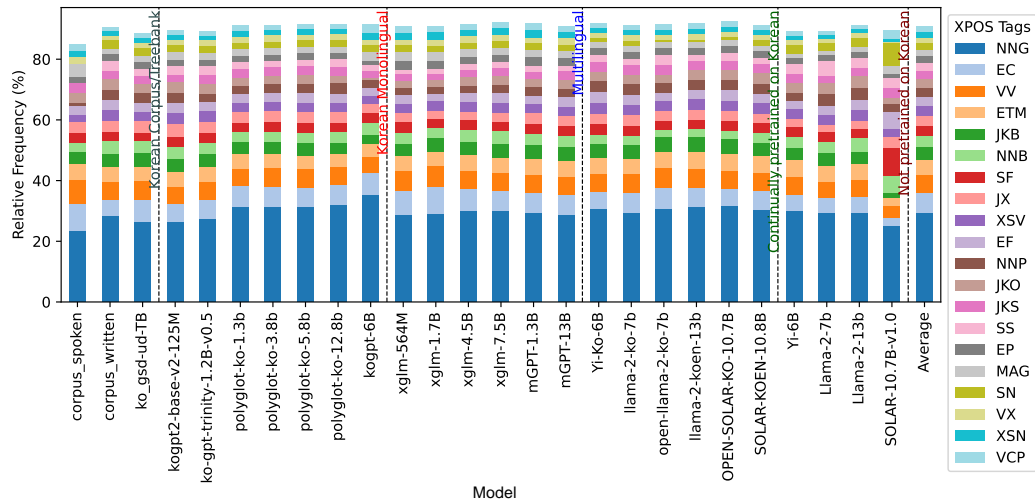


Figure 4.29: XPOS distribution (separated tag) in Korean corpora and Korean sentences generated by language models in the few-shot task.

Figure 4.28 and Figure 4.29 show the results of the tag distributions calculated by separating the tags connected by “+” in the original tag forms. For the underlying tables, refer to Table A.22 and Table A.23 in the appendix. Similar to what was observed in other POS distributions, there are no substantial differences between model groups in the zero-shot task, except for non-Korean-trained models. In the few-shot task, the difference is reduced, and the proportion of NNG (general nouns) slightly increases in the Polyglot-Ko and KoGPT models of the Korean monolingual models.

Universal dependency relations (deprel)

Figure 4.30 and Figure 4.31 show the distribution of the universal dependency relations in Korean sentences from Korean corpora and those generated by language models. The universal dependency relations also have a fine-grained list,⁷ so the Figures show the distribution of up to the 20 most frequently occurring relations. For the underlying tables and the description of the tags, refer to Table A.24, Table A.25, and Table A.26 in the appendix.

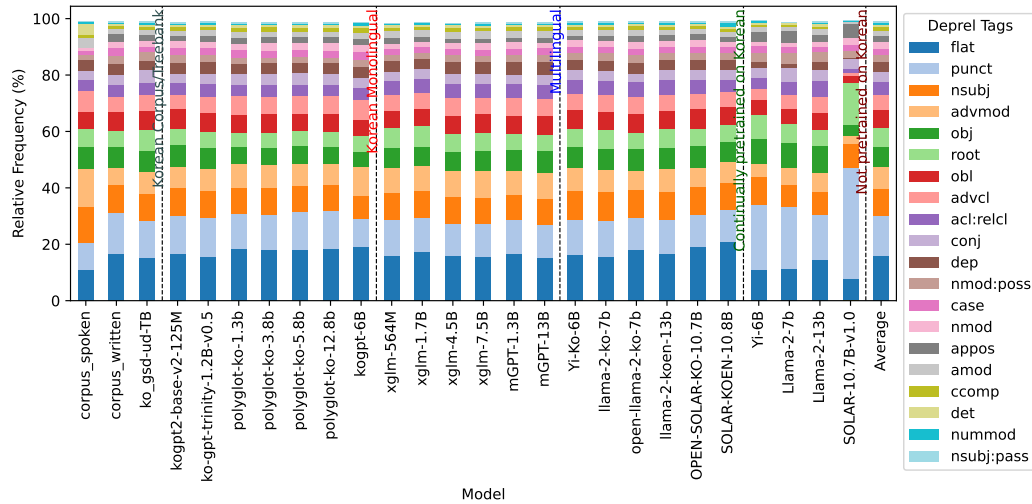


Figure 4.30: Dependency relation distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.

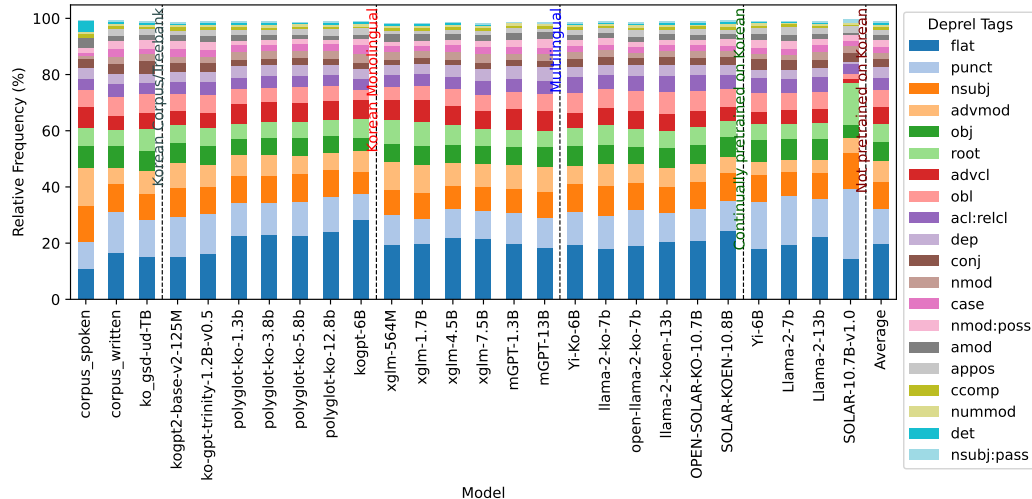


Figure 4.31: Dependency relation distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.

In the zero-shot task, similar to previous observations, there is a noticeable difference in the non-Korean-trained models and the spoken corpus compared to other model groups.

⁷<https://universaldependencies.org/u/dep/index.html>

It can be seen that the dependency relations in the Korean sentences generated by non-Korean-trained models are not typical.

While there are no substantial differences among the three model groups, multilingual models tend to have a slightly higher proportion of adverbial modifiers (advmod) and a slightly lower proportion of flat expressions (flat) compared to Korean monolingual models or Korean continually pre-trained models.

In the few-shot task, again similar to previous observations, it can be seen that the dependency relation distribution of the non-Korean-trained models becomes more similar to other models. It is noticeable that the differences between models within other model groups have increased. Compared to the zero-shot setting, there is an overall increase in flat expression (Flat) in the few-shot setting. This might be related to the increased proportion of nouns observed in the UPOS and XPOS distributions. The examples provided in the few-shot setting might have led to a higher proportion of nouns, such as proper nouns, which are difficult to assign specific dependency relations. This could have resulted in a relative increase in the flat expression dependency relation.

Overall, in terms of dependency relations in Korean sentences, the differences between language model groups are not substantial, while there are some variations among individual models within each group.

Dependency arc

Figure 4.32 and Figure 4.33 show the direction and lengths of dependency arcs in Korean sentences from the Korean corpora and those generated by language models. Refer to Table A.27 and Table A.28 in the appendix for the underlying tables.

In the zero-shot task, there are no substantial differences in the distribution of dependency arc directions among the models, except for the spoken corpus and the SOLAR-10.7B model. The proportion of left-pointing arcs is higher, around 60%. In the few-shot task, the proportion of left-pointing arcs slightly decreases, except for the Ko-GPT2-base, XGLM-1.7B, and SOLAR-10.7B models. There is some variation in the changes among models within the Korean monolingual models and multilingual models, with the changes in the Polyglot-Ko and Ko-GPT models being particularly noticeable.

In terms of arc length, multilingual models tend to have slightly shorter arcs compared to Korean monolingual models, but overall, there is almost no difference.

Considering that the structure and complexity of sentences can be related from the direction and length of dependency arcs, there seem to be no substantial differences among the model groups trained on Korean in terms of the structure or complexity of the generated Korean sentences. Instead, there may be differences between individual models depending on the specific task or input prompt.

4.2. Analysis for generated Korean sentences

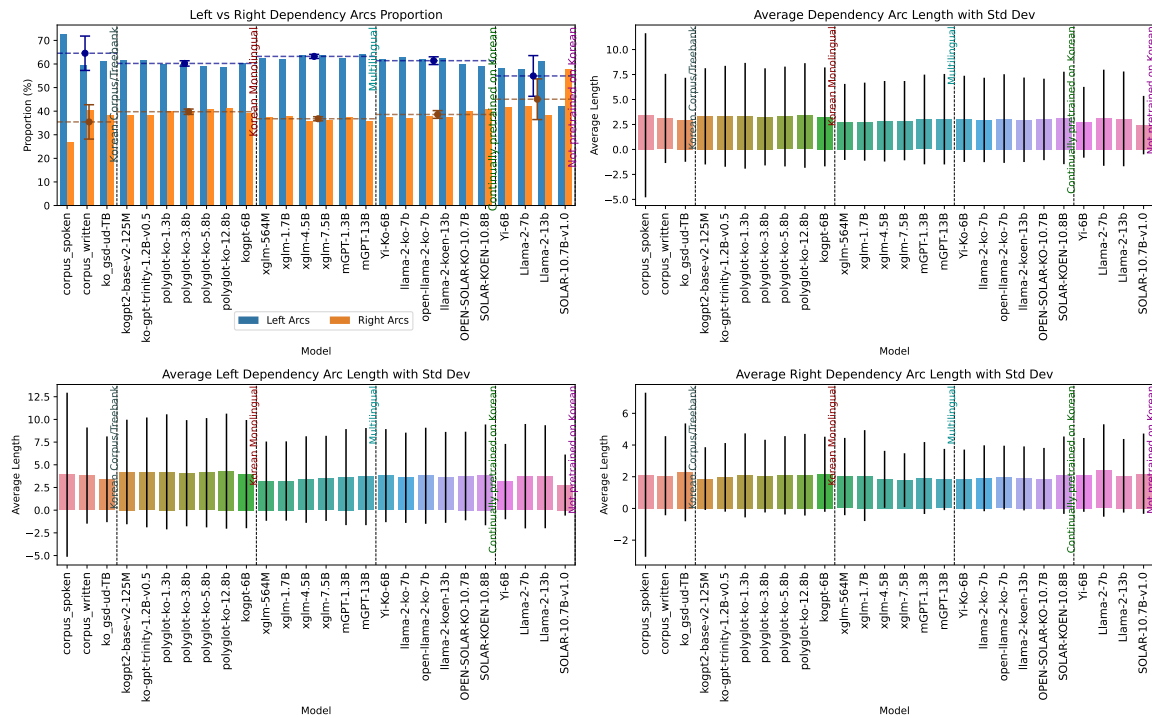


Figure 4.32: Direction and lengths of the dependency arcs in Korean corpora and Korean sentences generated by language models in the zero-shot task.

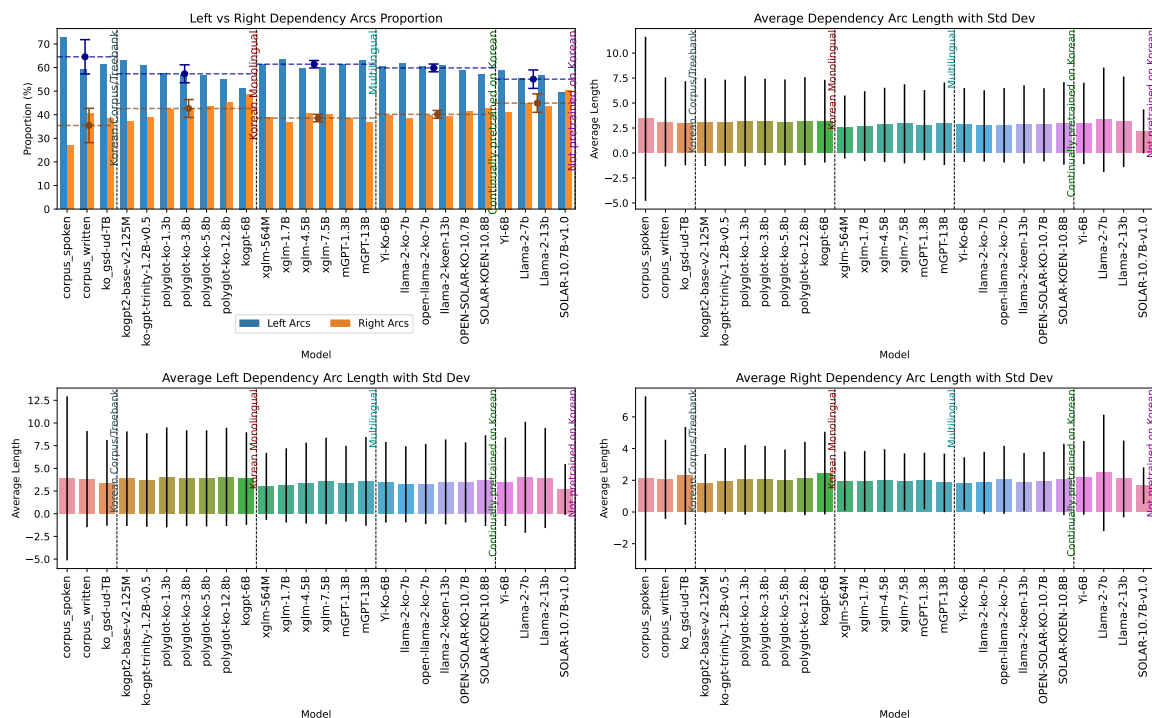


Figure 4.33: Direction and lengths of the dependency arcs in Korean corpora and Korean sentences generated by language models in the few-shot task.

4.2.4 English translationese

This section presents the results of an investigation into whether the Korean sentences generated by language models exhibit characteristics of translationese from English, focusing on several grammatical features related to the differences between Korean and English. For the underlying tables, refer to Table A.30 and Table A.29 in the appendix.

Articles

Articles are an essential grammatical component in English but do not exist as a grammatical category in Korean. When the English articles ‘a/an’ and ‘the’ are directly translated into Korean at the word level, they are translated as ‘han(한)’ or ‘eotteon(어떤)’ (meaning ‘one’ or ‘a certain’) and ‘geu(그)’ (meaning ‘that’), respectively.

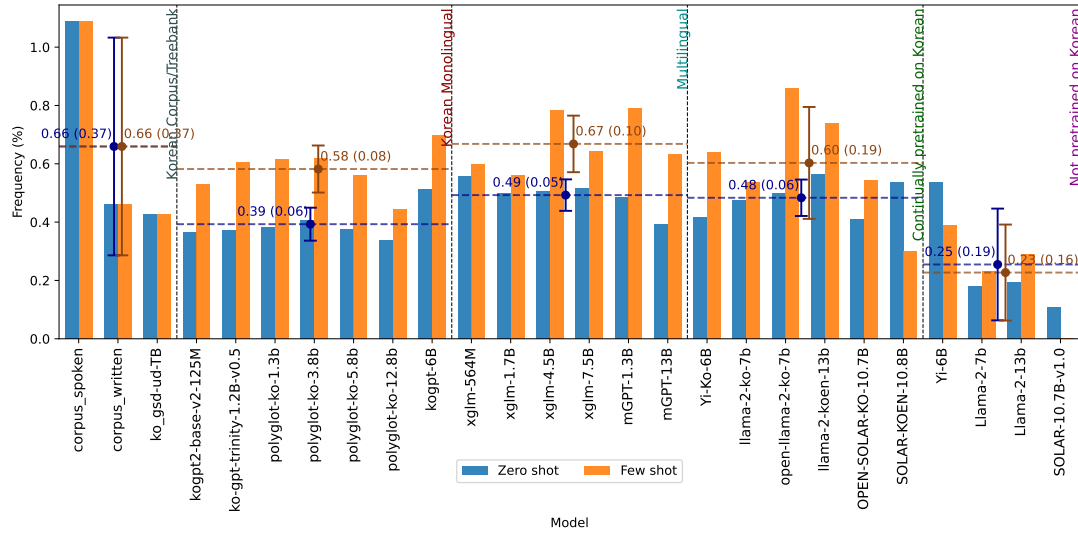


Figure 4.34: Proportion of the word ‘han(한)’ as a translationese of ‘a’ to the nouns in Korean corpora and Korean sentences generated by language models.

Figure 4.34 shows the frequency of the word ‘han(한)’ as a numeric modifier (dependency relation tag: nummod) used as a translationese of the English article ‘a/an’ in Korean corpora and sentences generated by language models, presented as a proportion to the total noun frequency.

The relatively high rate of over 1% in the spoken corpus is noticeable, while it is around 0.4% in the written corpus and treebank. In the language models, the non-Korean-trained models have a somewhat lower rate of around 0.25%, and the Korean monolingual models have a rate of around 0.4-0.6%, which is slightly lower than the multilingual models or continually pre-trained models with rates of around 0.5-0.7%, but there are no remarkable differences between the language model groups, considering the percentages. There is a slight tendency for the frequency to increase in the few-shot setting compared to the zero-shot setting.

The relatively high frequency in the spoken corpus may be influenced by the fact that this word is often used in spoken language to indicate an approximate quantity, meaning

‘roughly’.

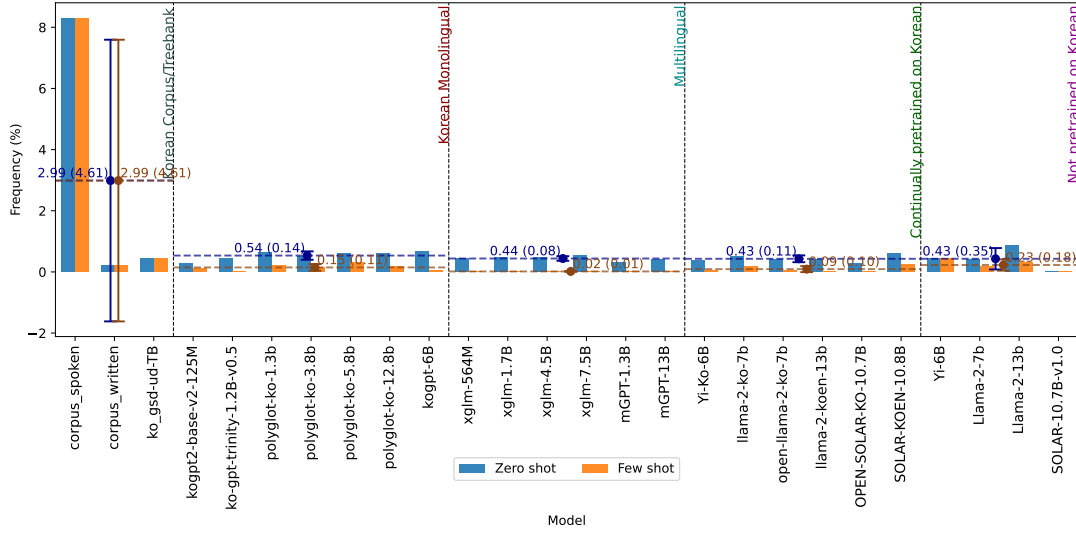


Figure 4.35: Proportion of the word ‘geu(그)’ as a translation of ‘the’ to the nouns in Korean corpora and Korean sentences generated by language models.

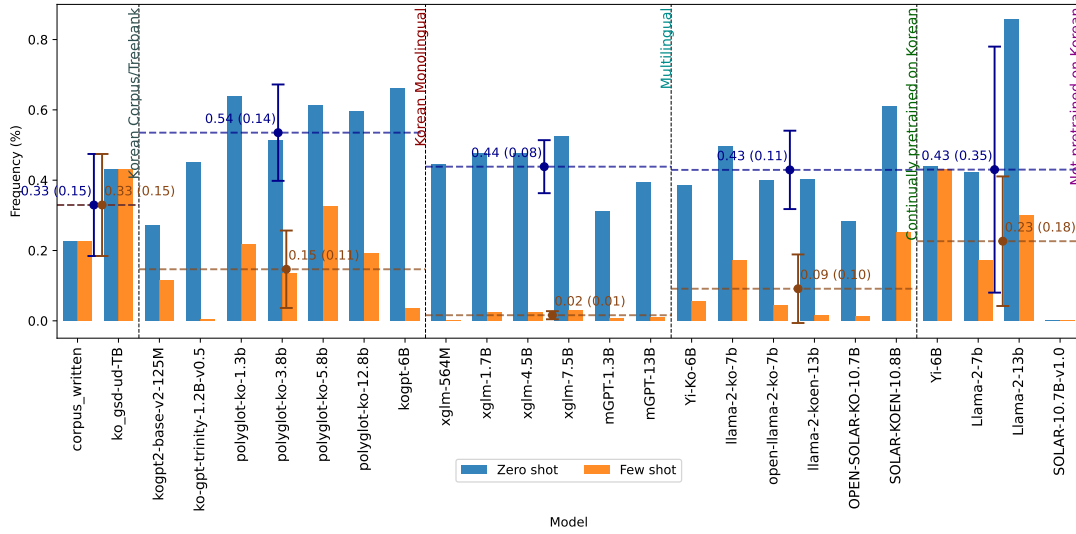


Figure 4.36: Proportion of the word ‘geu(그)’ as a translation of ‘the’ to the nouns in Korean corpora and Korean sentences generated by language models, without the spoken language corpus.

Figure 4.35 shows the frequency of the word ‘geu(그)’ as a determiner (universal part-of-speech tag DET) used as a translation of the English article ‘the’ in Korean corpora and sentences generated by language models, presented as a proportion to the total noun frequency.

The rate in the spoken corpus is significantly high at around 8%, which is because this word is also used as a kind of filler word in spoken language. When the distribution is re-examined excluding the spoken corpus, it appears as shown in Figure 4.36. In the zero-shot setting, the frequency in Korean monolingual models is slightly higher at around 0.5% compared to other model groups at around 0.4%, and in the few-shot setting, there

is a tendency for the frequency to decrease overall. However, considering the percentages, it is hard to say that there are noticeable differences between the language model groups overall.

If there were English translationese in the generated sentences, we expected that translationese of articles would appear with higher frequency in models trained on English data, but the results do not necessarily correspond to that expectation.

Possible reasons for this may include that the target words ‘han(한)’ and ‘geu(그)’ are also frequently used in Korean for purposes other than as translationese of ‘a/an’ and ‘the’, which may have prevented an accurate analysis. It may also indicate that English translationese do not manifest as word-level literal translationese.

Plurals

The natural distinction between singular and plural in English is not essential in Korean. In Korean, the basic noun form, i.e., the singular form, can also be understood as having a collective or plural meaning depending on the context. When English plural forms are translated into Korean at the word level, they are generally translated using the noun-deriving suffix ‘deul(들)’.

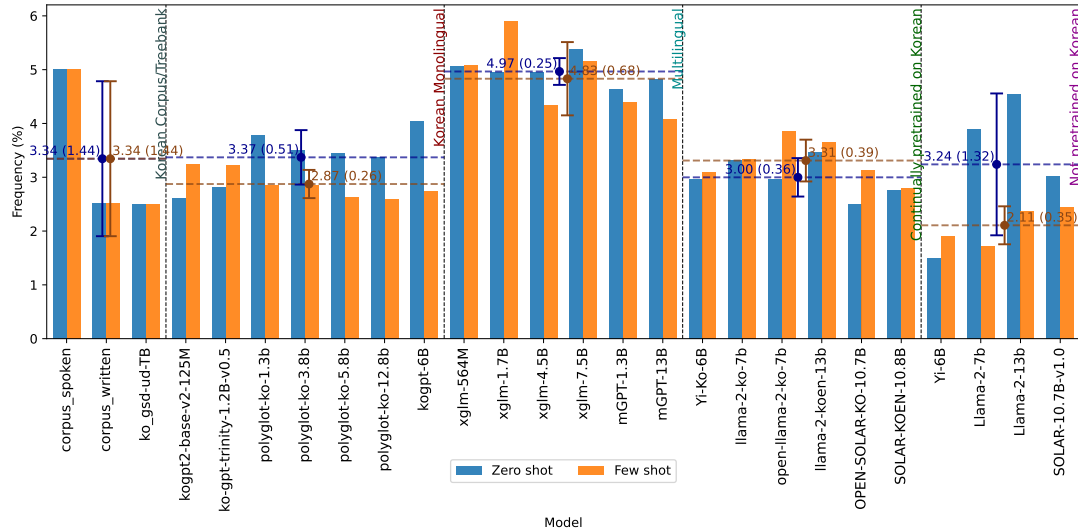


Figure 4.37: Proportion of the word ‘deul(들)’ as a translationese of plurals to the nouns in Korean corpora and Korean sentences generated by language models.

Figure 4.37 shows the frequency of the word ‘deul(들)’ as a derivational suffix (Korean-specific part of speech tag: XSN) used as a translationese of English plural forms in Korean corpora and sentences generated by language models, presented as a proportion to the total noun frequency.

The rate in the spoken corpus and multilingual models is around 5%, which is relatively higher than the written corpus and treebank of around 2.5%, and other language model groups of around 3%. The pattern change between zero-shot and few-shot is not consistent.

As with articles, if ‘deul(들)’ appeared as an English translationese, it could be assumed that the rate would be higher in models trained primarily on English data. Although the frequency is slightly lower in models specifically trained on Korean (Korean monolingual models and Korean continually pre-trained models) than in multilingual models, considering also the low frequencies in models not trained on Korean, further investigation is needed to verify this assumption. Additionally, the high frequency of the word ‘deul(들)’ in spoken language suggests that the composition of the training datasets might also influence the results.

Passive Voice

In English, passive voice are commonly used, but in Korean, passive voice are relatively less recommended.

Figure 4.38 shows the frequency of sentences with passive relations (containing ‘pass’ in the dependency relation tag, but mostly passive nominal subject (nsubj:pass)) in Korean corpora and sentences generated by language models, presented as a rate to the total number of sentences.

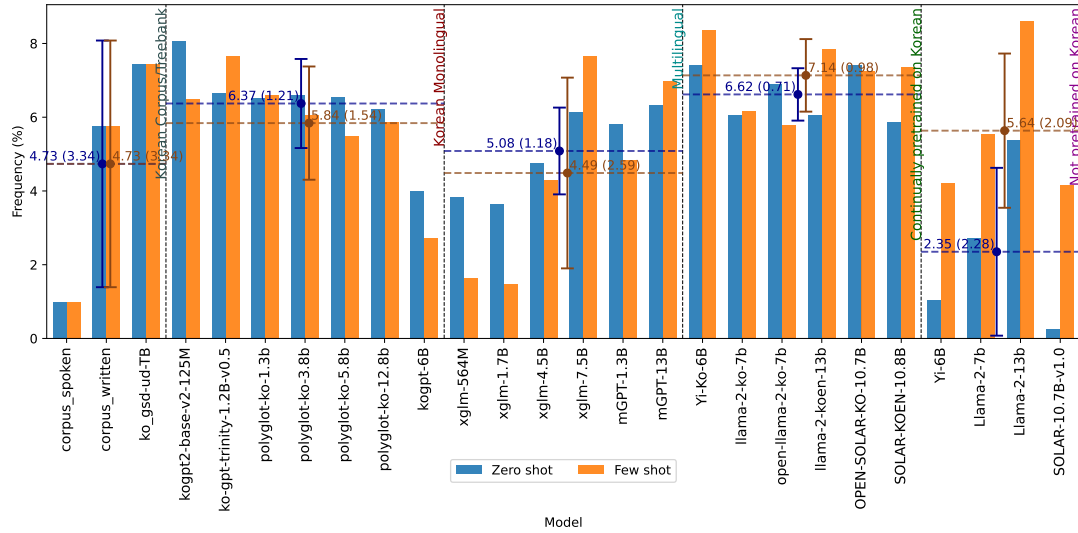


Figure 4.38: Proportion of the passive subjects in Korean corpora and Korean sentences generated by language models.

The rate is generally higher in the written corpus, treebank, Korean monolingual models, and Korean continually pre-trained models at around 6-7%, compared to the spoken corpus, multilingual models, and non-Korean-trained models. There are some variations across individual models in the Korean monolingual models, multilingual models, and non-Korean-trained models. The pattern change between zero-shot and few-shot is not consistent.

It can be observed that the frequency of passive voice is quite high in the written corpus and treebank of around 6-7% compared to the spoken corpus of around 1%. Moreover, upon reviewing the actual generated sentences, passive voices are often used in fact-delivering sentences such as news. Considering these points, it may be difficult to simply

regard passive voice as an English translationese. Rather, the fluent use of passive sentences in accordance with the required style may demonstrate proficiency in Korean. This is supported by the relatively high rates of passive voices in Korean monolingual models and Korean continually pre-trained models that are specifically trained on Korean. It is also supported by the low rates of passive voices of the XGLM-564M and 1.7B models, which performed relatively poorly among the multilingual models, and non-Korean-trained models in the zero-shot setting. Also, the increased rates of passive voices of non-Korean-trained models in the few-shot setting demonstrates their ability to learn Korean sentence structures.

Pro-drop

Unlike English, which is a non-pro-drop language where essential sentence components are rarely omitted, Korean is one of the pro-drop languages where sentences are not grammatically incorrect even if the subject or object is omitted.

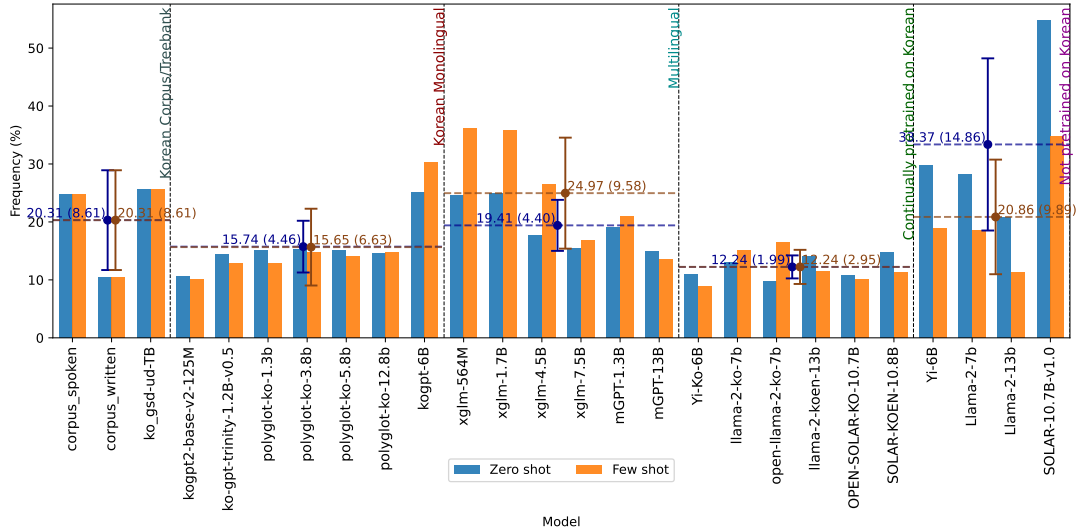


Figure 4.39: Proportion of sentences without subjects in Korean corpora and Korean sentences generated by language models.

Figure 4.39 shows the frequency of sentences without subjects in Korean corpora and sentences generated by language models, presented as a proportion to the total number of sentences. The spoken corpus and treebank have rates of around 25%, and multilingual models and non-Korean-trained models have rates of around 20%-30%, which is higher than the written corpus at around 10% and Korean monolingual models or Korean continually pre-trained models at around 10-15%. Between zero-shot and few-shot settings, there are differences in the pattern of changes across individual models.

This differs from the assumption that sentences without subjects would be more frequent in models trained on Korean. Various factors may contribute to this. Firstly, based on the results from the spoken and written corpora, it can be speculated that spoken language has a higher frequency of sentences without subjects because contextual information is easily utilized, while written language is relatively less so, resulting in a lower frequency of sentences without subjects. For example, upon manual reviewing

the generated sentences by KoGPT-6B, which has a relatively higher frequency among Korean monolingual models, they are more colloquial compared to the sentences by other Korean monolingual models. Considering that the given input prompts were literal texts such as news articles or columns, it might be difficult to simply associate the frequency of subject presence/absence with English translationese, unlike the assumption. Also, the relatively high rate of sentences without subjects in XGLM-564M and 1.7B models may be related to the relatively shorter length of sentences generated by these models (Figure 4.20 and Figure 4.21). Furthermore, the high rate of sentences without subjects of non-Korean-trained models in the zero-shot setting may be due to their difficulty in generating proper Korean sentences.

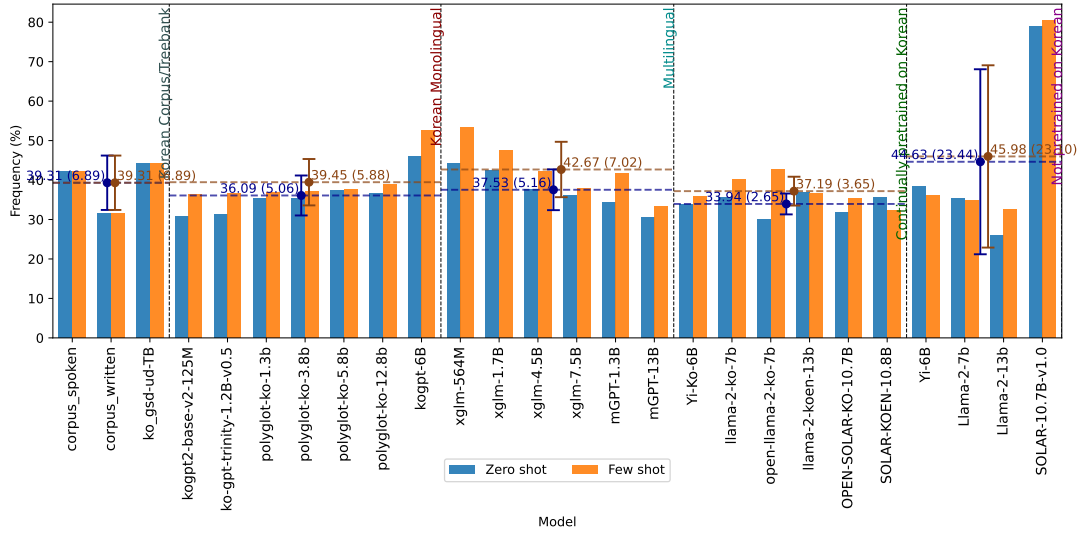


Figure 4.40: Proportion of sentences without objects in Korean corpora and Korean sentences generated by language models.

Figure 4.40 shows the frequency of sentences without objects in Korean corpora and sentences generated by language models, presented as a proportion to the total number of sentences.

The pattern is generally similar to the frequency of sentences without subjects, but the difference is smaller. Considering that the overall rates of sentences without objects reach around 30-50%, it may also be difficult to simply associate the frequency of object presence/absence with English translationese.

Figure 4.41 shows the frequency of sentences without both subjects and objects in Korean corpora and sentences generated by language models, presented as a ratio to the total number of sentences.

It shows a generally similar pattern to the frequency of sentences without subjects among language model groups, although the overall rates have decreased by about 10%. Overall, the rates of sentences without both subjects and objects are around 5-15% of the generated sentences. The high rate of over 40% in the zero-shot setting and over 20% in the few-shot setting of SOLAR-10.7B suggests that many of the Korean sentences generated by this model are short or incomplete.

4.2. Analysis for generated Korean sentences

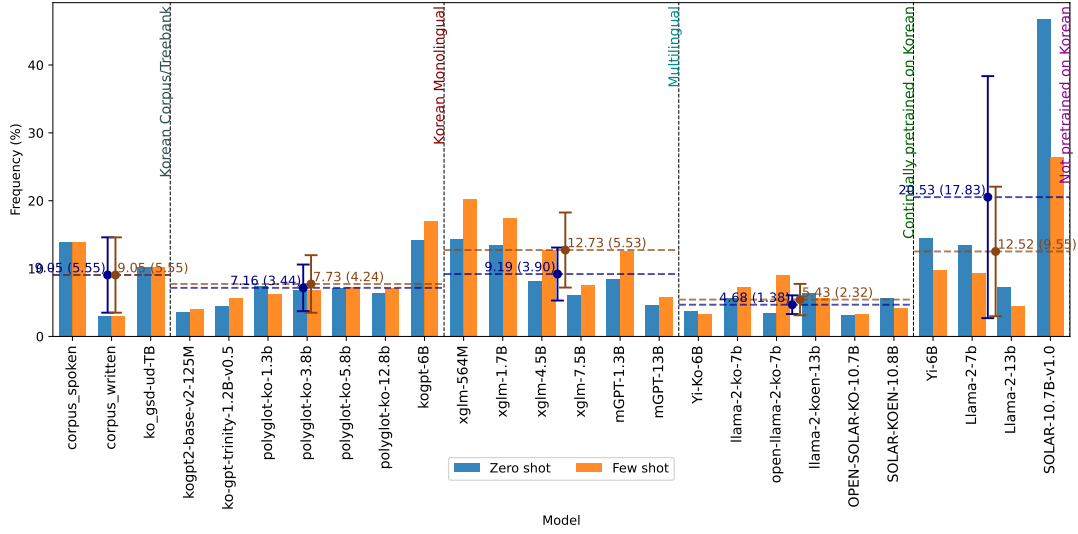


Figure 4.41: Proportion of sentences without both subjects and objects in Korean corpora and Korean sentences generated by language models.

Word order

One of the main differences between English and Korean is that English follows a subject-verb-object (SVO) word order, while Korean follows a subject-object-verb (SOV) word order.

Figure 4.42 and Figure 4.43 show the frequency of sentences with object-verb (OV) word order and verb-object (VO) word order in Korean corpora and sentences generated by language models, presented as a proportion to the number of sentences containing both objects and verbs.

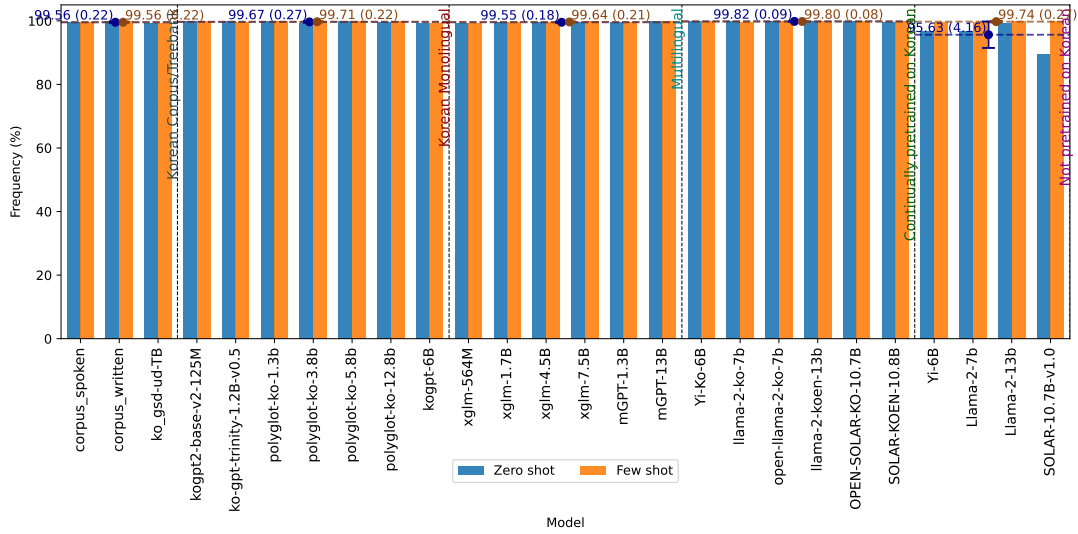


Figure 4.42: Proportion of Object-Verb order sentences to the sentences with object and verb in Korean corpora and Korean sentences generated by language models.

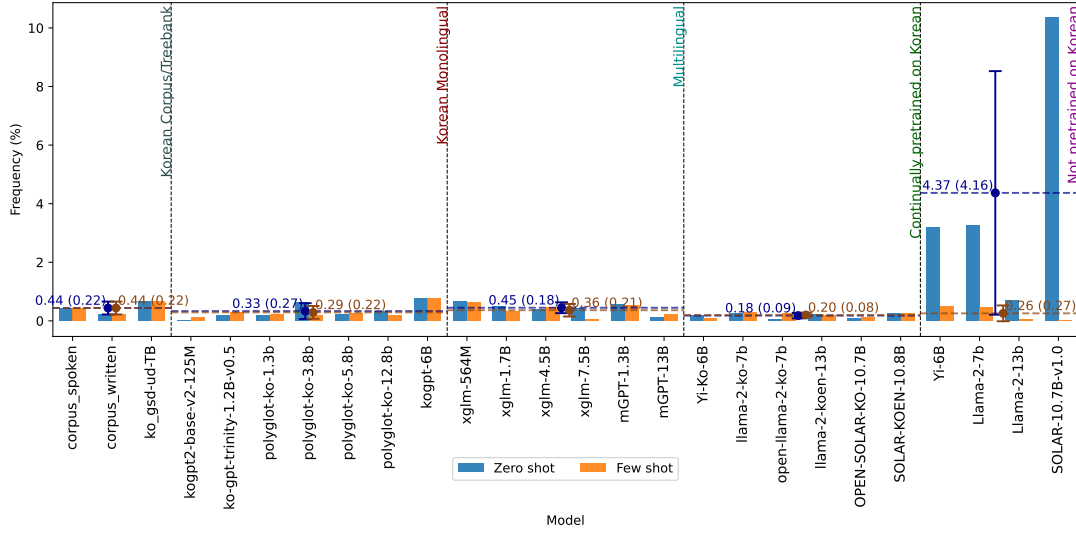


Figure 4.43: Proportion of Verb-Object order sentences to the sentences with object and verb in Korean corpora and Korean sentences generated by language models.

It is observed that most models generate Korean sentences with an object-verb word order, accounting for more than 99%. Among the models not trained on Korean, in the zero-shot setting, the Yi-6B and Llama-2-7B models show a frequency of verb-object word order sentences of around 3%, and the SOLAR-10.7B model shows a frequency of around 10%.

Upon examining the sentences classified as having a verb-object word order, it was found that they were either due to words being incorrectly tagged as something else when they should have been tagged as verbs, the verb of the second clause being omitted in a sentence with multiple clauses, or the sentence ending with a nominalized form. There were almost no sentences that actually had a verb-object word order. The relatively lower rates of non-Korean-trained models in the zero-shot task suggest that the generation of such incomplete sentences was higher for non-Korean-trained models in the zero-shot task.

In other words, there were hardly any instances of directly adopting the English word order or mixing the word orders of the two languages in the generated sentences. This suggests that language models are learning the syntactic rules of Korean quite well, and even if there is an influence from the English training data, the influence does not directly manifest at the surface level, such as word order.

4.2.5 Semantic Evaluation

This section presents the results of a semantic evaluation of Korean sentences generated by language models, conducted through sentiment classification. Considering that some of the sentences generated by language models may not convey clear meaning, and that the semantics of a sentence greatly influence its overall quality, semantic evaluation of sentences is crucial. We recognize the need for a comprehensive evaluation that goes beyond sentiment classification, assessing aspects such as fluency, sentence coherence,

and appropriateness of responses. However, due to time and resource constraints, this study could not address these aspects, and we plan to leave a more comprehensive semantic evaluation as a follow-up research.

Sentiment analysis

Figure 4.44 shows the proportion of sentences from Korean corpora and those generated by language models classified as positive, negative, or neutral in terms of sentiment. For comparison, the y-axis is displayed with the same intervals. Refer to Table A.31 in the appendix for the underlying table.

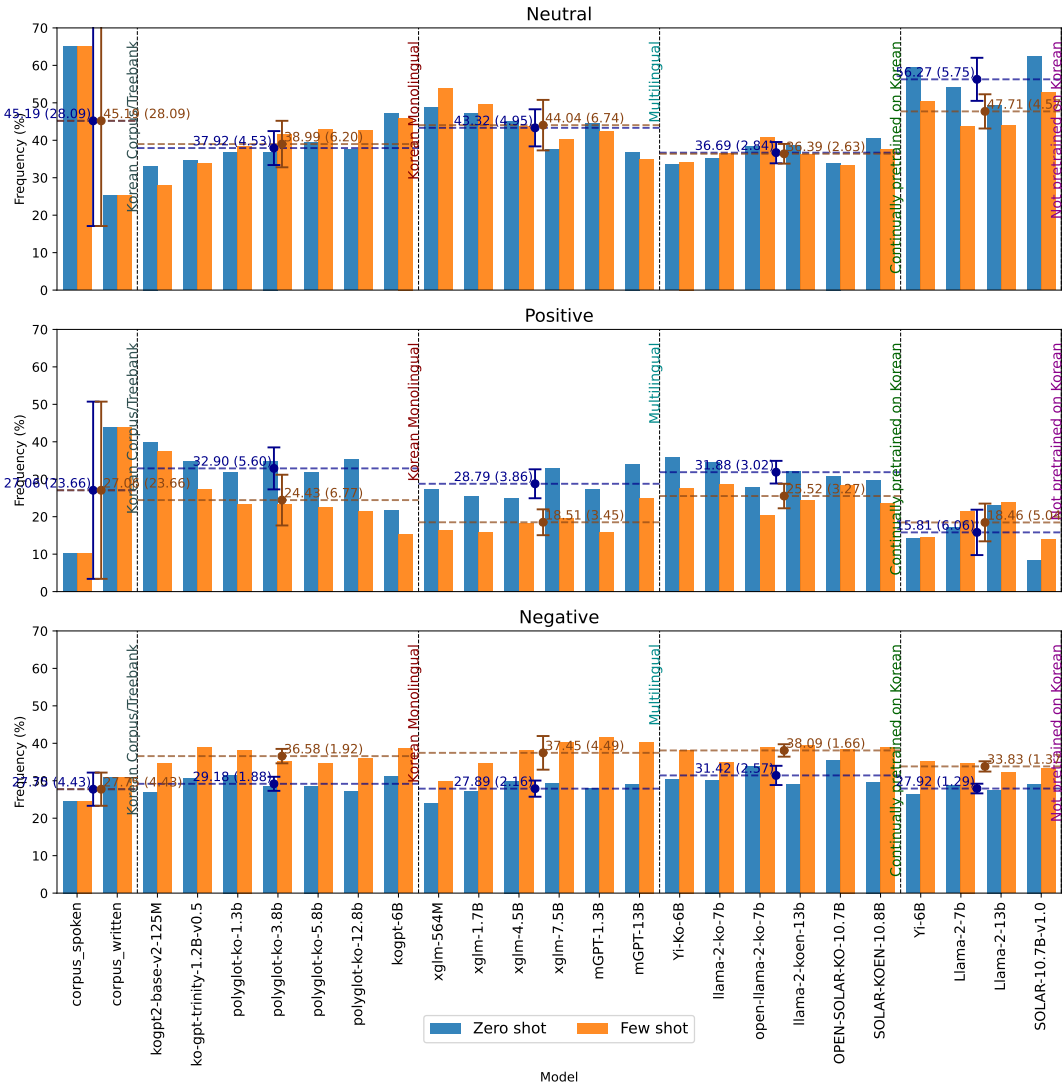


Figure 4.44: Distribution of sentiment classification in Korean corpora and Korean sentences generated by language models.

There is a significant difference between spoken and written corpora. In the spoken corpus, the neutral ratio is around 65%, with around 10% positive and 25% negative, while in the written corpus, the neutral is around 25%, positive is around 45%, and

negative is around 30%. In the sentences generated by language models, the neutral is around 35-50%, positive is around 15-30%, and negative is around 30-40%. Compared to the written corpus, the sentences generated by language models have a higher proportion of neutral sentiment and a lower proportion of positive sentiment, and this tendency is more prominent in multilingual models and especially, non-Korean-trained models. The change between zero-shot and few-shot settings is also noticeable. In most language models, the few-shot setting shows a decrease of about 10% in the positive sentiment and an increase of about 10% in the negative sentiment compared to the zero-shot setting.

The high neutral ratio in the spoken corpus may be due to the inclusion of many short sentences. XGLM-564M, 1.7B models, and models not trained for Korean, which tended to have relatively short sentence lengths (Figure 4.20 and Figure 4.21), also show a slightly higher neutral ratio.

The higher neutral sentiment ratio and lower positive sentiment ratio in sentences generated by language models compared to the written corpus may suggest that language models generate relatively fewer sentences expressing sentiments. This trend is stronger in multilingual models and non-Korean-trained models, which could be influenced by the characteristics or size of the language data they are trained on. Additionally, considering that non-Korean-trained models generated somewhat incomplete sentences, the sentences that are difficult to understand semantically may be classified as neutral.

It is interesting that the negative sentiment ratio generally increases in the few-shot setting compared to the zero-shot setting. The input prompts provided in the few-shot setting may have had an influence. Although the examples themselves (Table 3.2) in the few-shot setting were not particularly negative,⁸ this might suggest that the examples can influence the model in generating more emotionally expressive sentences, implying that they could affect how the model processes sentiments in sentence generation or provide a specific direction.

⁸according to the sentiment classification model used, they are classified as positive, positive, and neutral, and according to our manual review, they are classified as positive, negative, and positive.

Chapter 5

Discussion

In this section, we discuss the answers to the research questions by synthesizing the results from Chapter 4, along with further findings derived from the results, limitations of the study, and future works.

5.1 Answers to Research Questions

5.1.1 Differences in generated texts

RQ1. When various types of language models are prompted to generate text in Korean, what texts do the language models generate?

Language models trained on Korean (Korean monolingual models, multilingual models, Korean continually pre-trained models) were able to answer in Korean without much difficulty in both zero-shot and few-shot settings (Figure 4.3, Figure 4.4). Language models not trained on Korean required longer Korean input prompts to generate Korean text (Figure 4.3, Figure 4.4). This suggests that models lacking prior knowledge of Korean need more information and context to generate Korean text. In the zero-shot setting, models not trained on Korean often failed to generate Korean and generated primarily Latin alphabet characters, producing text in English or sometimes a mix of English and Korean. A manual review of the generated text indicates that it is not entirely unrelated to the input prompt, suggesting that even if the model fails to generate Korean, it understands the content of the provided Korean input prompt to some extent.¹ In the few-shot setting, when provided with longer Korean prompts containing three examples, models not trained on Korean showed highly improved Korean generation capabilities (Figure 4.4). Models not trained on Korean have very few Korean tokens in their vocabulary, so they attempt to assemble byte-level tokens to generate Korean characters (Figure 4.6), which sometimes results in the generation of characters that do not exist in Korean or only partial Korean characters.

¹A deeper exploration of this is also an interesting research topic, but as it goes beyond the scope of the study, it is left for future work.

In terms of the length of the generated texts or the number of tokens, there are no substantial differences among the model groups in the zero-shot setting (Figure 4.1, Figure 4.2).² However, in the few-shot setting, there is a tendency for the text length to decrease in the Korean monolingual models, Korean continuously pre-trained models, and non-Korean-trained models (Figure 4.1). The number of tokens increases in the non-Korean-trained models (Figure 4.2), but this is because these models lack Korean tokens and need to combine byte-level tokens to create Korean characters. The length of the generated text in these models is notably reduced compared to the zero-shot setting (Figure 4.1). It is speculated that the influence of the prompt leads to the compressed generation of sentences similar in length to the prompt examples. Also, as shown here, non-Korean-trained models require 2-3 times more tokens than Korean-trained models to generate Korean text of the same length. This suggests that the efficiency of processing Korean is relatively lower in non-Korean-trained models. If the cost is charged proportionally to the number of tokens, using non-Korean-trained models may result in higher expenses.

At the surface level, there were some differences in the generated sentences among the language model groups. Multilingual models had a wider distribution of Korean character ratios within the generated sentences (Figure 4.10), a higher rate of including unusual patterns for sentences, such as URLs, emails, news bylines, and symbols (Figure 4.11), and a higher percentage of sentences lacking typical Korean sentence-ending formats (Figure 4.14). These tendencies were more prominent in smaller models and tended to decrease as the model size increased. Models not trained on Korean showed a substantial difference in Korean character ratios in the generated sentences and the percentage of sentences with proper Korean sentence-ending formats compared to other model groups in the zero-shot setting (Figure 4.10, Figure 4.14). However, in the few-shot setting, their distribution became quite similar to that of other Korean-trained models, close to that of multilingual models. One of the models (SOLAR-10.7B-v1.0) was an exception and did not learn well to generate typical Korean.

At the semantic level, Korean monolingual models and Korean continually pre-trained models generated texts more similar to the original text compared to multilingual models and non-Korean-trained models (Figure 4.16, Table 4.3). In the zero-shot setting, Korean monolingual models and Korean continually pre-trained models showed almost similar scores, followed by multilingual models, with non-Korean-trained models falling far behind. In the few-shot setting, the scores of non-Korean-trained models increased significantly, resulting in the order of Korean continually pre-trained models, Korean monolingual models, non-Korean-trained models, and multilingual models. The visualization of the semantic embedding distributions of the original text and the generated texts also supports this (Figure 4.17, Figure 4.18).

In summary, Korean monolingual models and Korean continually pre-trained models (i.e., Korean bilingual/trilingual models generated more formally complete sentences and produced texts that were semantically more appropriate to the task. However, as the model size increased and depending on the architecture, rapid Korean learning ability was also observed in few-shot examples even without pre-training. Therefore, with changes in model architecture or further increases in model size, and with the provision of more input examples, there is potential for multilingual models and non-Korean-

²The non-Korean-trained models are given a maximum number of new tokens twice as large, so they cannot be compared on the same basis.

trained models to produce superficially complete and semantically task-appropriate results. Considering the amount of Korean data in the pre-training data of multilingual models, the results of multilingual models and non-Korean-trained models suggest that a large amount of Korean data may not be necessary to include in the pre-training data, but a minimum level of data should be included.³

5.1.2 Differences in generated Korean sentences

RQ2. Are there differences in the Korean sentences generated by various types of language models?

The average text length and the number of words in the generated Korean sentences did not differ substantially between the language model groups and the Korean corpora (Figure 4.20, Figure 4.21), although there were slight variations within the language model groups. Non-Korean-trained models generated shorter sentences in the zero-shot setting, but in the few-shot setting, they generated sentences of similar lengths to other language model groups.

In terms of lexical diversity based on the morphemes of the generated Korean sentences, the Korean continually pre-trained models showed the highest diversity, followed by the Korean monolingual models, and then the multilingual models and non-Korean-trained models, which exhibited almost similar levels of lexical diversity (Figure 4.23). However, it should be noted that the differences were not substantial, and there were within-group variations in the Korean monolingual models and non-Korean-trained models. Furthermore, the high lexical diversity of the non-Korean-trained models might indicate the generation of non-existent or nonsensical vocabulary at the word level, which requires additional analysis.

From a syntactic perspective, in terms of the distribution of universal POS, Korean-specific POS, dependency relations, and the direction and length of dependency arcs in the generated sentences, similar patterns were repeatedly observed. There seems to be no substantial difference in the Korean sentences generated by the language model groups trained on Korean. The Korean sentences generated by non-Korean-trained models show differences from those generated by other models in the zero-shot setting, but in the few-shot setting, the difference decreases, and the distribution becomes quite similar to those of other groups. Instead, in the few-shot setting, the variance within the language group slightly increases, especially in the Korean monolingual models. (from Figure 4.24 to Figure 4.33)

This may be related to whether the style of the generated text is colloquial or literary. Actually the difference between the spoken and written corpora was greater than the difference between the natural language and the texts generated by the language models in syntactic evaluation (Figure 4.25, Figure 4.27, Figure 4.27, Figure 4.31, Figure 4.33). It seems that the characteristics of individual models are further triggered in the few-shot setting.

Regarding translationese from English, it can be said that there were limitations in

³Research on the specific amount or ratio was beyond the scope of this study, but we believe this could be a topic for future research.

accurate measurement due to the influence of several different factors. The differences in the frequency of words corresponding to definite/indefinite article translations among language model groups were confounded with other usages of the words (Figure 4.34, Figure 4.36), and passive voices were also frequently used in Korean depending on the text style, such as fact-delivering news (Figure 4.38). The characteristics of pro-drop languages, such as subject or object omission, were less evident in written language where the use of contextual information is limited, and there were also some noises such as colloquial texts or actually incomplete sentences (Figure 4.39, Figure 4.40, Figure 4.41). For the basic differences between Korean and English, such as object-verb/verb-object word order, most of the sentences generated by the models followed the Korean word order (Figure 4.42).

However, even considering the need for more accurate analysis that controls for these various factors, the above results suggest that the influence of English in Korean sentences generated by language models does not appear to be very evident at the surface-level, such as in vocabulary and word order. This is somewhat different from another study that suggest multilingual models generate more sentences with English nuances compared to monolingual models (Papadimitriou et al., 2023). Although there are limitations in making direct comparisons due to differences in experimental methods, it can be speculated that the language differences between English and Korean are greater than the language differences between English and Spanish or Greek tested in the study, making it more difficult for the influence of English to directly affect the generated sentences.

From a semantic perspective, in the sentiment analysis of the generated Korean sentences, the sentences generated by multilingual models and non-Korean-trained models were classified as neutral at a rate about 6-10% higher than the sentences generated by the Korean monolingual model and the Korean continually pre-trained models. (Figure 4.44) The sentences generated by these models may express less sentiment due to the influence of the training data, or they may be semantically unclear sentences classified as neutral sentiment. Overall, the sentences generated by language models had a neutral sentiment rate more than 10% higher and a positive sentiment rate about 10% lower than the Korean written corpus in the zero-shot setting. In the few-shot setting, the positive sentiment rate decreased by about 10%, and the rate classified as negative sentiment increased by about 10%, which is interesting in terms of how the input prompt in the few-shot setting influences the generation of Korean sentences expressing sentiment.

In summary, there were no substantial differences in sentence length and syntactic aspects among the Korean sentences generated by various language model groups. However, there was a tendency for within-group variance to increase in the few-shot setting. In terms of lexical diversity and sentiment analysis, multilingual models showed slightly lower scores and a higher proportion of neutral sentences, respectively, compared to Korean monolingual models and Korean continually pre-trained models. Regarding the translationese from English, although there were some limitations in accurate analysis, English translationese at the surface-level, such as vocabulary or word order, did not seem to be evident. It should be noted that the filtered Korean sentences, which were the target of evaluation, contain some noise such as nonsensical vocabulary and sentences. Further evaluation of the Korean sentence quality in terms of semantic aspects is necessary.

5.2 Further Findings

5.2.1 Influence of Few-Shot Prompts

The input prompts provided in the few-shot task influenced the generation of Korean text in various dimensions. As observed in previous results, in the few-shot setting, the non-Korean-trained models generate outcomes that are considerably close to outcomes of the Korean-trained models (with the exception of the SOLAR-10.7B model). Few-shot prompts contributed to enhancing the ability of non-Korean-trained models to generate Korean (Figure 4.4, Figure 4.10), producing plausible Korean sentences at the surface level (Figure 4.15), and generating sentences that are more appropriate for the intent of the task at the semantic level (Figure 4.16). They also helped the non-Korean-trained models learn the syntactic composition of Korean sentences (Figure 4.25, Figure 4.27, Figure 4.29, Figure 4.31) and generate sentences that can be more clearly interpreted sentimentally (Figure 4.44). In most evaluations conducted in this study, the non-Korean-trained were influenced by the few-shot prompts.

Just as few-shot prompts helped non-Korean-trained models learn Korean, few-shot prompts also influenced Korean-trained models in various ways and degrees. At the surface level, the headline patterns included in the few-shot prompts appeared significantly in the generated texts (Figure 4.12, Figure 4.13). Semantically, the few-shot prompts contributed to obtaining higher similarity with the original text by enabling the models to learn the relationships of the examples in the task (Table 4.3). The few-shot prompts also affected the sentimental expression of the generated sentences, such as reducing the generation of sentences classified as positive sentiment and increasing the generation of sentences classified as negative sentiment (Figure 4.44).

Few-shot prompts also drove the models in diverse directions and to varying degrees, resulting in different patterns of change in their results compared to the zero-shot setting, which increased the within-group variance (Refer to many few-shot task results, including Figure 4.12, Figure 4.13, Figure 4.15, etc.). This demonstrated that even with the same few-shot prompts, the direction and degree of influence on the models could differ, which seems to be related to the characteristics of each individual model.

This study did not investigate how changes in the number or modifications of few-shot prompts might affect the outcomes. However, the findings suggest that few-shot prompts can significantly influence the performance and characteristics of language models. In particular, improvements in the Korean generation capabilities of non-Korean-trained models and the increase in within-group variance in Korean-trained models demonstrate the potential of optimizing language model performance through prompt engineering. Therefore, exploring how prompts with various forms, lengths, and contents operate on language models and their impact on model performance could be a promising topic for future research.

5.2.2 Influence of Model Size

A noticeable finding observed in the experimental results was the impact of language model size on performance. Among the models used in the experiment, the Korean monolingual models and multilingual models included relatively small-sized models compared to the Korean continually pre-trained models and non-Korean-trained models. While all the Korean continually pre-trained models and non-Korean-trained models were over 6B parameters in size, only 2 out of 7 Korean monolingual models and 2 out of 6 multilingual models were over 6B parameters (Table 3.1). These differences in model size may have influenced the performance of these model groups. In the comparisons between language model groups, the Korean continually pre-trained models generally showed good performance in various evaluations despite some variations within the group. The non-Korean-trained models also demonstrated remarkable Korean learning ability in the few-shot setting, except for SOLAR-10.7B.

Comparisons of results from the models that share the same architecture but differ in size support this finding. For the multilingual models, the XGLM and mGPT models generally showed better performance with increasing model sizes (Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.10, Figure 4.11, Figure 4.14, Figure 4.15, Table 4.3, Figure 4.17, Figure 4.18, Figure 4.43, Figure 4.44). Similarly, among the non-Korean-trained models, the Llama2 models also exhibited better performance with larger model sizes (Refer to the previous sources). In the Korean monolingual models, the results for the Polyglot-Ko models were inconsistent, but these models were not trained on the same size of data, limiting direct comparisons (Table 3.1). When examining the Polyglot-Ko 1.3b and 3.8b models, which were trained on almost similar-sized data, the 3.8b model generally showed slightly better performance (Figure 4.4, Figure 4.6, Figure 4.11, Figure 4.14, Figure 4.15, Figure 4.22, Figure 4.23), although there were cases where the 1.3b model performed partially a bit better in the few-shot setting (Table 4.3, Figure 4.44).

The impact of language model size on performance has been introduced in several studies (Devlin et al., 2019; Brown et al., 2020; Kaplan et al., 2020; Rae et al., 2022). In this study, overall, the Korean monolingual models and Korean continually pre-trained models showed better results than the multilingual models or non-Korean-trained models in terms of the generated outcome itself and the quality of generated Korean sentences. However, if the model size increases, there is potential for changes in the performance differences between these groups.

Another interesting finding is that KoGPT2, a 125M-sized Korean monolingual model, generated fluent Korean sentences at the surface level but showed relatively low performance at the semantic level (Figure 4.15, Table 4.3). In contrast, Llama-2-13b, a 13B-sized non-Korean-trained model, demonstrated decent performance at the surface level and even surpassed some Korean monolingual models and Korean continually pre-trained models at the semantic level in the few-shot setting (Figure 4.15, Table 4.3). This suggests that learning at the surface-level is possible even with small-sized language models, but learning at the semantic level may require larger-sized language models. Future research could involve using models of various sizes to analyze the impact of model size on the performance of specific tasks across a range of language processing applications.

5.2.3 Influence of Pre-training Data

The pre-training data of the models also appeared to influence their performance and their individual characteristics. This can be examined in more detail between or within model groups as follows.

Differences within Korean Monolingual Models

Korean monolingual models often showed some variations within the group, especially in the few-shot setting. For example, among the Korean monolingual models, KoGPT2, KoGPT-Trinity, and KoGPT models seemed to have some differences in the characteristics of the generated Korean sentences. In several results, KoGPT2 and KoGPT-Trinity exhibit characteristics closer to the written corpus, while Ko-GPT shows characteristics more similar to the spoken corpus (Figure 4.38, Figure 4.39, Figure 4.40, Figure 4.41, Figure 4.44). However, in some results, Ko-GPT also displays aspects that differ from the spoken corpus, such as having a higher noun ratio (Figure 4.25, Figure 4.27, Figure 4.29, Figure 4.31, Figure 4.33). Upon manual inspection, the texts generated by Ko-GPT are more colloquial, while those generated by KoGPT2 and KoGPT-Trinity are more literary. Although the training datasets of these models are not publicly disclosed, it can be assumed that the Korean datasets on which they were trained contributed to these results. Regarding the Polyglot-Ko models, the amount of pre-training data for these models varies, and possibly the data composition was different as well. This may be why, despite sharing the same architecture, these models do not exhibit consistent performance with changes in model size.

Differences within Multilingual Models

The two model series that constitute the multilingual models, XGLM and mGPT, generally showed better performance as the model size increased in most evaluations. Between XGLM and mGPT, mGPT generally showed better results (Table 4.2, Table 4.3, and others). One potential reason that could explain this is that the Korean data in XGLM was “inadvertently over-sampled”, as noted in Lin et al. (2022), which may have led to the model overfitting on these specific data. When manually reviewing the data generated by the XGLM models, certain text styles such as news bylines or ellipses are quite evident, and this is observed as a relatively high proportion of unusual patterns (Figure 4.11). This seems to be the effect of the Korean data the model is pre-trained on, and it may have been involved in the generation results or performance of the XGLM models.

Differences within Korean Continually Pre-trained Models

The Korean datasets used for continual pre-training can be categorized into open-access data (15B tokens), data including non-public sources (40B tokens), and Korean-English 1:1 mixed data (approximately 30B tokens for Korean). However, the impact of these data differences on model performance was inconsistent.

Looking at the results between models based on the same architecture, Llama-2-ko-7B, trained on 40B tokens of Korean data, and Open-Llama-2-ko-7B, trained on 15B tokens of open-access Korean data, showed varying results with each model outperforming the other in different evaluations. However, Open-Llama-2-Ko-7B generally showed a tendency to be more sensitive to few-shot prompts. In the case of OPEN-SOLAR-KO-10.7B, trained on 15B tokens of Korean open-access data, and SOLAR-KOEN-10.8B, which was trained on approximately 30B tokens for Korean in a 1:1 mixed data of Korean and English, OPEN-SOLAR-KO-10.7B generally showed better performance compared to SOLAR-KOEN-10.8B.⁴ OPEN-SOLAR-KO-10.7B was also sensitive to few-shot prompts, but in a different manner from Open-Llama-2-ko-7B.

The impact of training data size or proportion in Korean continual pre-training was not clearly identified. For example, even when trained on the same Korean-English mixed data, the performance changes of Yi-Ko-6B, SOLAR-KOEN-10.8B, and Llama-2-koen-13B compared to their base models were all different. Similarly, even when trained on the same Korean open-access data, the performance changes of Open-Llama-2-ko-7B and OPEN-SOLAR-10.7B compared to their base models were different (Figure 4.15, Figure 4.16, and many others). This may suggest that the size and architecture of the base model, as well as the size of the pre-training data for the base model and the inclusion or exclusion of Korean data in the base model’s pre-training data, all interact in a complex manner. Given the decent performance of Korean continually pre-trained models from the experiments, continual pre-training could be a useful strategy for training language models. Therefore, more research is needed on the impact of training data size and composition in continual pre-training.

Differences within Non-Korean-Trained Models

Among the non-Korean-trained models, Yi-6B, Llama-2-7b, and Llama-2-13b quickly learned the ability to generate Korean text in the few-shot setting, while SOLAR-10.7B seemed to struggle with acquiring Korean text generation skills even in the few-shot setting. This could be influenced by the composition of their pre-training data. It has been reported that Llama includes a small amount of Korean data in its training data (Touvron, Martin, et al., 2023), which may explain the relatively fast adaptation to Korean by Llama-based models. Although Yi-6B was trained on English and Chinese data, the lexical affinity between Chinese and Korean might have helped Yi-6B adapt more easily to Korean. The pre-training data of SOLAR-10.7B or Mistral-7B, which SOLAR-10.7B used as initial weights, is unknown but presumed to be mainly English, which could be the reason why SOLAR-10.7B faces more difficulty in learning Korean. However, considering that the SOLAR-KO-10.7B model, which was continually pre-trained on 15B of Korean open-access data using SOLAR-10.7B as the base model, shows significantly improved performance (Figure 4.15, Figure 4.16, Figure 4.44), it is suggested that continual pre-training for a specific language can work regardless of whether the base model’s pre-training data includes data from that language.

⁴SOLAR-KOEN displayed spacing issues, but the cause remains unclear, as other models trained on the same Korean-English mixed data did not exhibit such problems.

Comparison between Korean Continually Pre-trained Models and Non-Korean-Trained Models

Korean continually pre-trained models and non-Korean-trained models are based on the same model architectures and differ only in the additionally pre-trained data. The performance differences observed between Korean continually pre-trained models and non-Korean-trained models in various evaluations in the study (from Figure 4.1 to Figure 4.44) suggest that additional pre-training using Korean data on base language models greatly improves performance on Korean language tasks.

Comparison between Korean Monolingual Models and Korean Continually Pre-trained Models

When comparing Korean monolingual models and Korean continually pre-trained models of the same size, Korean continually pre-trained models generally performed better (Table 4.2, Table 4.3). One possible reason for this could be the total amount of pre-training data involved. Although Korean monolingual models have a larger amount of Korean training data, Korean continually pre-trained models have a larger total amount of training data⁵ (Table 3.1, Touvron, Martin, et al., 2023; AI et al., 2024; D. Kim et al., 2024). If a sufficient amount of continual pre-training data is provided for a particular language, it appears that the knowledge learned from other languages could indirectly influence the model’s performance on that language. Additionally, in terms of architecture, these models do not have the same architecture, and the architectures of the Korean continual pre-training models are relatively more recent, which could have influenced the results (Table 3.1).

Comparison between Multilingual Models and Non-Korean-Trained Models

When comparing multilingual models and non-Korean-trained models of the same size, in a semantic level, non-Korean-trained models such as Llama-2-7B and Llama-2-13B showed almost similar or slightly higher scores than multilingual models like XGLM-7.5B and mGPT-13B in the few-shot setting (Table 4.3). Considering that Llama-7B and Llama-13B models were not trained on Korean, it is impressive that they show such performance with only few-shot prompts. This could also be possibly due to the total amount of pre-training data of the model. Similar to the comparison between Korean monolingual models and Korean continually pre-trained models, multilingual models have a larger amount of Korean training data, but Llama-2 models have a larger total amount of pre-training data (Table 3.1, Touvron, Martin, et al., 2023). Here, too, it appears that the knowledge learned from other languages might influence the model’s ability to acquire and adapt to other languages. Also, the architectures of these models are not identical, and the architectures of the non-Korean-trained models are more recent, which could have also had an impact. In addition, the pre-training data of the Llama-2

⁵Although the total amount of pre-training data for the SOLAR-10.7B model or the Mistral-7B model, which SOLAR-10.7B used as initial weights, is not known, it is assumed that they were likely trained on mainly English data, and thus it can be speculated that the size of their training data is probably larger than the Korean data size used for the Korean monolingual models.

model includes a very small amount of Korean data, which may have contributed to its quick adaptation to the Korean language.

5.3 Limitations & Future Work

This study has the following limitations that need to be addressed or expanded upon in future work:

5.3.1 Limitations in Evaluation Methodology

This study aimed to investigate whether language models trained on English-centric data exhibit linguistic biases when generating Korean text. As no established benchmarks or evaluation methods were readily available for such an analysis, we attempted a variety of assessments from different dimensions: surface-level, lexical, syntactic, semantic aspects, and translationese from English.

However, we acknowledge that these evaluations were not conducted in a more systematic and comprehensive manner. Especially, if more extensive semantic evaluations had been conducted on the generated Korean sentences, a more integrated understanding could have been obtained, along with the results from other dimensions.

Also, conducting qualitative analysis through manual examination of the generated texts, along with human evaluation, would have enabled a more comprehensive evaluation of the texts and an integrated interpretation of the experimental results.

In the analysis of Korean sentences, the study also has limitations in that data noise was not sufficiently removed, and other influencing factors were not controlled for, hindering accurate and clear evaluation and analysis.

Additionally, in analyzing the results, if statistical tests had been conducted, it would have been possible to identify statistically significant differences and to construct more convincing arguments and analyses.

5.3.2 Limitations in Scope of the Study

This study focused on a limited scope related to biases in language models. By mentioning these limitations, we aim to highlight potential avenues for future expanded research.

Firstly, the texts dealt with in this study were mainly focused on written language. However, as observed in the evaluations of Korean corpora in the study, the characteristics of written and spoken language are quite different. Considering that language models are used in various environments where both written and spoken language are employed, it would be necessary to investigate language biases in diverse text styles and genres.

Also, this study only covered pre-trained language models. Even if there are linguistic biases in pre-trained base models, the impact of linguistic biases in them can vary greatly depending on what data and methods are used for fine-tuning or instruction tuning on those models. From this perspective, exploring strategies to mitigate bias in language models will also be an important research topic.

While this study focused on linguistic biases, investigating biases in language models trained on English-centric data also requires an approach from a cultural perspective. Future research is needed to examine the cultural contexts that underlie the responses of language models, how these biases influence users, and how such biases can be mitigated.

Lastly, this study has limitations in dealing with a single language, Korean, as a case study for linguistic biases in language models. To determine the extent to which the findings in this study can be generalized to other languages, additional research expanding to other languages is necessary. Examining linguistic biases across multiple languages can provide valuable insights into the universal and language-specific aspects of bias in language models.

We hope that the limitations of this study will suggest potential directions for future research. We also leave the topics discovered during the research but not addressed due to the scope of the study as potential subjects for future research.

Chapter 6

Conclusion

This study was initiated by questioning whether there could be potential issues or concerns when using language models trained on English-centric data for tasks in non-English languages. We aimed to explore linguistic biases in language models by selecting the Korean text generation task as a case study. To achieve this, various types of language models, including Korean monolingual models, multilingual models, Korean continually pre-training models, and non-Korean-trained models, were examined to see what texts they generate in response to Korean input prompts and whether there were differences in the Korean sentences they produced, from a linguistic perspective. We evaluated the texts generated by the language models and the Korean sentences filtered from the generated texts from various dimensions, such as surface-level, lexical, syntactic, semantic, and English translationese aspects. We also analyzed the results in terms of the influences of the model size, training data, and input prompts on text generation. The main findings obtained from this study can be summarized as follows:

Regarding the generated texts themselves, all models trained on Korean (the Korean monolingual models, multilingual models, and Korean continually pre-trained models) generated texts in the correct language — they followed Korean input prompts with Korean completions. However, the Korean monolingual models and Korean continually pre-trained models generated more complete sentences at the surface level and more task-appropriate results at the semantic level compared to the multilingual models. This suggests that models specifically trained on Korean data have an advantage in generating linguistically appropriate Korean texts. The non-Korean-trained models struggled to generate Korean in the zero-shot setting but demonstrated rapid learning ability for Korean in the few-shot setting.

Concerning the differences in the Korean sentences generated by the language models, no significant differences were observed between the language models in terms of sentence length or syntactic aspects. Instead, there was a tendency for the within-group variance of the language models to increase in the few-shot setting. In terms of lexical diversity and sentiment classification, the multilingual model generated slightly less diverse vocabulary and more sentences classified with neutral sentiment compared to the Korean monolingual models or continually pre-trained models. This suggests that the Korean monolingual models or Korean continually pre-trained models have a wider range of vocabulary choices and generate sentences that can be more clearly interpreted in terms

of sentiment. The translationese from English did not seem to be very apparent at the surface-level, such as in vocabulary and word order.

One of the factors influencing text generation was the input prompt. The few-shot prompt provided in the study, which included three examples, contributed to the rapid adaptation and learning of Korean by the Korean non-trained model. For the Korean-trained models, it affected text generation in various directions and degrees depending on the individual model.

The size of the language model was also a crucial factor affecting the models' text generation results. The competent performance demonstrated by the Korean continually pre-trained models and the rapid Korean learning ability shown by the Korean non-trained model in the few-shot setting might be attributed to their relatively large sizes. Among the models sharing the same architecture, including the multilingual models, performance generally improved with increasing model size.

The pre-training data of the language models also influenced their text generation results. The composition of the pre-training Korean data (i.e., the kinds of data it consisted of) seems to have affected the style of the generated texts by the language models. Moreover, considering the performance between the Korean monolingual models and the Korean continually pre-trained models, as well as the performance between the multilingual models and the non-Korean-trained models, the total amount of pre-training data of the language models also seems to impact performance. However, the influence of the size or ratio of the Korean data in the training data appears to interact with other factors such as the model size, architecture, and total pre-training data size. More research is needed to disentangle these interrelated factors.

The study provides insights into the relationship between the size and training data of language models and their linguistic abilities in various aspects. The results showed that a small-sized Korean monolingual model performs well in the formal aspects of Korean, and a large-sized non-Korean-trained model quickly learns Korean forms even with few-shot prompts. This suggests that learning the formal aspects of a language is possible with a small-sized model or limited training data. However, the results also showed that the small-sized Korean monolingual model struggled to understand the relationship between the headline and the main text in the input prompt in the few-shot setting, resulting in lower performance in the semantic aspect. In contrast, the large-sized non-Korean-trained model understood what was required in the input prompt and showed relatively high performance in the semantic aspect. This implies that learning the semantic aspects of a language, or the reasoning abilities based on semantic understanding, is possible with large-sized models and extensive pre-training data.

Revisiting the initial question that motivated this study—whether there could be potential issues or concerns when using language models trained on English-centric data for tasks in non-English languages—, a review of the results from the Korean continually pre-trained models suggests that even language models trained on English-centric data can generate good-quality sentences with linguistic features of Korean at both the surface and semantic levels if they are further trained with a considerable amount of Korean data. If the base model is learned from a large amount of high-quality data, it is speculated that the language knowledge learned from them can also be indirectly utilized in Korean

generation, especially in the semantic aspect. Therefore, for Korean tasks, employing a base model trained on a large amount of high-quality data and then further training it on Korean could be a strategy worth attempting without serious concerns about linguistic bias. This would be a practical consideration when applying language models trained on English-centric data to non-English language tasks.

This study aimed to shed light on the issue of linguistic bias, which can be a potential problem in the rapidly growing applications of large language models. Especially focusing on Korean as a case study, a non-English language that differs significantly from English in linguistic aspects, this study sought to broaden the scope of linguistic bias research. The methods used in this study to explore linguistic biases in language models can be extended and applied to other languages. The empirical results presented in this study can also be compared and analyzed with the results of applications in other languages. Analyses from various languages would further deepen our understanding of linguistic biases in language models, and contribute to the development of more unbiased and inclusive language models.

Appendix A

Appendix

당신은 한국어 전문지식을 가지고 있고 한국어를 모국어로 하는 언어학자로서, 주어진 한국어 텍스트의 품질을 평가할 것입니다. 첫 열 개 어절은 입력으로 주어진 프롬프트이며, 이어지는 부분은 LLM에 의해 생성된 텍스트입니다. LLM이 생성한 파트에 중점을 두어 평가하십시오. 텍스트를 주의깊게 읽고 다음의 ‘평가기준’에 따라 충분한 시간을 들여 면밀히 검토하십시오.

‘평가기준’

문법 정확성: 텍스트가 문법적으로 정확한지 평가합니다.

유창성: 텍스트가 자연스럽고 유창한지 평가합니다.

의미 명확성: 텍스트의 의미가 명확하고 이해하기 쉬운지 평가합니다.

적절성: 텍스트가 입력 텍스트로 주어진 맥락에 적절한지 평가합니다.

창의성: 텍스트에 창의성과 독창성이 있는지 평가합니다.

응집성: 텍스트 내 연결이 자연스럽고 일관된 흐름을 유지하는지 평가합니다.

종합적 평가: 전체 텍스트의 품질을 종합적으로 평가합니다.

이전 케이스들과 동일하고 일관된 척도로 평가하도록 유의하고, 답변 전 모든 평가가 타당한지 다시 한번 검토하십시오. 답변은 ‘평가기준’의 각 항목 점수를 1(낮음)에서 5(높음) 사이로 부여해, 오직 다음 형식으로 출력하십시오:

<답변 예>

문법 정확성: 5

유창성: 4

의미 명확성: 4

적절성: 4

창의성: 3

응집성: 3

종합적 평가: 4

Figure A.1: Input prompt provided to LLM as a judge in hyperparameter search.

Model	Zero-shot				Few-shot			
	Avg	Std	Min	Max	Avg	Std	Min	Max
<i>Korean Monolingual Models</i>								
kogpt2-base-v2-125M	644.9	41.2	384	832	664.5	37.8	521	833
ko-gpt-trinity-1.2B-v0.5	441.1	229.8	14	830	412.1	224.4	17	804
polyglot-ko-1.3b	416.1	177.8	12	608	420.8	149.6	62	609
polyglot-ko-3.8b	396.9	183.7	17	610	358.8	158.9	24	778
polyglot-ko-5.8b	409.1	179.9	18	639	293.7	148.1	32	602
polyglot-ko-12.8b	431.7	164.3	14	607	397.8	148.8	45	627
kogpt-6B	510.2	154.3	26	803	292.2	168.9	21	698
<i>Multilingual Models</i>								
xglm-564M	505.4	55.5	38	816	509.2	31.2	366	672
xglm-1.7B	506.6	42.6	361	828	513.3	34.2	391	905
xglm-4.5B	325.6	176.6	23	760	469.2	80.8	62	713
xglm-7.5B	272.8	173.2	21	620	475.9	78.6	55	604
mGPT-1.3B	331.6	119.0	14	536	209.4	86.8	18	590
mGPT-13B	359.9	108.1	35	695	282.1	95.3	16	466
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	466.2	164.0	15	681	195.2	82.0	14	609
llama-2-ko-7b	537.7	60.6	282	1042	544.2	59.9	279	987
open-llama-2-ko-7b	260.7	209.5	10	751	213.0	104.8	14	645
llama-2-koen-13b	419.3	182.0	23	709	200.7	74.2	28	642
OPEN-SOLAR-KO-10.7B	308.8	215.5	11	671	147.6	40.9	37	484
SOLAR-KOEN-10.8B	400.5	174.1	14	668	206.1	101.0	12	626
<i>Non-Korean-trained Models</i>								
Yi-6B	509.5	475.4	11	3086	274.5	153.9	40	2217
Llama-2-7b-hf	625.9	701.3	13	2911	317.4	244.1	21	2081
Llama-2-13b-hf	775.0	699.9	48	2923	271.0	157.2	14	1525
SOLAR-10.7B-v1.0	1170.6	734.9	12	3089	365.8	231.7	14	1911

Table A.1: Statistics on the length of texts generated by language models.

Model	Zero-shot				Few-shot			
	Avg	Std	Min	Max	Avg	Std	Min	Max
<i>Korean Monolingual Models</i>								
kogpt2-base-v2	261.0	2.7	255	275	261.4	2.8	256	275
ko-gpt-trinity-1.2B-v0.5	168.5	85.1	6	272	156.8	81.7	11	272
polyglot-ko-1.3b	205.5	84.2	7	278	201.0	70.6	38	279
polyglot-ko-3.8b	197.0	88.2	7	279	171.0	75.4	13	274
polyglot-ko-5.8b	204.5	86.6	7	279	143.6	72.9	16	275
polyglot-ko-12.8b	215.1	78.7	9	278	198.8	72.2	23	279
kogpt	228.9	66.3	13	278	125.9	72.0	10	271
<i>Multilingual Models</i>								
xglm-564M	262.6	18.5	22	280	265.4	3.2	259	281
xglm-1.7B	264.5	3.3	258	280	265.4	3.2	258	280
xglm-4.5B	174.6	89.8	12	279	250.4	41.5	30	280
xglm-7.5B	146.3	87.4	14	279	252.8	39.1	27	281
mGPT	209.3	70.9	10	280	141.9	70.2	11	288
mGPT-13B	227.4	64.1	21	282	178.1	62.5	11	283
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	220.2	72.5	11	283	89.8	38.8	11	266
llama-2-ko-7b	262.4	5.9	157	283	261.6	5.9	176	281
open-llama-2-ko-7b	118.0	86.2	6	277	96.7	46.6	11	274
llama-2-koen-13b	192.8	78.0	15	283	88.3	32.3	18	270
OPEN-SOLAR-KO-10.7B	143.2	95.3	7	281	66.6	18.1	17	196
SOLAR-KOEN-10.8B	204.6	81.3	10	282	132.0	60.5	10	271
<i>Non-Korean-trained Models</i>								
Yi-6B	356.0	194.6	19	623	452.4	117.0	55	571
Llama-2-7b-hf	289.8	174.5	18	556	370.9	141.5	27	557
Llama-2-13b-hf	367.7	162.5	34	558	364.2	132.2	19	556
SOLAR-10.7B-v1.0	420.3	143.6	16	548	445.3	144.7	19	548

Table A.2: Statistics on the number of tokens generated by language models.

Model	Total	HANGUL	LATIN	DIGIT	CJK	Symbol	Whitespace	Undecodable
<i>Korean Monolingual Models</i>								
kogpt2-base-v2	644873	441638	8321	21136	157	28920	144689	0
ko-gpt-trinity-1.2B-v0.5	441053	301439	6040	12149	241	19088	102088	0
polyglot-ko-1.3b	416102	281187	10987	10240	304	22198	91050	8
polyglot-ko-3.8b	396876	269345	9771	9501	403	20965	86812	11
polyglot-ko-5.8b	409077	276694	9624	10648	356	22896	88769	4
polyglot-ko-12.8b	431676	291195	11458	11290	300	24097	93192	8
kogpt	510168	344019	11043	12909	513	25959	115484	3
<i>Multilingual Models</i>								
xglm-564M	505381	301016	32276	28699	129	43367	98823	0
xglm-1.7B	506556	307810	28375	26371	403	40603	101596	0
xglm-4.5B	325602	212563	10783	11998	219	20990	68728	0
xglm-7.5B	272752	185019	5134	8262	68	14505	59663	0
mGPT	331550	227695	5783	7908	189	18459	71394	24
mGPT-13B	359860	252425	4183	7199	135	15117	80746	28
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	466169	324267	7352	9005	212	21240	104077	10
llama-2-ko-7b	537698	359442	18934	11420	705	31351	114704	429
open-llama-2-ko-7b	260659	169782	15734	3160	3498	11720	56045	1
llama-2-koen-13b	419342	289447	10695	5084	195	23429	90381	9
OPEN-SOLAR-KO-10.7B	308816	206007	13038	4819	2772	12333	69236	0
SOLAR-KOEN-10.8B	400490	290604	6564	8198	171	20866	74011	2
<i>Non-Korean-trained Models</i>								
Yi-6B	509516	82682	238555	22022	1132	69605	92373	307
Llama-2-7b-hf	625940	94826	393100	8278	1180	26153	100826	169
Llama-2-13b-hf	775026	120160	490070	9223	418	23994	130973	125
SOLAR-10.7B-v1.0	1170566	35596	791355	36761	16112	91972	175765	19719

Table A.3: Character type distribution in texts generated by language models in the zero-shot task.

Model	Total	HANGUL	LATIN	DIGIT	CJK	Symbol	Whitespace	Undecodable
<i>Korean Monolingual Models</i>								
kogpt2-base-v2	664539	455577	6161	19178	250	32465	150900	0
ko-gpt-trinity-1.2B-v0.5	412148	278680	6975	11173	221	19771	95309	0
polyglot-ko-1.3b	420803	283473	9454	10020	600	23335	93832	2
polyglot-ko-3.8b	358839	245721	5631	7850	449	17941	81182	3
polyglot-ko-5.8b	293732	200099	5477	6783	342	15461	65480	32
polyglot-ko-12.8b	397772	261499	10836	10896	1416	26135	86000	47
kogpt	292196	188014	9449	9803	217	18500	65990	1
<i>Multilingual Models</i>								
xglm-564M	509155	324587	20934	22414	153	37627	102694	0
xglm-1.7B	513346	330624	20429	19236	394	35918	106199	0
xglm-4.5B	469243	314244	10058	14866	211	27384	102100	0
xglm-7.5B	475935	327830	6756	11817	382	22205	106875	0
mGPT	209382	143638	3063	8957	69	10451	41620	1353
mGPT-13B	282140	198727	2613	6078	200	11352	62968	190
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	195189	138370	1930	3000	99	7976	43813	0
llama-2-ko-7b	544250	363592	17960	10048	999	33984	116578	597
open-llama-2-ko-7b	212986	141352	13140	1513	109	14752	42005	0
llama-2-koen-13b	200676	142621	3317	1396	61	8291	44981	0
OPEN-SOLAR-KO-10.7B	147603	105561	1589	1526	40	5425	33462	0
SOLAR-KOEN-10.8B	206126	177636	2569	4557	179	11425	9745	0
<i>Non-Korean-trained Models</i>								
Yi-6B	274528	153015	30536	10342	1486	21010	55787	1080
Llama-2-7b-hf	317377	160752	59029	7857	1326	24076	62992	654
Llama-2-13b-hf	270985	164331	24629	8217	440	13769	59368	210
SOLAR-10.7B-v1.0	365755	54966	72822	17711	8182	45335	40667	112651

Table A.4: Character type distribution in texts generated by language models in the few-shot task.

Model	Total	HANGUL	LATIN	DIGIT	CJK	Mixed	Symbol	Whitespace	Undecodable
<i>Korean Monolingual Models</i>									
ko-gpt-trinity-1.2B-v0.5	168507	140034	3193	3737	239	17007	3936	353	0
kogpt	228876	195703	4774	7248	435	74	18436	264	1795
kogpt2-base-v2	261016	214906	5708	7035	136	25087	7834	298	0
polyglot-ko-1.3b	205494	169656	6174	5965	149	42	19199	175	4041
polyglot-ko-12.8b	215122	176011	6599	6525	163	30	20543	159	4997
polyglot-ko-3.8b	197012	163033	5654	5615	175	45	18414	157	3866
polyglot-ko-5.8b	204547	168166	5510	6164	180	38	19934	170	4322
<i>Multilingual Models</i>									
xglm-564M	261570	208735	10931	9640	121	5219	25171	894	0
xglm-1.7B	263516	214357	10161	8618	330	4943	23241	915	0
xglm-4.5B	173554	147715	4000	4459	195	1730	14243	938	0
xglm-7.5B	145316	127040	1998	3231	68	1091	10887	907	0
mGPT	209336	169625	2659	4582	157	906	14644	1179	15502
mGPT-13B	227443	188231	1946	4101	124	847	14163	1649	16362
<i>Korean Continually Pre-trained Models</i>									
Yi-Ko-6B	219210	179590	3615	9005	175	1	20441	5071	1308
llama-2-ko-7b	261376	205402	7685	11453	456	2	27477	4945	3328
open-llama-2-ko-7b	117005	92854	4748	3162	2978	0	9319	1521	1706
llama-2-koen-13b	191757	160083	5020	5102	143	22	18469	1624	1230
OPEN-SOLAR-KO-10.7B	142250	113026	4126	4819	2493	0	11394	4760	1023
SOLAR-KOEN-10.8B	203559	168476	3404	8198	132	4	18128	3268	1918
<i>Non-Korean-trained Models</i>									
Yi-6B	355985	36865	67936	22022	785	174	54985	31894	139197
Llama-2-7b-hf	288786	74824	93707	8278	770	3	22580	25950	61946
Llama-2-13b-hf	366727	90207	115510	9222	257	0	21930	38855	90693
SOLAR-10.7B-v1.0	419260	33231	189899	36760	12741	28	74395	21823	48456

Table A.5: Token type distribution in texts generated by language models in the zero-shot task.

Model	Total	HANGUL	LATIN	DIGIT	CJK	Mixed	Symbol	Whitespace	Undecodable
<i>Korean Monolingual Models</i>									
ko-gpt-trinity-1.2B-v0.5	156809	127996	3678	3181	217	16556	4979	186	0
kogpt	125880	102877	4095	5470	202	27	12048	134	873
kogpt2-base-v2	261365	214663	3962	5831	202	28219	8415	66	0
polyglot-ko-1.3b	200975	166593	5303	5942	445	17	20858	100	1647
polyglot-ko-12.8b	198830	151532	6161	6374	715	35	22256	232	11448
polyglot-ko-3.8b	170988	145083	3149	4702	291	35	16655	48	992
polyglot-ko-5.8b	143571	117760	3017	4050	216	43	14459	91	3911
<i>Multilingual Models</i>									
xglm-564M	264373	219451	7776	9265	153	3349	23524	226	0
xglm-1.7B	264400	222579	7944	8049	377	2579	22238	214	0
xglm-4.5B	249351	216866	3720	7145	206	1408	19345	381	0
xglm-7.5B	251764	223190	2850	6262	380	641	17857	517	0
mGPT	141854	99544	1726	5025	40	748	7939	167	26572
mGPT-13B	178075	144260	1324	3599	160	888	10693	508	16636
<i>Korean Continually Pre-trained Models</i>									
Yi-Ko-6B	88780	75152	1000	3000	91	0	7854	1490	192
llama-2-ko-7b	260608	204395	7295	10073	713	4	29314	4662	3757
open-llama-2-ko-7b	95737	77752	5846	1513	66	2	9831	271	345
llama-2-koen-13b	87299	76391	1640	1397	45	2	7165	506	146
OPEN-SOLAR-KO-10.7B	65623	57349	869	1526	30	0	5207	594	48
SOLAR-KOEN-10.8B	130997	113041	1437	4557	140	0	10349	1060	402
<i>Non-Korean-trained Models</i>									
Yi-6B	452442	48164	10218	10342	1034	3	20006	44974	316775
Llama-2-7b-hf	369939	106302	17997	7857	990	4	21227	48546	166467
Llama-2-13b-hf	363182	106766	7790	8217	282	0	13143	53477	173487
SOLAR-10.7B-v1.0	444257	21274	23753	17619	3901	5	28084	28815	319555

Table A.6: Token type distribution in texts generated by language models in the few-shot task.

Model	Zero-shot				Few-shot			
	Length		Words		Length		Words	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
<i>Korean Monolingual Models</i>								
kogpt2-base-v2	60.5	32.0	13.6	7.3	55.7	32.1	12.7	7.2
ko-gpt-trinity-1.2B-v0.5	60.0	36.5	13.7	8.3	52.6	34.3	12.0	7.8
polyglot-ko-1.3b	70.5	47.9	15.8	10.8	69.6	45.3	15.8	10.3
polyglot-ko-3.8b	67.5	45.0	15.1	10.2	63.1	42.1	14.6	9.9
polyglot-ko-5.8b	66.1	45.8	14.7	10.2	62.4	43.1	14.1	9.9
polyglot-ko-12.8b	69.5	48.3	15.4	10.7	67.6	47.8	14.8	10.7
kogpt	67.1	56.2	15.4	13.3	64.0	52.4	14.8	12.3
<i>Multilingual Models</i>								
xglm-564M	67.6	51.3	13.5	10.4	69.6	56.3	14.5	11.9
xglm-1.7B	67.2	51.3	13.8	10.5	67.1	57.6	14.3	12.4
xglm-4.5B	62.5	44.6	13.5	9.6	67.9	51.8	15.1	11.7
xglm-7.5B	57.1	35.1	12.7	7.9	67.0	43.9	15.3	10.1
mGPT	63.5	46.4	14.0	10.4	64.7	47.9	13.2	9.8
mGPT-13B	59.7	39.1	13.6	9.0	61.6	40.8	14.0	9.4
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	56.9	31.9	12.9	7.3	57.6	30.8	13.2	7.0
llama-2-ko-7b	59.5	41.4	13.0	8.7	59.1	42.3	13.0	9.0
open-llama-2-ko-7b	70.7	51.9	15.5	10.9	72.3	54.3	14.8	10.8
llama-2-koen-13b	60.4	41.1	13.4	9.1	64.1	36.8	14.7	8.4
OPEN-SOLAR-KO-10.7B	59.8	32.9	13.5	7.4	59.6	30.0	13.8	6.8
SOLAR-KOEN-10.8B	53.1	36.4	10.3	8.2	48.7	30.3	3.1	4.9
<i>Non-Korean-trained Models</i>								
Yi-6B	64.6	117.9	8.6	16.6	62.4	66.8	12.9	11.9
Llama-2-7b-hf	132.9	209.5	21.5	31.7	75.0	83.7	15.2	14.8
Llama-2-13b-hf	138.7	205.2	23.5	31.1	64.5	57.2	14.3	11.0
SOLAR-10.7B-v1.0	113.6	188.3	16.9	31.0	186.7	196.1	21.2	29.6

Table A.7: Average text length and average number of words per sentence generated by language models.

Model	Zero-shot			Few-shot		
	Total Sent.	Pattern Sent.	Ratio	Total Sent.	Pattern Sent.	Ratio
<i>Korean Monolingual Models</i>						
kogpt2-base-v2	10681	197	0.018	11956	261	0.022
ko-gpt-trinity-1.2B-v0.5	7320	227	0.031	7830	699	0.089
polyglot-ko-1.3b	5925	967	0.163	6074	893	0.147
polyglot-ko-3.8b	5909	941	0.159	5721	582	0.102
polyglot-ko-5.8b	6221	883	0.142	4731	511	0.108
polyglot-ko-12.8b	6242	1087	0.174	5911	1618	0.274
kogpt	7623	1994	0.262	4599	1672	0.364
<i>Multilingual Models</i>						
xglm-564M	7516	4190	0.557	7381	3343	0.453
xglm-1.7B	7575	4174	0.551	7711	3039	0.394
xglm-4.5B	5236	1917	0.366	6961	2050	0.294
xglm-7.5B	4802	1019	0.212	7129	896	0.126
mGPT	5252	1564	0.298	3257	1068	0.328
mGPT-13B	6047	603	0.100	4597	456	0.099
<i>Korean Continually Pre-trained Models</i>						
Yi-Ko-6B	8208	624	0.076	3407	102	0.030
llama-2-ko-7b	9089	1318	0.145	9260	1294	0.140
open-llama-2-ko-7b	3707	381	0.103	2975	691	0.232
llama-2-koen-13b	6993	1398	0.200	3150	212	0.067
OPEN-SOLAR-KO-10.7B	5174	231	0.045	2495	20	0.008
SOLAR-KOEN-10.8B	7609	944	0.124	4302	168	0.039
<i>Non-Korean-trained Models</i>						
Yi-6B	7846	4158	0.530	4423	883	0.200
Llama-2-7b-hf	4716	1235	0.262	4252	1154	0.271
Llama-2-13b-hf	5595	950	0.170	4220	332	0.079
SOLAR-10.7B-v1.0	10265	4158	0.405	1963	1471	0.749

Table A.8: Total sentences, sentences with unusual patterns, and their ratio in texts generated by language models.

Model	Zero-shot			Few-shot		
	Total Sent.	Pattern Sent.	Ratio	Total Sent.	Pattern Sent.	Ratio
<i>Korean Monolingual Models</i>						
kogpt2-base-v2	10681	13	0.001	11956	163	0.014
ko-gpt-trinity-1.2B-v0.5	7320	10	0.001	7830	0	0.000
polyglot-ko-1.3b	5925	0	0.000	6074	127	0.021
polyglot-ko-12.8b	6242	0	0.000	5911	127	0.021
polyglot-ko-3.8b	5909	0	0.000	5721	254	0.044
polyglot-ko-5.8b	6221	0	0.000	4731	257	0.054
kogpt	7623	192	0.025	4599	165	0.036
<i>Multilingual Models</i>						
xglm-564M	7516	451	0.060	7381	534	0.072
xglm-1.7B	7575	383	0.051	7711	489	0.063
xglm-4.5B	5236	188	0.036	6961	599	0.086
xglm-7.5B	4802	102	0.021	7129	712	0.100
mGPT	5252	41	0.008	3257	0	0.000
mGPT-13B	6047	43	0.007	4597	128	0.028
<i>Korean Continually Pre-trained Models</i>						
Yi-Ko-6B	8208	66	0.008	3407	64	0.019
llama-2-ko-7b	9089	339	0.037	9260	680	0.073
open-llama-2-ko-7b	3707	61	0.016	2975	607	0.204
llama-2-koen-13b	6993	208	0.030	3150	103	0.033
OPEN-SOLAR-KO-10.7B	5174	24	0.005	2495	2	0.001
SOLAR-KOEN-10.8B	7609	65	0.009	4302	287	0.067
<i>Non-Korean-trained Models</i>						
Yi-6B	7846	1683	0.215	4423	434	0.098
Llama-2-7b-hf	4716	93	0.020	4252	172	0.040
Llama-2-13b-hf	5595	70	0.013	4220	217	0.051
SOLAR-10.7B-v1.0	10265	218	0.021	1963	90	0.046

Table A.9: Total sentences, sentences with the headline pattern “<...>”, and their ratio in texts generated by language models.

Model	Zero-shot			Few-shot		
	Total Sent.	Pattern Sent.	Ratio	Total Sent.	Pattern Sent.	Ratio
<i>Korean Monolingual Models</i>						
kogpt2-base-v2	10681	13	0.001	11956	221	0.018
ko-gpt-trinity-1.2B-v0.5	7320	23	0.003	7830	59	0.008
polyglot-ko-1.3b	5925	12	0.002	6074	471	0.078
polyglot-ko-12.8b	6242	28	0.004	5911	448	0.076
polyglot-ko-3.8b	5909	14	0.002	5721	708	0.124
polyglot-ko-5.8b	6221	17	0.003	4731	636	0.134
kogpt	7623	274	0.036	4599	563	0.122
<i>Multilingual Models</i>						
xglm-564M	7516	667	0.089	7381	1747	0.237
xglm-1.7B	7575	565	0.075	7711	1981	0.257
xglm-4.5B	5236	254	0.049	6961	2104	0.302
xglm-7.5B	4802	115	0.024	7129	2087	0.293
mGPT	5252	62	0.012	3257	0	0.000
mGPT-13B	6047	52	0.009	4597	274	0.060
<i>Korean Continually Pre-trained Models</i>						
Yi-Ko-6B	8208	83	0.010	3407	147	0.043
llama-2-ko-7b	9089	371	0.041	9260	1453	0.157
open-llama-2-ko-7b	3707	81	0.022	2975	884	0.297
llama-2-koen-13b	6993	259	0.037	3150	168	0.053
OPEN-SOLAR-KO-10.7B	5174	26	0.005	2495	2	0.001
SOLAR-KOEN-10.8B	7609	72	0.009	4302	512	0.119
<i>Non-Korean-trained Models</i>						
Yi-6B	7846	2135	0.272	4423	865	0.196
Llama-2-7b-hf	4716	118	0.025	4252	503	0.118
Llama-2-13b-hf	5595	84	0.015	4220	417	0.099
SOLAR-10.7B-v1.0	10265	334	0.033	1963	168	0.086

Table A.10: Total sentences, sentences with the headline pattern “<” or “>”, and their ratio in texts generated by language models.

Model	Zero-shot				Few-shot			
	Total Sent.	Complete Sent.	Comp. Ratio	Incomp. Ratio	Total Sent.	Complete Sent.	Comp. Ratio	Incomp. Ratio
<i>Korean Monolingual Models</i>								
kogpt2-base-v2	10681	9605	0.899	0.101	11956	10783	0.902	0.098
ko-gpt-trinity-1.2B-v0.5	7320	6675	0.912	0.088	7830	6576	0.840	0.160
polyglot-ko-1.3b	5925	4906	0.828	0.172	6074	5107	0.841	0.159
polyglot-ko-3.8b	5909	4913	0.831	0.169	5721	4990	0.872	0.128
polyglot-ko-5.8b	6221	5221	0.839	0.161	4731	3950	0.835	0.165
polyglot-ko-12.8b	6242	5139	0.823	0.177	5911	4668	0.790	0.210
kogpt	7623	6275	0.823	0.177	4599	3873	0.842	0.158
<i>Multilingual Models</i>								
xglm-564M	7516	5270	0.701	0.299	7381	4950	0.671	0.329
xglm-1.7B	7575	5425	0.716	0.284	7711	5431	0.704	0.296
xglm-4.5B	5236	4049	0.773	0.227	6961	4874	0.700	0.300
xglm-7.5B	4802	3982	0.829	0.171	7129	5258	0.738	0.262
mGPT	5252	3748	0.714	0.286	3257	2592	0.796	0.204
mGPT-13B	6047	4871	0.806	0.194	4597	3890	0.846	0.154
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	8208	7294	0.889	0.111	3407	3267	0.959	0.041
llama-2-ko-7b	9089	7673	0.844	0.156	9260	7282	0.786	0.214
open-llama-2-ko-7b	3707	3204	0.864	0.136	2975	2035	0.684	0.316
llama-2-koen-13b	6993	5658	0.809	0.191	3150	2893	0.918	0.082
OPEN-SOLAR-KO-10.7B	5174	4677	0.904	0.096	2495	2474	0.992	0.008
SOLAR-KOEN-10.8B	7609	6618	0.870	0.130	4302	3722	0.865	0.135
<i>Non-Korean-trained Models</i>								
Yi-6B	7846	2202	0.281	0.719	4423	3055	0.691	0.309
Llama-2-7b-hf	4716	3041	0.645	0.355	4252	3170	0.746	0.254
Llama-2-13b-hf	5595	3990	0.713	0.287	4220	3547	0.841	0.159
SOLAR-10.7B-v1.0	10265	4724	0.460	0.540	1963	690	0.352	0.648

Table A.11: Total sentences, sentences with formally complete endings, and their ratio in texts generated by language models.

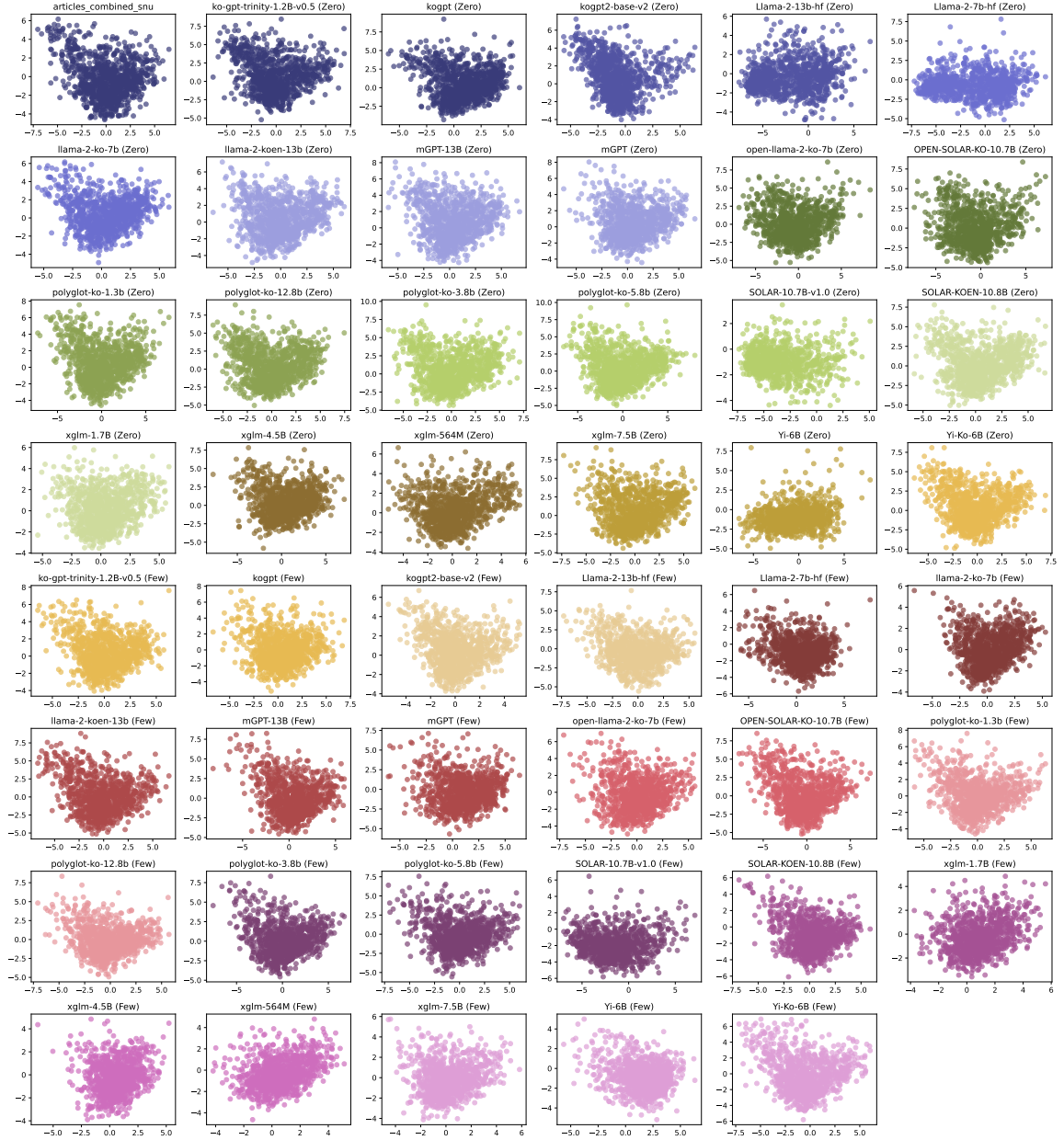


Figure A.2: Visualization using PCA of semantic embedding distribution for the original texts and the texts generated by language models.

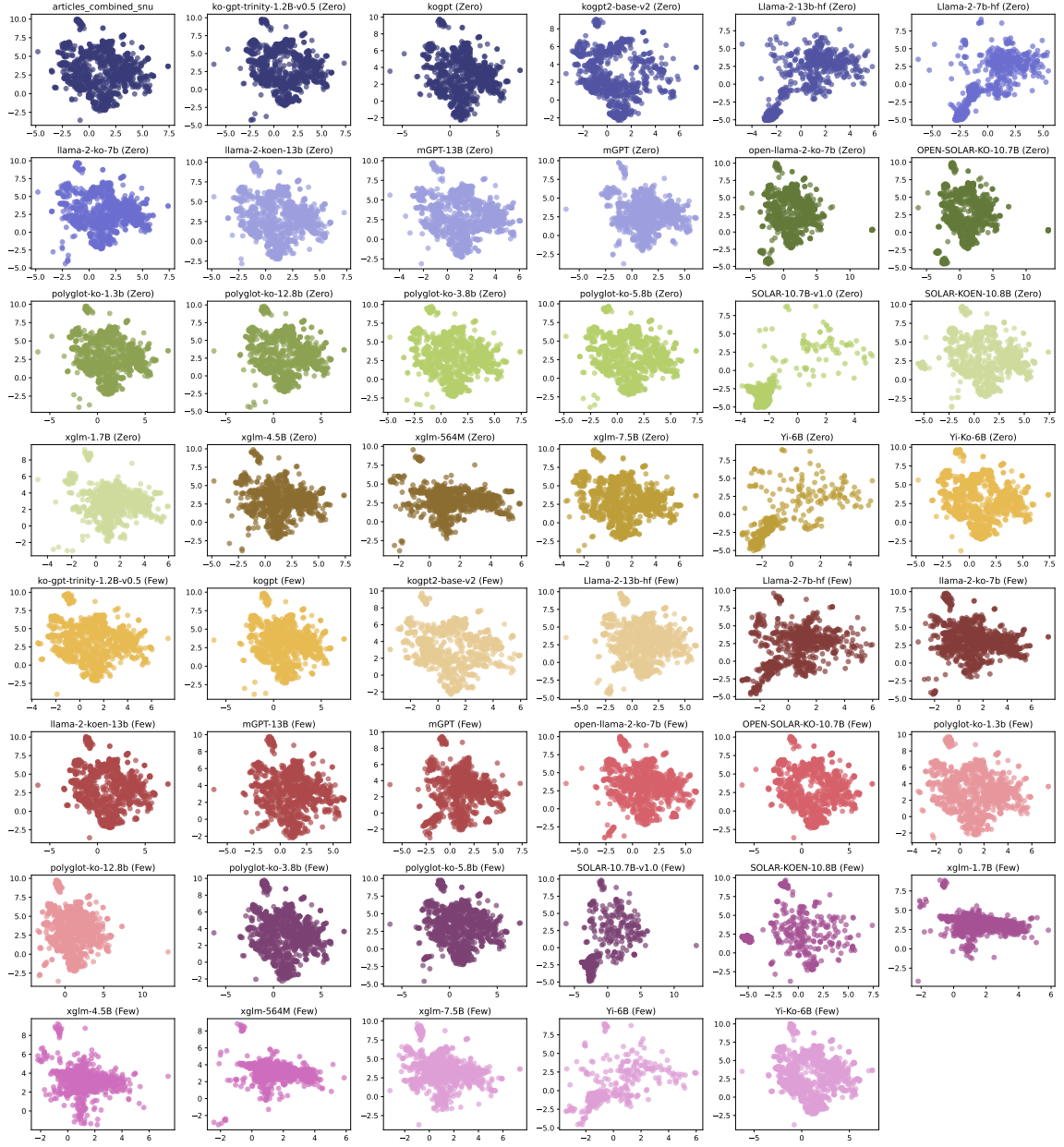


Figure A.3: Visualization using UMAP of semantic embedding distribution for the original texts and the texts generated by language models.

Model	Zero-shot Sentence count	Few-shot Sentence count
<i>Korean Corpora</i>		
corpus spoken	10000	10000
corpus written	10000	10000
<i>Korean Monolingual Models</i>		
kogpt2-base-v2	9396	10361
ko-gpt-trinity-1.2B-v0.5	6481	6315
polyglot-ko-1.3b	4270	4297
polyglot-ko-12.8b	4396	3487
polyglot-ko-3.8b	4334	4200
polyglot-ko-5.8b	4651	3299
kogpt	4627	2238
<i>Multilingual Models</i>		
xglm-1.7B	2529	2553
xglm-4.5B	2812	2650
xglm-564M	2477	2181
xglm-7.5B	3358	3726
mGPT-13B	4566	3620
mGPT	2964	2003
<i>Korean Continually Pre-trained Models</i>		
Yi-Ko-6B	6910	3142
llama-2-ko-7b	6754	5779
open-llama-2-ko-7b	2765	1414
llama-2-koen-13b	4821	2727
OPEN-SOLAR-KO-10.7B	4279	2437
SOLAR-KOEN-10.8B	5911	3487
<i>Non-Korean-trained Models</i>		
Yi-6B	1249	2497
Llama-2-7b-hf	1354	2289
Llama-2-13b-hf	2134	3213
SOLAR-10.7B-v1.0	412	72

Table A.12: Number of Korean sentences in the Korean corpora and Korean sentences generated by language models.

	Zero-shot				Few-shot			
	Length		Words		Length		Words	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
<i>Korean Corpora</i>								
corpus spoken	53.8	56.5	14.1	15.4	53.8	56.5	14.1	15.4
corpus written	60.8	31.3	14.0	7.1	60.8	31.3	14.0	7.1
<i>Korean Monolingual Models</i>								
kogpt2-base-v2	61.6	31.0	14.2	7.1	56.3	30.7	13.1	6.9
ko-gpt-trinity-1.2B-v0.5	62.2	35.2	14.4	8.0	57.5	32.3	13.5	7.3
polyglot-ko-1.3b	70.1	45.2	16.2	10.4	68.7	39.4	16.3	9.3
polyglot-ko-3.8b	66.3	40.9	15.4	9.5	63.4	37.3	15.1	8.9
polyglot-ko-5.8b	64.9	41.9	15.0	9.5	61.8	36.8	14.5	8.6
polyglot-ko-12.8b	67.3	42.9	15.5	9.9	63.0	37.6	14.8	8.8
kogpt	62.7	49.1	14.9	11.7	54.6	40.1	13.5	10.1
<i>Multilingual Models</i>								
xglm-564M	49.6	35.3	11.4	8.0	42.6	34.7	10.0	8.1
xglm-1.7B	51.5	35.5	11.9	8.2	46.1	38.5	10.9	9.0
xglm-4.5B	57.6	36.5	13.4	8.5	55.1	39.1	13.0	9.2
xglm-7.5B	57.4	31.5	13.4	7.2	61.6	35.8	14.5	8.4
mGPT	62.0	41.3	14.3	9.5	61.8	37.9	14.2	8.7
mGPT-13B	62.5	36.4	14.5	8.4	65.6	37.5	15.3	8.7
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	57.1	30.2	13.3	6.9	58.4	29.7	13.6	6.7
llama-2-ko-7b	56.9	33.2	13.2	7.6	52.4	30.2	12.2	7.0
open-llama-2-ko-7b	65.3	36.8	15.3	8.5	57.1	39.1	13.3	8.9
llama-2-koen-13b	56.8	34.9	13.3	8.0	61.9	34.1	14.6	7.9
OPEN-SOLAR-KO-10.7B	58.7	28.3	13.8	6.6	58.2	29.7	13.8	6.7
SOLAR-KOEN-10.8B	56.1	35.2	14.2	8.9	63.1	35.5	15.8	8.9
<i>Non-Korean-trained Models</i>								
Yi-6B	39.3	29.0	8.3	5.9	60.7	41.8	13.6	9.2
Llama-2-7b-hf	56.5	49.3	10.8	9.1	67.4	49.2	14.9	10.7
Llama-2-13b-hf	63.2	44.3	13.6	9.3	64.8	38.3	15.1	8.8
SOLAR-10.7B-v1.0	21.3	17.4	3.3	2.4	21.3	16.1	4.7	3.4

Table A.13: Average text length and average number of words per Korean sentences in the Korean corpora and Korean sentences generated by language models.

Model	Zero-shot								Few-shot							
	Tokens	Types	TTR	MSTTR	MATTR	RTTR	CTTR	MTLD	Tokens	Types	TTR	MSTTR	MATTR	RTTR	CTTR	MTLD
<i>Corpora</i>																
corpus spoken	271665	10617	0.039	0.531	0.531	20.37	14.40	56.52	271665	10617	0.039	0.531	0.531	20.37	14.40	56.52
corpus written	281127	17867	0.064	0.567	0.567	33.70	23.83	74.16	281127	17867	0.064	0.567	0.567	33.70	23.83	74.16
<i>Korean Monolingual Models</i>																
kogpt2-base-v2-125M	276548	12025	0.043	0.591	0.592	22.87	16.17	78.61	269749	12455	0.046	0.594	0.595	23.98	16.96	80.47
ko-gpt-trinity-1.2B-v0.5	172581	11688	0.068	0.629	0.629	28.13	19.89	96.94	191937	12809	0.067	0.617	0.617	29.24	20.67	91.59
polyglot-ko-1.3b	136333	10635	0.078	0.680	0.681	28.80	20.37	140.66	140614	10453	0.074	0.658	0.658	27.88	19.71	123.84
polyglot-ko-3.8b	122548	10458	0.085	0.684	0.684	29.87	21.12	142.03	135583	10550	0.078	0.655	0.656	28.65	20.26	122.27
polyglot-ko-5.8b	93263	9119	0.098	0.679	0.679	29.86	21.11	135.62	141198	10882	0.077	0.644	0.645	28.96	20.48	111.81
polyglot-ko-12.8b	99148	9511	0.096	0.686	0.686	30.21	21.36	143.43	137911	10738	0.078	0.654	0.653	28.92	20.45	119.90
kogpt-6B	55300	7590	0.137	0.741	0.741	32.28	22.82	228.54	138013	11059	0.080	0.705	0.704	29.77	21.05	170.44
<i>Multilingual Models</i>																
xglm-564M	44097	5372	0.122	0.705	0.704	25.58	18.09	169.17	58966	6636	0.113	0.668	0.668	27.33	19.32	128.79
xglm-1.7B	56629	6423	0.113	0.700	0.699	26.99	19.09	159.47	62040	7254	0.117	0.674	0.673	29.12	20.59	133.92
xglm-4.5B	68667	7236	0.105	0.698	0.697	27.61	19.53	159.21	78820	7900	0.100	0.663	0.662	28.14	19.90	124.35
xglm-7.5B	107825	8481	0.079	0.677	0.676	25.83	18.26	137.19	93046	8060	0.087	0.648	0.647	26.42	18.68	112.77
mGPT-1.3B	59662	6798	0.114	0.690	0.689	27.83	19.68	151.77	88546	8007	0.090	0.668	0.667	26.91	19.03	131.31
mGPT-13B	114273	8929	0.078	0.663	0.663	26.41	18.68	124.17	137791	9691	0.070	0.644	0.644	26.11	18.46	111.44
<i>Korean Continually Pre-trained Models</i>																
Yi-Ko-6B	87126	8337	0.096	0.630	0.630	28.24	19.97	96.04	188584	12422	0.066	0.603	0.603	28.60	20.23	85.57
llama-2-ko-7b	145194	11797	0.081	0.634	0.634	30.96	21.89	98.58	184421	13385	0.073	0.608	0.608	31.17	22.04	86.63
open-llama-2-ko-7b	38719	5841	0.151	0.665	0.664	29.68	20.99	123.50	86586	8556	0.099	0.636	0.636	29.08	20.56	101.48
llama-2-koen-13b	80925	8226	0.102	0.651	0.651	28.92	20.45	112.34	131929	10897	0.083	0.633	0.633	30.00	21.21	100.48
OPEN-SOLAR-KO-10.7B	67338	7776	0.115	0.653	0.653	29.97	21.19	111.96	119267	10041	0.084	0.629	0.629	29.07	20.56	97.95
SOLAR-KOEN-10.8B	103501	9523	0.092	0.640	0.640	29.60	20.93	105.23	156089	11553	0.074	0.598	0.598	29.24	20.68	82.43
<i>Non-Korean-trained Models</i>																
Yi-6B	70418	9967	0.142	0.652	0.652	37.56	26.56	109.23	22359	4342	0.194	0.618	0.618	29.04	20.53	89.16
Llama-2-7b	68258	9005	0.132	0.661	0.660	34.47	24.37	116.53	33533	5957	0.178	0.676	0.676	32.53	23.00	137.64
Llama-2-13b	93735	9080	0.097	0.608	0.608	29.66	20.97	83.69	61967	7206	0.116	0.634	0.634	28.95	20.47	103.30
SOLAR-10.7B-v1.0	635	343	0.540	0.679	0.669	13.61	9.62	85.53	3727	1102	0.296	0.597	0.591	18.05	12.76	73.14

Table A.14: Lexical diversity metrics of Korean corpora and Korean sentences generated by language models.

Model	Zero-shot								Few-shot							
	Tokens	Types	TTR	MSTTR	MATTR	RTTR	CTTR	MTLD	Tokens	Types	TTR	MSTTR	MATTR	RTTR	CTTR	MTLD
<i>Corpora</i>																
corpus spoken	268552	9529	0.035	0.526	0.525	18.39	13.00	55.23	268552	9529	0.035	0.526	0.525	18.39	13.00	55.23
corpus written	276714	15459	0.056	0.561	0.561	29.39	20.78	71.40	276714	15459	0.056	0.561	0.561	29.39	20.78	71.40
<i>Korean Monolingual Models</i>																
kogpt2-base-v2-125M	273772	10707	0.039	0.587	0.587	20.46	14.47	76.82	266294	10886	0.041	0.590	0.589	21.10	14.92	78.16
ko-gpt-trinity-1.2B-v0.5	170656	10539	0.062	0.624	0.624	25.51	18.04	94.29	189677	11406	0.060	0.611	0.611	26.19	18.52	88.43
polyglot-ko-1.3b	134492	9617	0.072	0.676	0.676	26.22	18.54	135.19	138819	9467	0.068	0.654	0.653	25.41	17.97	120.73
polyglot-ko-3.8b	120950	9517	0.079	0.679	0.679	27.37	19.35	136.93	133934	9581	0.072	0.650	0.651	26.18	18.51	117.98
polyglot-ko-5.8b	91900	8342	0.091	0.674	0.673	27.52	19.46	130.20	139285	9810	0.070	0.640	0.640	26.29	18.59	108.73
polyglot-ko-12.8b	97540	8617	0.088	0.680	0.681	27.59	19.51	137.67	136024	9684	0.071	0.649	0.649	26.26	18.57	116.78
kogpt-6B	54453	7038	0.129	0.735	0.735	30.16	21.33	217.32	136212	10021	0.074	0.699	0.699	27.15	19.20	164.08
<i>Multilingual Models</i>																
xglm-564M	43374	4865	0.112	0.696	0.697	23.36	16.52	160.64	58121	6021	0.104	0.661	0.661	24.97	17.66	123.42
xglm-1.7B	55804	5832	0.105	0.695	0.693	24.69	17.46	154.86	61143	6584	0.108	0.667	0.667	26.63	18.83	128.36
xglm-4.5B	67748	6610	0.098	0.692	0.691	25.40	17.96	152.46	77848	7216	0.093	0.657	0.656	25.86	18.29	119.39
xglm-7.5B	106226	7619	0.072	0.671	0.669	23.38	16.53	129.84	92009	7396	0.080	0.642	0.642	24.38	17.24	108.74
mGPT-1.3B	58663	6063	0.103	0.683	0.682	25.03	17.70	144.28	87367	7272	0.083	0.661	0.661	24.60	17.40	126.35
mGPT-13B	112697	7975	0.071	0.657	0.657	23.76	16.80	119.48	136133	8784	0.065	0.638	0.638	23.81	16.83	107.66
<i>Korean Continually Pre-trained Models</i>																
Yi-Ko-6B	85805	7615	0.089	0.625	0.624	26.00	18.38	91.72	186093	11093	0.060	0.597	0.598	25.71	18.18	83.36
llama-2-ko-7b	143121	10599	0.074	0.627	0.628	28.02	19.81	94.40	182037	11997	0.066	0.603	0.602	28.12	19.88	84.18
open-llama-2-ko-7b	38115	5407	0.142	0.660	0.658	27.70	19.58	117.77	85378	7749	0.091	0.630	0.630	26.52	18.75	98.03
llama-2-koen-13b	79762	7519	0.094	0.645	0.645	26.62	18.83	107.80	130219	9816	0.075	0.628	0.627	27.20	19.23	96.73
OPEN-SOLAR-KO-10.7B	66329	7144	0.108	0.647	0.646	27.74	19.61	106.65	117598	9030	0.077	0.623	0.623	26.33	18.62	94.80
SOLAR-KOEN-10.8B	102222	8769	0.086	0.635	0.635	27.43	19.39	101.50	153978	10455	0.068	0.593	0.593	26.64	18.84	79.42
<i>Non-Korean-trained Models</i>																
Yi-6B	67947	7791	0.115	0.637	0.637	29.89	21.13	98.51	21620	3668	0.170	0.604	0.601	24.95	17.64	81.08
Llama-2-7b	66236	7439	0.112	0.647	0.647	28.90	20.44	106.13	32753	5264	0.161	0.665	0.665	29.09	20.57	128.62
Llama-2-13b	91832	7909	0.086	0.599	0.600	26.10	18.45	80.06	61092	6530	0.107	0.627	0.627	26.42	18.68	99.35
SOLAR-10.7B-v1.0	594	302	0.508	0.628	0.638	12.39	8.76	65.39	3625	1015	0.280	0.594	0.576	16.86	11.92	66.47

Table A.15: Lexical diversity metrics of Korean corpora and Korean sentences generated by language models, considering only tokens included in the Korean dictionary entry morpheme set.

Model	ADV	ADJ	NOUN	VERB	PUNCT	ADP	DET	NUM	PRON	CCONJ	AUX	PROPN	SYM	INTJ	PART	X
<i>Corpora</i>																
corpus spoken	18.24	4.61	30.74	26.04	9.42	1.09	4.16	1.20	4.24	0.17	0.00	0.08	0.01	0.00	0.00	0.00
corpus written	11.61	2.40	42.31	21.52	14.45	3.22	0.58	1.34	0.78	0.20	0.32	0.88	0.36	0.00	0.02	0.00
ko gsd-ud-TB	14.02	3.37	40.42	22.85	12.94	2.27	0.69	1.08	0.85	0.30	0.15	0.63	0.35	0.01	0.04	0.02
<i>Korean Monolingual Models</i>																
kogpt2-base-v2-125M	13.61	2.19	41.01	22.03	13.55	2.95	0.59	1.20	1.44	0.23	0.43	0.56	0.20	0.00	0.02	0.00
ko-gpt-trinity-1.2B-v0.5	13.75	2.62	39.45	22.87	14.00	2.67	0.74	1.05	1.31	0.27	0.36	0.70	0.17	0.00	0.02	0.00
polyglot-ko-1.3b	13.92	2.58	40.17	23.25	12.67	2.90	0.86	1.03	1.29	0.37	0.25	0.52	0.18	0.00	0.01	0.00
polyglot-ko-3.8b	13.91	2.67	40.66	23.29	12.49	2.81	0.77	0.91	1.24	0.33	0.24	0.49	0.20	0.00	0.01	0.00
polyglot-ko-5.8b	13.34	2.49	40.72	22.74	13.36	2.88	0.81	0.99	1.26	0.42	0.25	0.51	0.20	0.00	0.01	0.00
polyglot-ko-12.8b	13.33	2.43	41.36	22.36	13.17	2.96	0.78	1.00	1.16	0.48	0.28	0.48	0.19	0.00	0.01	0.00
kogpt-6B	15.24	3.04	40.70	23.69	9.78	2.77	0.94	1.15	1.30	0.62	0.17	0.46	0.13	0.00	0.01	0.00
<i>Multilingual Models</i>																
xglm-564M	14.23	3.83	38.30	24.19	12.53	2.31	0.78	0.99	1.69	0.32	0.24	0.52	0.07	0.00	0.01	0.00
xglm-1.7B	14.46	3.46	39.31	24.30	12.01	2.15	0.68	0.97	1.56	0.33	0.24	0.44	0.07	0.00	0.01	0.00
xglm-4.5B	14.86	3.51	38.37	25.45	11.30	2.31	0.78	0.71	1.67	0.37	0.21	0.39	0.06	0.00	0.02	0.00
xglm-7.5B	15.58	3.56	37.93	24.47	11.57	2.38	0.72	0.84	1.76	0.36	0.25	0.48	0.07	0.00	0.03	0.00
mGPT-1.3B	14.69	3.22	39.25	24.77	11.93	2.43	0.47	0.83	1.15	0.38	0.33	0.44	0.10	0.00	0.01	0.00
mGPT-13B	15.30	3.53	38.43	24.63	11.76	2.39	0.63	0.74	1.40	0.37	0.30	0.40	0.10	0.00	0.01	0.00
<i>Korean Continually Pre-trained Models</i>																
Yi-Ko-6B	13.82	2.57	40.99	23.49	12.51	2.43	0.70	0.95	1.32	0.26	0.31	0.52	0.13	0.00	0.01	0.00
llama-2-ko-7b	13.82	2.89	39.97	23.83	12.59	2.39	0.84	1.05	1.41	0.26	0.22	0.54	0.18	0.00	0.01	0.00
open-llama-2-ko-7b	13.25	2.30	42.99	23.85	11.20	2.42	0.73	0.88	1.20	0.29	0.22	0.59	0.08	0.00	0.02	0.00
llama-2-koen-13b	14.22	2.88	40.78	24.05	11.78	2.13	0.74	0.89	1.49	0.27	0.22	0.47	0.10	0.00	0.01	0.00
OPEN-SOLAR-KO-10.7B	13.46	2.31	44.20	22.48	11.50	2.45	0.60	0.90	1.04	0.31	0.18	0.47	0.09	0.00	0.02	0.00
SOLAR-KOEN-10.8B	12.75	2.67	43.62	22.55	11.27	2.17	0.83	1.83	1.37	0.26	0.17	0.39	0.13	0.00	0.01	0.00
<i>Non-Korean-trained Models</i>																
Yi-6B	9.27	3.12	35.44	19.58	23.19	3.84	0.73	2.19	1.17	0.16	0.21	0.80	0.24	0.00	0.04	0.00
Llama-2-7b	10.89	2.40	38.11	18.42	21.92	3.41	0.55	0.82	0.93	0.39	0.29	1.24	0.61	0.00	0.04	0.00
Llama-2-13b	11.98	2.94	39.54	22.46	15.83	2.82	0.92	0.81	1.07	0.28	0.30	0.74	0.29	0.00	0.02	0.00
SOLAR-10.7B-v1.0	5.39	1.98	33.91	9.09	39.33	3.63	0.04	2.13	1.43	0.40	0.11	1.80	0.73	0.04	0.00	0.00

Table A.16: UPOS distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.

Model	ADV	ADJ	NOUN	VERB	PUNCT	ADP	DET	NUM	PRON	CCONJ	AUX	PROPN	SYM	INTJ	PART	X
<i>Corpora</i>																
corpus spoken	18.24	4.61	30.74	26.04	9.42	1.09	4.16	1.20	4.24	0.17	0.00	0.08	0.01	0.00	0.00	0.00
corpus written	11.61	2.40	42.31	21.52	14.45	3.22	0.58	1.34	0.78	0.20	0.32	0.88	0.36	0.00	0.02	0.00
ko gsd-ud-TB	14.02	3.37	40.42	22.85	12.94	2.27	0.69	1.08	0.85	0.30	0.15	0.63	0.35	0.01	0.04	0.02
<i>Korean Monolingual Models</i>																
kogpt2-base-v2-125M	14.02	2.68	38.93	22.31	14.31	3.09	0.58	1.06	1.69	0.19	0.49	0.50	0.14	0.00	0.01	0.00
ko-gpt-trinity-1.2B-v0.5	13.46	2.77	39.30	23.10	14.30	2.53	0.64	1.10	1.52	0.24	0.20	0.70	0.14	0.00	0.01	0.00
polyglot-ko-1.3b	12.15	2.33	44.70	22.56	11.86	2.61	0.74	1.01	0.98	0.31	0.24	0.38	0.13	0.00	0.01	0.00
polyglot-ko-3.8b	12.22	2.34	44.97	22.76	11.40	2.49	0.76	0.99	1.10	0.33	0.21	0.33	0.11	0.00	0.01	0.00
polyglot-ko-5.8b	11.67	2.21	45.32	22.10	12.34	2.61	0.70	1.03	0.98	0.31	0.22	0.39	0.12	0.00	0.01	0.00
polyglot-ko-12.8b	11.07	2.07	46.22	21.61	12.57	2.58	0.65	1.03	0.95	0.40	0.21	0.51	0.13	0.00	0.00	0.00
kogpt-6B	10.49	2.29	50.01	20.94	9.38	3.11	0.60	1.28	0.78	0.55	0.15	0.34	0.09	0.00	0.00	0.00
<i>Multilingual Models</i>																
xglm-564M	14.36	4.31	39.85	24.12	10.69	2.03	0.49	1.17	2.00	0.34	0.24	0.30	0.07	0.00	0.02	0.00
xglm-1.7B	15.07	4.20	40.27	25.09	9.17	1.86	0.48	0.94	1.99	0.42	0.18	0.25	0.05	0.00	0.03	0.00
xglm-4.5B	13.40	3.27	43.45	23.72	10.02	2.25	0.45	1.13	1.27	0.36	0.22	0.34	0.09	0.00	0.04	0.00
xglm-7.5B	13.61	2.91	43.87	23.54	9.68	2.35	0.48	0.98	1.41	0.35	0.29	0.38	0.07	0.00	0.05	0.00
mGPT-1.3B	14.42	3.40	42.70	23.55	10.69	2.13	0.23	1.12	0.69	0.30	0.36	0.30	0.08	0.00	0.03	0.00
mGPT-13B	14.78	3.22	42.12	23.95	10.55	2.20	0.19	0.89	0.95	0.38	0.38	0.28	0.09	0.00	0.03	0.00
<i>Korean Continually Pre-trained Models</i>																
Yi-Ko-6B	13.01	2.30	43.63	23.58	11.47	2.28	0.63	0.88	1.06	0.24	0.36	0.45	0.07	0.00	0.03	0.00
llama-2-ko-7b	13.63	2.79	41.88	24.05	11.89	2.10	0.55	0.92	1.22	0.23	0.20	0.42	0.10	0.00	0.02	0.00
open-llama-2-ko-7b	13.00	2.84	41.39	24.14	12.65	1.92	0.48	0.87	1.48	0.27	0.40	0.49	0.04	0.00	0.01	0.00
llama-2-koen-13b	13.73	2.56	43.05	24.75	10.20	2.00	0.55	0.89	1.21	0.30	0.32	0.38	0.04	0.00	0.03	0.00
OPEN-SOLAR-KO-10.7B	13.28	2.48	43.96	23.28	11.37	2.09	0.60	0.82	1.01	0.25	0.29	0.46	0.06	0.00	0.03	0.00
SOLAR-KOEN-10.8B	11.67	2.40	46.96	22.55	10.66	2.09	0.52	1.66	0.74	0.18	0.18	0.31	0.09	0.00	0.01	0.00
<i>Non-Korean-trained Models</i>																
Yi-6B	10.53	2.85	42.23	20.46	16.57	3.07	0.61	1.52	0.75	0.33	0.17	0.57	0.33	0.00	0.00	0.00
Llama-2-7b	10.02	2.90	43.74	19.22	17.55	3.12	0.30	0.90	0.46	0.26	0.35	0.97	0.21	0.00	0.01	0.00
Llama-2-13b	9.99	2.48	46.74	20.77	13.62	3.02	0.44	1.04	0.51	0.22	0.39	0.61	0.14	0.01	0.01	0.00
SOLAR-10.7B-v1.0	7.31	2.30	42.80	15.03	26.10	1.88	0.00	2.51	0.84	0.21	0.00	0.42	0.63	0.00	0.00	0.00

Table A.17: UPOS distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.

Tag	Description
ADJ	Adjective
ADP	Adposition
ADV	Adverb
AUX	Auxiliary
CCONJ	Coordinating conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper noun
PUNCT	Punctuation
SCONJ	Subordinating conjunction
SYM	Symbol
VERB	Verb
X	Other

Table A.18: *Universal Part-of-Speech tags and descriptions.*

Model	NNG	SF	SS	NNG+JKO	VV+EC	NNG+JKB	MAG	NNP	VV+ETM	NNG+JX	NNG+JKS	NNG+XSV+EC	NNG+XSV+ETM	NNG+NNG	SP	MM	SN+NNB	NNB	NNG+JKG	VA+ETM
Average	12.04	6.27	5.27	4.02	4.01	3.93	3.79	3.18	3.02	2.48	2.47	2.18	2.04	2.02	1.96	1.79	1.45	1.43	1.39	0.95
<i>Corpora</i>																				
corpus spoken	10.19	6.07	0.14	4.52	6.25	4.14	7.24	1.20	3.51	2.41	3.24	1.53	1.45	0.63	0.03	6.55	0.06	1.72	1.34	1.01
corpus written	12.22	5.87	6.23	4.68	3.97	4.08	2.70	4.07	3.33	2.84	2.73	1.77	2.17	1.73	1.83	1.41	1.84	1.58	1.57	0.83
ko gsd-ud-TB	11.84	5.53	4.34	3.78	4.54	4.40	3.98	3.43	2.99	2.92	2.71	1.98	1.70	1.35	2.56	1.47	1.66	1.45	1.72	1.01
<i>Korean Monolingual Models</i>																				
kogpt2-base-v2-125M	11.68	5.93	5.29	4.57	3.84	3.82	3.56	3.58	3.28	2.75	2.38	2.21	2.49	2.15	1.73	1.49	1.66	1.61	1.45	0.76
ko-gpt-trinity-1.2B-v0.5	11.33	5.80	5.81	4.81	4.03	4.02	3.80	3.14	3.33	2.45	2.76	2.27	2.45	1.87	2.07	1.66	1.53	1.38	1.68	0.88
polyglot-ko-1.3b	13.52	5.16	5.00	3.28	4.11	4.07	4.13	3.30	2.60	2.64	2.22	2.47	1.87	2.38	2.11	1.81	1.40	1.66	1.14	0.95
polyglot-ko-3.8b	13.40	5.44	4.63	3.43	4.05	4.17	4.04	3.44	2.81	2.69	2.33	2.29	1.89	2.30	2.05	1.68	1.46	1.81	1.12	0.97
polyglot-ko-5.8b	13.30	5.64	4.88	3.31	3.93	4.01	3.92	3.32	2.62	2.79	2.29	2.35	1.80	2.56	2.37	1.64	1.52	1.70	1.18	0.90
polyglot-ko-12.8b	13.94	5.31	5.04	3.10	3.83	4.06	3.87	3.54	2.59	2.65	2.14	2.36	1.75	2.75	2.29	1.55	1.42	1.82	1.04	0.87
kogpt-6B	16.63	4.19	4.05	1.56	4.22	3.45	5.61	2.74	2.26	2.44	1.72	2.34	1.51	3.00	1.22	2.12	1.26	1.78	1.42	0.94
<i>Multilingual Models</i>																				
xglm-564M	11.16	6.12	4.56	3.62	4.23	3.80	4.33	3.20	3.41	2.21	2.46	2.30	2.10	2.42	1.45	1.89	1.36	1.50	1.24	1.26
xglm-1.7B	12.47	5.78	4.40	3.79	4.28	3.91	4.43	3.27	3.56	2.15	2.53	2.29	2.20	2.35	1.36	1.66	1.26	1.46	1.21	1.22
xglm-4.5B	11.76	6.02	3.73	4.01	4.21	4.26	4.51	2.93	3.55	2.21	2.50	2.52	2.36	2.10	1.32	1.91	1.28	1.48	1.23	1.12
xglm-7.5B	11.09	6.43	3.54	4.00	4.25	4.19	4.42	2.78	3.38	2.40	2.43	2.43	2.37	2.00	1.34	1.84	1.28	1.62	1.33	1.19
mGPT-1.3B	12.07	5.71	4.26	3.94	3.98	3.81	3.92	3.28	3.37	2.41	2.06	2.43	2.35	1.97	1.43	1.42	1.32	1.60	1.40	1.12
mGPT-13B	10.91	6.01	3.79	4.37	4.05	4.16	4.19	2.94	3.34	2.50	2.12	2.47	2.50	1.82	1.64	1.57	1.15	1.66	1.51	1.23
<i>Korean Continually Pre-trained Models</i>																				
Yi-Ko-6B	11.88	6.37	4.31	4.84	4.05	4.15	3.69	3.45	3.35	2.88	2.87	2.29	2.42	2.00	1.50	1.75	1.19	1.50	1.38	0.91
llama-2-ko-7b	11.36	6.31	4.24	4.65	4.39	4.26	3.77	3.10	3.43	2.79	3.05	2.19	2.26	1.95	1.61	1.87	1.20	1.42	1.60	0.98
open-llama-2-ko-7b	13.54	4.71	4.90	4.80	3.93	4.42	3.24	3.91	3.54	2.65	2.95	2.34	2.56	2.33	1.29	1.81	1.02	1.42	1.47	0.92
llama-2-koen-13b	12.57	6.16	4.26	4.61	4.27	4.17	4.03	3.45	3.57	2.69	2.97	2.12	2.28	2.12	1.06	1.94	0.82	1.43	1.37	1.03
OPEN-SOLAR-KO-10.7B	14.73	6.23	3.88	4.80	3.61	4.49	3.38	4.07	3.31	2.88	2.90	2.14	2.53	2.23	1.11	1.70	1.21	1.40	1.38	0.85
SOLAR-KOEN-10.8B	17.84	5.79	3.43	4.96	4.51	4.36	3.68	3.38	3.08	3.08	3.09	2.09	1.96	0.44	1.70	1.93	1.23	1.62	1.61	0.99
<i>Non-Korean-trained Models</i>																				
Yi-6B	9.27	8.32	9.87	4.46	3.78	3.50	1.81	2.14	2.76	1.77	2.66	1.76	1.53	1.70	3.55	1.43	1.33	0.92	1.77	0.85
Llama-2-7b	7.77	6.60	10.42	4.08	3.06	3.15	2.23	2.70	2.26	1.83	1.83	2.58	1.66	2.45	4.07	0.79	1.46	0.47	1.57	0.62
Llama-2-13b	8.86	5.96	6.07	5.31	4.25	4.12	2.67	2.97	2.76	1.90	2.03	3.05	2.40	1.75	3.38	1.40	1.73	1.10	1.81	0.82
SOLAR-10.7B-v1.0	7.66	15.47	15.91	1.25	0.66	1.10	1.32	3.48	0.59	1.58	1.32	0.22	0.37	2.13	4.91	0.15	5.06	0.11	0.66	0.37

Table A.19: XPOS distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.

Model	NNG	SF	SS	MAG	VV+EC	NNG+JKO	NNG+JKB	NNP	VV+ETM	NNG+JKS	NNG+NNG	NNG+JX	NNG+XSV+EC	MM	NNG+XSV+ETM	SN+NNB	NNB	SP	NNG+JKG	VA+ETM
Average	15.46	6.33	4.27	4.03	4.02	3.88	3.74	3.46	3.17	2.92	2.26	2.17	2.02	1.74	1.67	1.55	1.34	1.30	1.25	1.04
<i>Corpora</i>																				
corpus spoken	10.19	6.07	0.14	7.24	6.25	4.52	4.14	1.20	3.51	3.24	0.63	2.41	1.53	6.55	1.45	0.06	1.72	0.03	1.34	1.01
corpus written	12.22	5.87	6.23	2.70	3.97	4.68	4.08	4.07	3.33	2.73	1.73	2.84	1.77	1.41	2.17	1.84	1.58	1.83	1.57	0.83
ko gsd-ud-TB	11.84	5.53	4.34	3.98	4.54	3.78	4.40	3.43	2.99	2.71	1.35	2.92	1.98	1.47	1.70	1.66	1.45	2.56	1.72	1.01
<i>Korean Monolingual Models</i>																				
kogpt2-base-v2-125M	10.67	6.45	5.07	4.22	4.10	4.57	3.89	3.17	3.25	2.64	1.68	2.96	1.85	1.63	2.28	1.82	1.31	2.27	1.55	0.92
ko-gpt-trinity-1.2B-v0.5	12.08	6.22	5.32	4.12	3.96	4.65	4.09	2.83	3.48	3.25	1.63	2.60	2.23	1.70	1.82	1.81	0.88	2.53	1.88	0.98
polyglot-ko-1.3b	17.91	5.35	4.25	4.12	3.82	3.43	3.60	3.61	2.82	2.93	2.59	2.65	2.07	1.87	1.61	1.45	1.38	1.91	1.12	1.06
polyglot-ko-3.8b	18.06	5.69	3.78	4.24	4.08	3.73	3.64	3.52	2.90	2.98	2.61	2.73	2.09	1.86	1.72	1.56	1.52	1.63	1.06	1.00
polyglot-ko-5.8b	17.44	5.95	4.14	3.51	3.88	3.64	3.80	4.01	2.89	2.92	2.62	2.76	2.01	1.70	1.58	1.74	1.49	1.93	1.16	0.90
polyglot-ko-12.8b	18.89	5.73	4.63	3.50	3.90	3.32	3.47	4.19	2.57	2.79	2.92	2.65	2.11	1.59	1.40	1.61	1.40	1.81	0.94	0.91
kogpt-6B	27.31	5.25	3.42	4.79	3.48	0.64	1.73	3.48	1.94	1.32	3.68	2.19	1.70	1.76	0.93	1.29	2.08	0.45	1.13	0.83
<i>Multilingual Models</i>																				
xglm-564M	16.76	6.68	3.01	5.68	4.40	2.71	2.66	3.01	3.21	2.06	2.86	1.20	2.11	1.91	0.94	1.20	1.03	0.37	0.67	1.41
xglm-1.7B	17.09	5.59	2.80	5.84	4.90	2.96	3.06	2.86	3.14	2.17	2.73	1.22	2.47	1.91	1.10	0.96	1.20	0.30	0.71	1.43
xglm-4.5B	17.81	6.07	3.19	4.57	4.08	3.48	3.43	3.34	3.40	2.60	2.70	1.34	2.42	1.75	1.32	1.47	1.20	0.40	1.00	1.24
xglm-7.5B	16.57	5.93	3.15	4.31	3.98	3.89	3.39	3.67	3.43	3.01	2.61	1.31	2.38	1.71	1.48	1.73	1.27	0.36	1.03	1.11
mGPT-1.3B	16.20	6.31	3.62	4.49	3.51	3.46	3.50	3.20	3.21	2.16	2.54	1.42	2.13	1.25	1.46	1.27	1.55	0.40	1.19	1.31
mGPT-13B	13.82	5.89	3.19	4.85	3.97	4.46	3.71	3.05	3.33	2.94	2.21	1.73	2.28	1.19	1.76	1.32	1.48	1.22	1.64	1.21
<i>Korean Continually Pre-trained Models</i>																				
Yi-Ko-6B	13.63	6.41	4.14	3.16	3.69	4.69	4.65	4.15	3.71	3.66	2.42	2.44	2.31	1.70	2.00	1.06	1.26	0.73	1.35	0.85
llama-2-ko-7b	13.06	6.69	3.77	3.72	4.56	4.53	4.41	3.75	3.51	3.71	2.12	2.46	2.12	1.59	1.76	1.01	1.23	1.09	1.38	0.98
open-llama-2-ko-7b	14.90	5.98	5.93	3.61	3.94	4.02	4.80	3.36	3.94	3.40	2.60	2.17	2.26	1.81	1.47	0.38	1.13	0.54	1.17	1.00
llama-2-koen-13b	14.87	5.86	3.83	3.60	4.22	4.54	5.10	3.90	3.87	3.34	2.33	2.34	2.28	1.86	1.87	0.42	1.21	0.32	1.23	0.95
OPEN-SOLAR-KO-10.7B	15.52	6.37	4.58	3.39	3.83	4.85	4.99	4.05	3.61	3.65	2.51	2.18	2.11	1.74	1.82	0.70	1.26	0.29	1.10	0.95
SOLAR-KOEN-10.8B	20.19	5.48	3.36	3.16	4.18	5.16	4.33	3.98	3.39	3.89	0.48	2.42	2.15	1.47	2.05	1.12	1.70	1.56	1.73	0.94
<i>Non-Korean-trained Models</i>																				
Yi-6B	14.03	6.04	6.30	2.67	3.57	4.01	4.08	2.70	3.17	2.85	2.75	1.96	1.55	1.27	2.21	1.96	1.23	3.00	1.59	1.09
Llama-2-7b	13.80	5.74	8.41	2.46	3.55	4.14	3.87	4.42	2.99	2.59	2.38	1.58	1.84	0.81	2.08	1.98	1.23	2.52	1.38	1.30
Llama-2-13b	15.26	5.77	5.33	2.57	3.92	4.95	4.16	4.82	3.10	2.78	2.09	1.79	1.91	1.24	2.50	2.81	1.47	1.89	1.33	1.10
SOLAR-10.7B-v1.0	11.69	15.66	5.22	4.18	2.30	2.09	0.21	2.09	1.67	3.55	2.09	2.09	0.84	0.42	1.04	6.05	0.63	1.88	0.63	0.63

Table A.20: XPOS distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.

Major Category	Subcategory	Fine-grained Category
(1) Nominal	Noun (NN)	General Noun (NNG)
		Proper Noun (NNP)
		Dependent Noun (NNB)
	Pronoun (NP)	Pronoun (NP)
	Numeral (NR)	Numeral (NR)
(2) Predicate	Verb (VV)	Verb (VV)
	Adjective (VA)	Adjective (VA)
	Auxiliary Verb (VX)	Auxiliary Verb (VX)
	Copula (VC)	Affirmative Copula (VCP)
		Negative Copula (VCN)
(3) Modifier	Determiner (MM)	Descriptive Determiner (MMA)
		Demonstrative Determiner (MMD)
		Numeral Determiner (MMN)
	Adverb (MA)	General Adverb (MAG)
		Conjunctive Adverb (MAJ)
(4) Interjection	Interjection (IC)	Interjection (IC)
(5) Particle	Case Particle (JK)	Nominative Case Particle (JKS)
		Complement Case Particle (JKC)
		Genitive Case Particle (JKG)
		Accusative Case Particle (JKO)
		Adverbial Case Particle (JKB)
		Vocative Case Particle (JKV)
		Quotative Case Particle (JKQ)
		Auxiliary Particle (JX)
	Conjunctive Particle (JC)	Auxiliary Particle (JX)
		Conjunctive Particle (JC)
(6) Dependent Morpheme	Ending (EM)	Pre-final Ending (EP)
		Final Ending (EF)
		Connective Ending (EC)
		Nominal ending (ETN)
		Adnominal Ending (ETM)
	Prefix (XP)	Nominal Prefix (XPN)
		Noun-Deriving Suffix (XSN)
		Verb-Deriving Suffix (XSV)
	Suffix (XS)	Adjective-Deriving Suffix (XSA)
		Root (XR)
(7) Symbol	General Symbol (ST)	Period, Question Mark, Exclamation Mark (SF)
		Comma, Middle Dot, Colon, Slash (SP)
		Quotation Marks, Parentheses, Dash (SS)
		Ellipsis (SE)
		Tilde (SO)
		Other Symbols (SW)
	Foreign Language (SL)	Foreign Language (SL)
	Hanja (Chinese Characters) (SH)	Hanja (Chinese Characters) (SH)
	Number (SN)	Number (SN)
	Uncategorizable Category (NA)	Uncategorizable Category (NA)

Table A.21: Korean Part-of-Speech tags and descriptions.

Model	NNG	EC	VV	ETM	JKB	JX	NNB	XSV	SF	JKO	EF	SS	NNP	JKS	SN	EP	MAG	VX	XSN	JKG
Average	27.88	6.40	5.71	4.91	4.48	3.59	3.59	3.56	3.42	3.24	3.05	2.89	2.87	2.42	2.21	2.08	2.06	1.82	1.65	1.61
<i>Corpora</i>																				
corpus spoken	23.34	9.03	7.91	5.24	4.10	3.35	2.98	2.29	3.34	3.24	2.96	0.07	1.12	3.46	0.11	1.98	4.01	2.31	2.08	1.25
corpus written	28.35	5.41	5.84	4.96	4.42	3.68	4.05	3.38	3.20	3.66	3.19	3.40	3.54	2.54	2.83	2.34	1.48	1.54	1.43	1.74
ko gsd-ud-TB	26.47	7.36	6.12	4.59	4.63	3.68	3.74	3.26	2.92	2.97	2.83	2.30	3.42	2.36	2.53	2.29	2.27	1.79	1.49	1.86
<i>Korean Monolingual Models</i>																				
kogpt2-base-v2-125M	28.64	5.56	5.45	4.82	4.43	3.82	4.02	4.07	3.19	3.71	3.24	2.85	3.20	2.35	2.77	2.31	1.93	1.58	1.33	1.63
ko-gpt-trinity-1.2B-v0.5	27.94	5.99	5.70	5.15	4.53	3.54	3.82	3.88	3.12	3.69	3.15	3.12	2.77	2.51	2.30	2.12	2.05	1.79	1.45	1.75
polyglot-ko-1.3b	28.72	7.26	5.73	4.67	4.59	3.66	3.66	3.63	2.79	2.54	2.69	2.70	2.78	2.29	2.08	2.33	2.24	2.04	1.81	1.27
polyglot-ko-3.8b	28.66	6.99	5.98	4.84	4.70	3.62	3.70	3.52	2.93	2.64	2.84	2.50	2.84	2.33	2.05	2.36	2.19	1.99	1.71	1.26
polyglot-ko-5.8b	28.94	6.86	5.70	4.56	4.49	3.69	3.70	3.56	3.05	2.65	2.93	2.65	2.83	2.32	2.26	2.33	2.13	1.95	1.66	1.30
polyglot-ko-12.8b	29.49	6.88	5.58	4.50	4.61	3.77	3.72	3.57	2.88	2.48	2.79	2.74	3.00	2.21	2.23	2.30	2.11	1.83	1.72	1.18
kogpt-6B	29.90	8.52	5.74	4.53	4.18	3.37	3.66	3.29	2.26	1.33	2.08	2.19	2.23	2.03	1.99	2.29	3.04	2.19	2.06	1.79
<i>Multilingual Models</i>																				
xglm-564M	27.51	7.07	6.04	5.39	4.46	3.44	3.43	3.60	3.23	2.90	2.98	2.41	2.83	2.39	1.92	2.12	2.31	2.03	1.86	1.48
xglm-1.7B	28.15	7.30	6.16	5.57	4.47	3.24	3.31	3.59	3.07	2.94	2.76	2.34	2.76	2.42	1.74	2.18	2.38	2.09	1.92	1.41
xglm-4.5B	27.87	6.97	5.98	5.74	4.65	3.42	3.58	3.91	3.14	3.07	3.01	1.95	2.48	2.42	1.58	2.17	2.36	2.09	1.92	1.42
xglm-7.5B	27.56	6.33	5.86	5.67	4.89	3.67	3.61	3.82	3.36	3.15	3.29	1.85	2.46	2.33	1.73	1.97	2.33	2.06	2.01	1.50
mGPT-1.3B	27.62	7.00	5.92	5.53	4.70	3.70	3.77	3.76	2.99	3.15	2.83	2.23	2.89	2.09	1.88	2.43	2.08	2.02	2.03	1.62
mGPT-13B	27.30	6.45	5.80	5.73	4.79	3.88	3.61	3.79	3.14	3.39	3.13	1.98	2.67	2.17	1.66	2.11	2.21	1.98	2.01	1.71
<i>Korean Continually Pre-trained Models</i>																				
Yi-Ko-6B	28.94	5.96	5.76	5.09	4.56	3.88	3.34	3.96	3.38	3.62	3.39	2.29	3.09	2.59	1.79	2.42	1.97	1.73	1.43	1.47
llama-2-ko-7b	28.42	6.34	6.05	5.24	4.55	3.68	3.34	3.62	3.35	3.46	3.24	2.25	2.86	2.69	1.80	2.10	2.02	1.88	1.51	1.66
open-llama-2-ko-7b	30.68	6.54	5.76	5.37	4.75	3.57	3.01	4.04	2.50	3.65	2.42	2.60	3.51	2.66	1.28	2.53	1.73	1.63	1.55	1.59
llama-2-koen-13b	29.18	6.35	6.10	5.48	4.60	3.59	3.01	3.66	3.28	3.43	3.14	2.27	3.08	2.70	1.09	2.22	2.16	1.89	1.60	1.51
OPEN-SOLAR-KO-10.7B	31.30	5.23	5.49	5.15	4.86	3.75	3.26	4.01	3.35	3.60	3.35	2.08	3.44	2.53	1.58	2.40	1.82	1.50	1.50	1.42
SOLAR-KOEN-10.8B	28.47	6.61	6.06	5.00	4.42	3.64	3.61	3.58	3.29	3.50	3.17	1.95	2.98	2.64	1.85	2.13	2.11	1.83	1.48	1.63
<i>Non-Korean-trained Models</i>																				
Yi-6B	25.38	5.48	5.97	4.37	4.40	3.05	2.92	3.10	4.77	4.35	3.80	5.67	2.26	2.70	3.19	1.04	1.05	1.89	0.87	2.11
Llama-2-7b	28.57	5.25	4.36	4.07	4.50	3.52	2.56	3.78	3.58	4.16	2.81	5.64	2.98	2.09	2.43	0.88	1.21	1.33	1.55	2.27
Llama-2-13b	27.40	6.29	5.27	5.11	4.49	3.16	3.23	4.14	3.16	4.41	2.88	3.22	2.86	2.10	2.41	1.18	1.43	2.10	1.84	2.04
SOLAR-10.7B-v1.0	19.99	1.50	2.03	1.39	2.80	4.03	6.65	1.69	9.62	2.46	4.26	9.89	3.83	2.10	8.39	1.50	0.89	0.25	1.09	1.91

Table A.22: XPOS distribution (separated tag) in Korean corpora and Korean sentences generated by language models in the zero-shot task.

Model	NNG	EC	VV	ETM	JKB	NNB	SF	JX	XSV	EF	NNP	JKO	JKS	SS	EP	MAG	SN	VX	XSN	VCP
Average	29.55	6.41	5.97	4.95	4.22	3.53	3.47	3.24	3.21	3.15	2.93	2.92	2.83	2.34	2.23	2.21	2.19	1.81	1.76	1.75
<i>Corpora</i>																				
corpus spoken	23.34	9.03	7.91	5.24	4.10	2.98	3.34	3.35	2.29	2.96	1.12	3.24	3.46	0.07	1.98	4.01	0.11	2.31	2.08	2.02
corpus written	28.35	5.41	5.84	4.96	4.42	4.05	3.20	3.68	3.38	3.19	3.54	3.66	2.54	3.40	2.34	1.48	2.83	1.54	1.43	1.29
ko gsd-ud-TB	26.47	7.36	6.12	4.59	4.63	3.74	2.92	3.68	3.26	2.83	3.42	2.97	2.36	2.30	2.29	2.27	2.53	1.79	1.49	1.41
<i>Korean Monolingual Models</i>																				
kogpt2-base-v2-125M	26.40	5.90	5.74	4.84	4.30	3.86	3.50	4.14	3.49	3.47	3.32	3.48	2.56	2.75	2.23	2.31	2.52	1.78	1.43	1.44
ko-gpt-trinity-1.2B-v0.5	27.42	6.11	5.88	5.23	4.35	3.50	3.37	3.56	3.42	3.25	2.56	3.38	2.83	2.88	2.14	2.25	2.26	1.98	1.46	1.59
polyglot-ko-1.3b	31.40	6.74	5.79	4.81	4.01	3.32	2.96	3.51	3.09	2.85	2.78	2.57	2.96	2.35	1.93	2.29	2.01	2.12	1.85	1.76
polyglot-ko-3.8b	31.40	6.75	5.95	4.75	3.97	3.42	3.17	3.50	3.17	3.07	2.70	2.73	2.91	2.10	1.84	2.37	2.09	2.14	1.79	1.62
polyglot-ko-5.8b	31.48	6.34	5.93	4.56	4.15	3.40	3.30	3.40	3.06	3.21	3.15	2.77	2.90	2.30	2.07	1.96	2.22	1.91	1.72	1.59
polyglot-ko-12.8b	32.08	6.60	5.75	4.25	3.91	3.30	3.22	3.47	3.07	3.10	3.34	2.55	2.89	2.60	2.02	1.97	2.19	1.97	1.73	1.61
kogpt-6B	35.37	7.12	5.42	4.19	3.00	4.05	3.04	3.26	2.37	2.74	2.72	0.62	2.15	1.98	1.60	2.79	2.26	1.91	2.00	2.92
<i>Multilingual Models</i>																				
xglm-564M	28.69	8.14	6.61	4.68	4.03	3.58	3.57	3.00	2.91	3.14	2.24	2.14	2.30	1.61	2.79	3.05	2.11	1.95	2.32	2.10
xglm-1.7B	29.08	9.05	6.73	4.71	4.38	3.29	2.96	2.88	3.04	2.54	2.12	2.28	2.28	1.48	2.65	3.10	1.65	2.22	2.40	1.96
xglm-4.5B	30.08	7.16	6.19	4.98	4.51	3.84	3.26	2.77	3.17	2.98	2.54	2.67	2.59	1.72	2.45	2.46	2.24	2.06	2.18	1.83
xglm-7.5B	30.04	6.51	6.00	5.11	4.66	3.98	3.14	3.00	3.38	3.05	2.72	2.90	2.87	1.67	2.44	2.29	2.33	1.96	2.31	1.68
mGPT-1.3B	29.44	6.58	5.90	5.26	4.52	3.92	3.30	3.38	2.97	2.93	2.66	2.57	2.48	1.89	3.00	2.35	2.23	1.84	2.39	2.13
mGPT-13B	28.80	6.42	6.03	5.41	4.36	3.90	3.08	3.34	3.16	3.01	2.53	3.25	2.83	1.67	2.54	2.54	2.08	1.98	2.19	1.90
<i>Korean Continually Pre-trained Models</i>																				
Yi-Ko-6B	30.69	5.73	5.85	5.25	4.72	2.99	3.41	3.40	3.75	3.39	3.55	3.27	3.17	2.20	2.54	1.68	1.51	1.72	1.53	1.60
llama-2-ko-7b	29.31	6.55	6.39	5.03	4.58	2.84	3.56	3.35	3.36	3.44	3.31	3.25	3.21	2.01	2.34	1.99	1.55	1.93	1.63	1.59
open-llama-2-ko-7b	30.84	6.92	6.34	5.48	4.74	2.28	3.23	3.05	3.27	2.87	3.09	2.89	3.14	3.20	2.42	1.95	0.53	1.89	1.66	1.77
llama-2-koen-13b	31.24	6.44	6.16	5.68	5.01	2.42	3.11	3.29	3.64	3.02	3.33	3.20	2.90	2.04	2.27	1.92	0.71	1.94	1.77	1.82
OPEN-SOLAR-KO-10.7B	31.71	5.58	6.06	5.43	5.02	2.55	3.43	3.14	3.49	3.41	3.42	3.27	3.14	2.47	2.41	1.83	1.01	1.69	1.60	1.73
SOLAR-KOEN-10.8B	30.53	6.03	6.11	5.52	4.23	3.48	3.12	3.05	3.51	3.05	3.20	3.50	3.25	1.91	2.21	1.81	1.75	1.80	1.67	1.63
<i>Non-Korean-trained Models</i>																				
Yi-6B	30.13	5.32	5.79	5.56	4.17	3.42	3.39	2.56	3.22	2.93	2.41	3.38	2.80	3.54	1.75	1.51	2.83	1.64	1.33	1.46
Llama-2-7b	29.28	4.94	5.44	5.30	4.26	3.54	3.27	2.32	3.43	2.89	3.83	3.56	2.64	4.79	2.01	1.41	2.64	1.23	1.49	1.05
Llama-2-13b	29.54	5.14	5.52	5.25	4.08	4.41	3.23	2.47	3.84	3.11	4.03	3.84	2.86	2.99	2.27	1.44	3.04	1.49	1.53	1.16
SOLAR-10.7B-v1.0	25.21	2.67	3.76	2.67	1.58	5.82	9.09	3.76	2.67	5.45	2.67	1.94	3.52	3.03	1.45	2.42	7.76	0.36	0.85	2.79

Table A.23: XPOS distribution (separated tag) in Korean corpora and Korean sentences generated by language models in the few-shot task.

Model	flat	punct	nsubj	advmod	obj	root	obl	advcl	acl:relcl	conj	dep	nmod:poss	case	nmod	appos	amod	ccomp	det	nummod	nsubj:pass
Average	15.99	14.15	9.53	7.85	7.18	6.60	6.18	5.60	4.51	3.64	3.55	2.48	2.29	2.17	2.12	2.06	1.17	0.81	0.61	0.37
<i>Corpora</i>																				
corpus spoken	11.01	9.40	13.08	13.18	7.97	6.44	5.74	7.66	3.81	2.99	4.23	1.93	1.08	1.18	0.12	3.31	1.18	3.82	0.97	0.07
corpus written	16.59	14.45	10.15	5.83	7.58	5.82	6.57	5.08	4.56	3.62	3.76	2.75	2.95	2.31	2.53	1.76	1.15	0.57	0.70	0.38
ko gsd-ud-TB	15.35	12.94	9.51	7.88	7.27	7.76	6.79	5.70	3.99	4.79	3.09	3.00	2.05	2.18	1.91	1.97	0.79	0.69	0.70	0.67
<i>Korean Monolingual Models</i>																				
kogpt2-base-v2-125M	16.64	13.56	9.65	7.67	7.71	5.83	6.74	5.30	4.19	3.27	3.95	2.59	2.61	2.46	1.80	1.71	1.55	0.57	0.54	0.53
ko-gpt-trinity-1.2B-v0.5	15.59	14.01	9.30	7.78	7.64	5.72	6.61	5.71	4.54	3.55	4.10	2.68	2.44	2.20	2.18	1.99	1.18	0.73	0.50	0.42
polyglot-ko-1.3b	18.25	12.67	9.23	8.27	5.99	5.19	6.31	6.52	3.90	3.89	3.89	1.86	2.78	2.45	2.05	1.93	1.69	0.85	0.56	0.38
polyglot-ko-3.8b	18.11	12.50	9.42	8.06	6.16	5.49	6.48	6.21	4.07	3.90	3.96	1.84	2.66	2.49	2.02	2.05	1.65	0.74	0.56	0.40
polyglot-ko-5.8b	18.05	13.36	9.38	7.85	6.14	5.58	6.11	6.03	4.00	4.16	3.68	1.95	2.67	2.46	2.07	1.86	1.55	0.79	0.51	0.44
polyglot-ko-12.8b	18.63	13.17	9.27	7.61	5.88	5.41	6.42	6.18	3.81	4.20	3.69	1.76	2.80	2.52	2.12	1.85	1.68	0.77	0.52	0.38
kogpt-6B	19.13	9.79	8.33	10.33	5.06	5.93	5.54	7.29	4.06	3.73	3.57	2.71	2.69	2.62	2.07	2.29	1.51	0.94	0.76	0.26
<i>Multilingual Models</i>																				
xglm-564M	16.04	12.54	9.64	8.90	6.91	7.43	5.87	5.72	4.99	3.27	3.64	2.31	1.97	2.06	1.83	2.66	1.11	0.76	0.64	0.32
xglm-1.7B	17.42	12.02	9.38	9.06	6.88	7.18	6.00	5.75	5.25	3.19	3.78	2.16	1.94	1.90	1.80	2.35	1.08	0.66	0.60	0.29
xglm-4.5B	15.90	11.31	9.57	9.05	7.06	6.43	6.46	6.07	5.24	3.33	4.54	2.18	2.15	2.09	1.64	2.55	1.45	0.76	0.51	0.35
xglm-7.5B	15.66	11.57	9.29	9.49	7.17	6.43	6.66	5.57	5.17	3.42	4.15	2.35	2.23	2.28	1.57	2.71	1.15	0.71	0.53	0.43
mGPT-1.3B	16.70	11.94	8.69	8.80	7.19	5.96	6.46	6.18	4.94	3.36	4.39	2.61	2.25	2.20	1.61	2.47	1.55	0.45	0.51	0.39
mGPT-13B	15.33	11.76	8.99	9.41	7.60	5.90	6.57	5.97	4.98	3.57	4.50	2.76	2.25	2.02	1.46	2.76	1.27	0.61	0.45	0.41
<i>Korean Continually Pre-trained Models</i>																				
Yi-Ko-6B	16.35	12.51	10.29	7.85	7.65	6.36	6.70	5.68	4.86	3.47	3.80	2.35	2.20	2.37	1.68	1.85	1.19	0.69	0.54	0.52
llama-2-ko-7b	15.58	12.59	10.53	7.91	7.49	6.40	6.62	5.77	5.03	3.60	3.68	2.63	2.15	2.23	1.81	2.03	1.06	0.82	0.60	0.44
open-llama-2-ko-7b	18.14	11.19	9.78	7.00	7.78	5.65	6.94	5.70	5.38	3.06	3.91	2.56	2.25	2.51	2.21	1.81	1.41	0.71	0.61	0.43
llama-2-koen-13b	16.83	11.78	10.03	8.30	7.46	6.44	6.59	5.63	5.39	3.29	3.81	2.37	1.98	2.26	1.89	2.07	1.09	0.72	0.59	0.43
OPEN-SOLAR-KO-10.7B	19.06	11.50	9.70	6.97	7.53	6.21	7.33	5.01	5.08	3.34	3.85	2.21	2.22	2.65	1.94	1.76	1.09	0.58	0.56	0.51
SOLAR-KOEN-10.8B	21.01	11.27	9.71	7.23	7.20	6.10	6.12	5.48	4.22	3.55	3.56	2.42	2.07	2.08	1.47	1.82	0.98	0.81	1.48	0.41
<i>Non-Korean-trained Models</i>																				
Yi-6B	10.93	23.15	9.82	4.50	9.02	8.48	5.32	3.73	4.23	3.29	2.25	3.14	2.50	1.30	3.71	1.63	0.70	0.71	0.83	0.10
Llama-2-7b	11.17	21.91	8.14	5.92	8.98	6.61	5.34	4.97	4.61	4.95	1.56	3.64	2.29	1.33	4.19	1.61	0.59	0.55	0.29	0.21
Llama-2-13b	14.53	15.83	8.35	6.46	9.56	5.88	6.06	5.71	5.48	4.18	2.56	3.39	2.13	1.72	2.39	2.21	0.82	0.91	0.40	0.36
SOLAR-10.7B-v1.0	7.73	39.30	8.58	2.79	3.81	15.10	2.35	1.10	1.47	3.56	0.40	2.42	2.24	2.53	4.99	0.44	0.00	0.04	0.51	0.07

Table A.24: Dependency relation distribution in Korean corpora and Korean sentences generated by language models in the zero-shot task.

Model	flat	punct	nsubj	advmod	obj	root	advel	obl	acl:recl	dep	conj	nmod	case	nmod:poss	amod	appos	ccomp	nummod	det	nsubj:pass
Average	19.71	12.49	9.64	7.42	6.64	6.63	6.01	5.93	4.59	3.69	2.70	2.37	2.25	2.23	1.99	1.79	0.95	0.75	0.64	0.42
<i>Corpora</i>																				
corpus spoken	11.01	9.40	13.08	13.18	7.97	6.44	7.66	5.74	3.81	4.23	2.99	1.18	1.08	1.93	3.31	0.12	1.18	0.97	3.82	0.07
corpus written	16.59	14.45	10.15	5.83	7.58	5.82	5.08	6.57	4.56	3.76	3.62	2.31	2.95	2.75	1.76	2.53	1.15	0.70	0.57	0.38
ko gsd-ud-TB	15.35	12.94	9.51	7.88	7.27	7.76	5.70	6.79	3.99	3.09	4.79	2.18	2.05	3.00	1.97	1.91	0.79	0.70	0.69	0.67
<i>Korean Monolingual Models</i>																				
kogpt2-base-v2-125M	15.08	14.30	10.41	8.71	7.19	6.23	5.28	6.09	4.05	3.77	2.98	2.27	2.78	2.93	1.99	1.52	1.41	0.56	0.58	0.44
ko-gpt-trinity-1.2B-v0.5	16.11	14.30	9.66	7.86	6.94	6.13	5.45	6.24	4.84	3.54	3.21	2.25	2.21	2.95	2.04	2.38	0.91	0.66	0.64	0.52
polyglot-ko-1.3b	22.56	11.86	9.71	7.24	5.88	5.28	6.93	5.44	4.04	4.15	2.28	2.69	2.53	1.68	1.81	1.88	0.97	0.72	0.73	0.38
polyglot-ko-3.8b	22.98	11.39	9.65	7.47	6.12	5.71	6.78	5.26	4.25	3.89	2.28	2.63	2.36	1.55	1.79	1.88	0.90	0.74	0.75	0.39
polyglot-ko-5.8b	22.55	12.34	9.77	6.52	6.23	5.87	6.55	5.69	4.11	3.79	2.40	2.70	2.38	1.70	1.64	1.94	0.96	0.72	0.67	0.36
polyglot-ko-12.8b	24.11	12.57	9.42	6.25	5.83	5.76	6.64	5.37	3.90	3.52	2.36	2.80	2.44	1.45	1.51	2.17	1.00	0.66	0.64	0.39
kogpt-6B	28.20	9.40	7.68	7.44	4.79	6.63	6.92	3.65	3.85	3.58	1.94	3.37	3.12	2.15	1.55	2.18	0.67	0.98	0.59	0.20
<i>Multilingual Models</i>																				
xglm-564M	19.49	10.71	8.82	10.12	6.06	8.84	7.17	4.67	3.99	3.76	1.81	2.63	2.14	1.64	2.64	1.17	1.08	0.81	0.49	0.17
xglm-1.7B	19.76	9.16	8.91	10.28	6.78	8.22	7.71	5.26	4.24	3.66	1.92	2.40	1.96	1.63	2.70	1.01	1.16	0.67	0.50	0.13
xglm-4.5B	22.01	10.02	8.47	8.21	6.48	6.79	6.94	5.89	4.53	4.16	1.82	2.89	2.34	1.89	2.20	1.45	0.91	0.82	0.44	0.31
xglm-7.5B	21.66	9.69	8.80	8.08	6.34	6.10	6.25	6.01	4.91	4.40	1.94	3.28	2.44	2.09	2.24	1.43	1.04	0.73	0.47	0.49
mGPT-1.3B	20.00	10.69	8.62	8.73	6.17	6.15	7.35	6.35	4.34	4.95	1.82	2.44	2.21	2.69	2.26	1.71	0.98	0.79	0.21	0.32
mGPT-13B	18.49	10.56	9.25	8.97	7.15	5.71	6.86	6.29	4.44	5.07	2.30	2.29	2.20	3.09	2.41	1.36	0.86	0.67	0.19	0.44
<i>Korean Continually Pre-trained Models</i>																				
Yi-Ko-6B	19.56	11.47	10.08	6.66	6.81	6.30	5.61	7.08	5.43	3.83	2.86	2.35	2.24	2.26	1.61	1.67	1.13	0.65	0.62	0.59
llama-2-ko-7b	17.99	11.89	10.54	7.67	7.04	7.01	6.08	6.70	5.09	3.41	3.07	2.18	2.03	2.44	1.74	1.60	0.92	0.61	0.54	0.48
open-llama-2-ko-7b	19.15	12.66	9.70	6.65	6.21	6.37	6.43	7.02	5.36	3.85	2.54	1.91	2.03	1.98	1.81	2.17	1.05	0.75	0.48	0.41
llama-2-koen-13b	20.50	10.20	9.41	6.94	6.78	6.01	6.23	7.59	5.95	3.92	2.79	1.97	2.08	2.13	1.93	1.49	1.04	0.77	0.54	0.52
OPEN-SOLAR-KO-10.7B	20.89	11.38	9.57	6.55	6.68	6.28	5.64	7.37	5.61	3.67	2.54	2.21	2.18	1.83	1.82	2.11	0.88	0.62	0.59	0.50
SOLAR-KOEN-10.8B	24.43	10.67	9.72	6.01	7.07	5.48	5.32	6.23	4.86	3.88	2.57	2.07	2.01	2.49	1.76	1.41	0.87	1.36	0.51	0.44
<i>Non-Korean-trained Models</i>																				
Yi-6B	17.96	16.56	9.60	4.93	7.62	5.84	4.30	6.52	5.62	2.93	3.72	1.65	2.30	2.71	2.08	2.12	0.94	0.89	0.60	0.30
Llama-2-7b	19.39	17.55	8.22	4.62	7.49	5.25	4.86	6.10	4.98	3.09	3.68	1.83	2.31	2.60	2.26	3.00	0.75	0.46	0.30	0.33
Llama-2-13b	22.16	13.62	9.25	4.47	7.78	5.51	5.03	6.26	4.99	3.47	2.77	2.70	2.47	2.30	1.91	1.81	1.01	0.57	0.44	0.54
SOLAR-10.7B-v1.0	14.41	25.05	12.53	5.64	4.38	15.03	1.46	1.88	3.55	0.63	3.34	2.30	1.67	2.09	1.04	2.51	0.21	0.84	0.00	1.25

Table A.25: Dependency relation distribution in Korean corpora and Korean sentences generated by language models in the few-shot task.

Tag	Description
acl	clausal modifier of noun (adnominal clause)
acl:relcl	relative clause modifier
advcl	adverbial clause modifier
advcl:relcl	adverbial relative clause modifier
advmod	adverbial modifier
advmod:emph	emphasizing word, intensifier
advmod:lmod	locative adverbial modifier
amod	adjectival modifier
appos	appositional modifier
aux	auxiliary
aux:pass	passive auxiliary
case	case marking
cc	coordinating conjunction
cc:preconj	preconjunct
ccomp	clausal complement
clf	classifier
compound	compound
compound:lvc	light verb construction
compound:prt	phrasal verb particle
compound:redup	reduplicated compounds
compound:svc	serial verb compounds
conj	conjunct
cop	copula
csubj	clausal subject
csubj:outer	outer clause clausal subject
csubj:pass	clausal passive subject
dep	unspecified dependency
det	determiner
det:numgov	pronominal quantifier governing the case of the noun
det:nummod	pronominal quantifier agreeing in case with the noun
det:poss	possessive determiner
discourse	discourse element
dislocated	dislocated elements
expl	expletive
expl:impers	impersonal expletive
expl:pass	reflexive pronoun used in reflexive passive
expl:pvc	reflexive clitic with an inherently reflexive verb
fixed	fixed multiword expression
flat	flat expression
flat:foreign	foreign words
flat:name	names
goeswith	goes with
iobj	indirect object
list	list
mark	marker
nmod	nominal modifier
nmod:poss	possessive nominal modifier
nmod:tmod	temporal modifier
nsubj	nominal subject
nsubj:outer	outer clause nominal subject
nsubj:pass	passive nominal subject
nummod	numeric modifier
nummod:gov	numeric modifier governing the case of the noun
obj	object
obl	oblique nominal
obl:agent	agent modifier
obl:arg	oblique argument
obl:lmod	locative modifier
obl:tmod	temporal modifier
orphan	orphan
parataxis	parataxis
punct	punctuation
reparandum	overridden disfluency
root	root
vocative	vocative
xcomp	open clausal complement

Table A.26: Universal dependency relations tags and descriptions.

Model	Arc Direction		Total Arcs		Left Arcs		Right Arcs	
	Left	Right	Length	Std.	Length	Std.	Length	Std.
<i>Corpora</i>								
corpus spoken	72.88	27.12	3.42	8.21	3.90	9.04	2.13	5.18
corpus written	59.35	40.65	3.10	4.46	3.81	5.30	2.06	2.50
ko gsd-ud-TB	61.43	38.57	2.97	4.20	3.40	4.72	2.28	3.09
<i>Korean Monolingual Models</i>								
kogpt2-base-v2-125M	61.52	38.48	3.31	4.82	4.21	5.76	1.88	1.98
ko-gpt-trinity-1.2B-v0.5	61.65	38.35	3.32	5.05	4.16	6.05	1.96	2.16
polyglot-ko-1.3b	60.04	39.96	3.37	5.29	4.22	6.34	2.09	2.65
polyglot-ko-3.8b	60.03	39.97	3.26	4.86	4.07	5.85	2.04	2.29
polyglot-ko-5.8b	59.22	40.78	3.29	5.00	4.12	6.02	2.09	2.47
polyglot-ko-12.8b	58.55	41.45	3.40	5.24	4.30	6.34	2.13	2.58
kogpt-6B	60.62	39.38	3.26	4.95	3.97	5.96	2.16	2.37
<i>Multilingual Models</i>								
xglm-564M	62.65	37.35	2.75	3.81	3.20	4.37	2.01	2.44
xglm-1.7B	62.06	37.94	2.78	3.91	3.21	4.37	2.07	2.87
xglm-4.5B	63.97	36.03	2.82	4.03	3.37	4.77	1.84	1.80
xglm-7.5B	63.80	36.20	2.88	3.97	3.50	4.69	1.78	1.70
mGPT-1.3B	62.67	37.33	3.00	4.49	3.64	5.29	1.92	2.27
mGPT-13B	64.16	35.84	3.03	4.53	3.70	5.35	1.83	1.93
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	62.27	37.73	3.06	4.32	3.81	5.13	1.83	1.88
llama-2-ko-7b	63.09	36.91	2.95	4.22	3.57	4.97	1.90	2.08
open-llama-2-ko-7b	61.97	38.03	3.09	4.44	3.79	5.29	1.95	2.01
llama-2-koen-13b	62.36	37.64	2.97	4.23	3.61	5.01	1.89	2.02
OPEN-SOLAR-KO-10.7B	59.78	40.22	3.00	4.08	3.76	4.88	1.88	1.94
SOLAR-KOEN-10.8B	58.95	41.05	3.15	4.63	3.89	5.55	2.10	2.44
<i>Non-Korean-trained Models</i>								
Yi-6B	58.31	41.69	2.72	3.54	3.15	4.15	2.12	2.33
Llama-2-7b	57.74	42.26	3.17	4.81	3.74	5.75	2.39	2.91
Llama-2-13b	61.46	38.54	3.06	4.74	3.68	5.68	2.06	2.32
SOLAR-10.7B-v1.0	42.23	57.77	2.43	2.93	2.77	3.36	2.19	2.53

Table A.27: Dependency arc direction and lengths in Korean corpora and Korean sentences generated by language models in the zero-shot task.

Model	Direction		Total Arcs		Left Arcs		Right Arcs	
	Left	Right	Length	Std.	Length	Std.	Length	Std.
<i>Corpora</i>								
corpus spoken	72.88	27.12	3.42	8.21	3.90	9.04	2.13	5.18
corpus written	59.35	40.65	3.10	4.46	3.81	5.30	2.06	2.50
ko gsd-ud-TB	61.43	38.57	2.97	4.20	3.40	4.72	2.28	3.09
<i>Korean Monolingual Models</i>								
kogpt2-base-v2-125M	62.82	37.18	3.10	4.40	3.86	5.21	1.80	1.85
ko-gpt-trinity-1.2B-v0.5	61.03	38.97	3.02	4.31	3.71	5.15	1.95	2.08
polyglot-ko-1.3b	57.62	42.38	3.17	4.53	4.01	5.51	2.03	2.20
polyglot-ko-3.8b	57.20	42.80	3.11	4.34	3.91	5.28	2.03	2.14
polyglot-ko-5.8b	56.67	43.33	3.06	4.30	3.90	5.29	1.96	1.98
polyglot-ko-12.8b	54.83	45.17	3.18	4.41	4.06	5.42	2.11	2.31
kogpt-6B	51.30	48.70	3.19	4.14	3.88	5.10	2.46	2.61
<i>Multilingual Models</i>								
xglm-564M	61.26	38.74	2.60	3.16	3.01	3.69	1.95	1.86
xglm-1.7B	63.24	36.76	2.69	3.50	3.12	4.09	1.94	1.91
xglm-4.5B	59.48	40.52	2.81	3.72	3.37	4.46	1.99	1.97
xglm-7.5B	59.96	40.04	2.92	3.96	3.60	4.77	1.90	1.80
mGPT-1.3B	61.36	38.64	2.79	3.51	3.31	4.17	1.96	1.78
mGPT-13B	63.19	36.81	2.93	4.13	3.57	4.89	1.83	1.85
<i>Korean Continually Pre-trained Models</i>								
Yi-Ko-6B	60.28	39.72	2.80	3.70	3.46	4.45	1.80	1.65
llama-2-ko-7b	61.76	38.24	2.71	3.57	3.24	4.19	1.84	1.96
open-llama-2-ko-7b	60.41	39.59	2.78	3.73	3.27	4.41	2.03	2.15
llama-2-koen-13b	60.75	39.25	2.87	3.91	3.51	4.69	1.88	1.84
OPEN-SOLAR-KO-10.7B	58.71	41.29	2.82	3.67	3.46	4.41	1.91	1.88
SOLAR-KOEN-10.8B	57.13	42.87	2.96	4.13	3.64	5.00	2.06	2.25
<i>Non-Korean-trained Models</i>								
Yi-6B	58.86	41.14	2.96	4.08	3.52	4.87	2.15	2.33
Llama-2-7b	55.33	44.67	3.33	5.22	4.02	6.11	2.47	3.67
Llama-2-13b	56.47	43.53	3.13	4.53	3.93	5.51	2.09	2.42
SOLAR-10.7B-v1.0	49.63	50.37	2.16	2.19	2.67	2.80	1.66	1.15

Table A.28: Dependency arc direction and lengths in Korean corpora and Korean sentences generated by language models in the few-shot task.

Model	A/An	Plural	Det	Passive	No Subj.	No Obj.	No Subj.Obj.	Obj-Verb Order	Verb-Obj Order	Nouns	Words	Sentences with Obj&Verb	Sentences
<i>Corpora</i>													
corpus spoken	520	2389	3962	100	2484	4230	1391	5717	25	47706	155171	5742	10000
corpus written	335	1831	165	575	1038	3143	300	6807	15	72744	171912	6822	10000
ko gsd-ud-TB	98	575	99	328	1131	1945	451	2409	16	22915	56687	2425	4400
<i>Korean Monolingual Models</i>													
kogpt2-base-v2-125M	242	1730	180	759	995	2901	338	6456	1	66060	161075	6457	9396
ko-gpt-trinity-1.2B-v0.5	166	1258	202	431	932	2021	293	4415	9	44735	113386	4424	6481
polyglot-ko-1.3b	126	1250	211	278	645	1509	317	2705	5	33023	82201	2710	4270
polyglot-ko-3.8b	130	1124	165	286	662	1529	294	2742	17	32089	78916	2759	4334
polyglot-ko-5.8b	127	1168	208	305	703	1733	331	2871	6	33929	83328	2877	4651
polyglot-ko-12.8b	114	1136	200	273	639	1609	282	2743	9	33626	81292	2752	4396
kogpt-6B	163	1287	210	185	1164	2132	660	2420	19	31762	78031	2439	4627
<i>Multilingual Models</i>													
xglm-564M	71	648	57	95	608	1097	356	1341	9	12774	33355	1350	2477
xglm-1.7B	69	685	66	92	630	1074	339	1427	7	13851	35231	1434	2529
xglm-4.5B	85	831	80	134	495	1059	229	1718	7	16776	43721	1725	2812
xglm-7.5B	102	1064	104	206	517	1213	206	2101	9	19805	52210	2110	3358
mGPT-1.3B	95	905	61	172	567	1015	251	1910	11	19533	49771	1921	2964
mGPT-13B	117	1432	117	289	678	1389	212	3135	4	29723	77340	3139	4566
<i>Korean Continually Pre-trained Models</i>													
Yi-Ko-6B	185	1323	172	512	761	2331	261	4534	9	44518	108608	4543	6910
llama-2-ko-7b	201	1404	209	409	883	2410	378	4296	11	42202	105591	4307	6754
open-llama-2-ko-7b	105	623	84	191	271	830	95	1923	1	21052	48974	1924	2765
llama-2-koen-13b	172	1060	123	292	677	1780	308	3005	7	30533	74875	3012	4820
OPEN-SOLAR-KO-10.7B	125	762	86	317	461	1357	135	2910	2	30476	68958	2912	4279
SOLAR-KOEN-10.8B	227	1168	258	347	870	2102	337	3768	10	42291	96963	3778	5911
<i>Non-Korean-trained Models</i>													
Yi-6B	28	78	23	13	372	479	181	731	24	5219	14728	755	1249
Llama-2-7b	14	304	33	37	381	479	183	833	28	7806	20485	861	1354
Llama-2-13b	28	652	123	115	442	553	155	1563	11	14356	36306	1574	2134
SOLAR-10.7B-v1.0	1	28	0	1	226	325	193	52	6	925	2728	58	412

Table A.29: Detection of English translationese artifacts in Korean corpora and Korean sentences generated by language models in the zero-shot task.

Model	A/An	Plural	Det	Passive	No Subj.	No Obj.	No Subj.Obj.	Obj-Verb Order	Verb-Obj Order	Nouns	Words	Sentences with Obj&Verb	Sentences
<i>Corpora</i>													
corpus spoken	520	2389	3962	100	2484	4230	1391	5717	25	47706	155171	5742	10000
corpus written	335	1831	165	575	1038	3143	300	6807	15	72744	171912	6822	10000
ko gsd-ud-TB	98	575	99	328	1131	1945	451	2409	16	22915	56687	2425	4400
<i>Korean Monolingual Models</i>													
kogpt2-base-v2-125M	344	2094	74	672	1053	3767	418	6503	8	64727	166244	6511	10361
ko-gpt-trinity-1.2B-v0.5	245	1307	2	483	807	2307	358	3952	11	40517	103089	3963	6315
polyglot-ko-1.3b	224	1036	79	284	555	1587	268	2655	6	36361	81347	2661	4297
polyglot-ko-3.8b	205	942	45	254	618	1563	284	2589	5	33067	73535	2594	4200
polyglot-ko-5.8b	143	671	83	181	463	1238	240	2020	5	25490	56245	2025	3299
polyglot-ko-12.8b	124	725	54	205	513	1359	248	2080	4	28003	60580	2084	3487
kogpt-6B	118	463	6	61	677	1178	381	1029	8	16894	33781	1037	2238
<i>Multilingual Models</i>													
xglm-564M	59	501	0	36	790	1164	441	962	6	9837	24685	968	2181
xglm-1.7B	70	739	3	38	912	1210	445	1265	4	12507	31055	1269	2553
xglm-4.5B	133	736	4	114	703	1118	338	1484	7	16963	39041	1491	2650
xglm-7.5B	172	1383	8	285	627	1414	283	2250	1	26804	61093	2251	3726
mGPT-1.3B	110	612	1	97	421	836	253	1144	6	13901	32555	1150	2003
mGPT-13B	169	1092	3	253	488	1209	209	2379	5	26723	63451	2384	3620
<i>Korean Continually Pre-trained Models</i>													
Yi-Ko-6B	139	673	12	263	279	1129	102	1993	2	21751	49852	1995	3142
llama-2-ko-7b	185	1152	59	356	868	2314	419	3415	10	34500	82388	3425	5779
open-llama-2-ko-7b	79	354	4	82	234	604	127	794	2	9193	22209	796	1414
llama-2-koen-13b	144	713	3	214	315	1000	153	1710	3	19526	45355	1713	2727
OPEN-SOLAR-KO-10.7B	93	535	2	177	247	862	80	1560	2	17072	38834	1562	2437
SOLAR-KOEN-10.8B	90	835	75	257	394	1130	147	2338	6	29885	63643	2344	3487
<i>Non-Korean-trained Models</i>													
Yi-6B	70	343	78	105	473	899	245	1570	8	18062	42767	1578	2497
Llama-2-7b	44	328	33	127	424	799	214	1472	7	19080	43618	1479	2289
Llama-2-13b	79	646	82	277	362	1043	146	2151	1	27247	58291	2152	3213
SOLAR-10.7B-v1.0	0	5	0	3	25	58	19	12	0	205	479	12	72

Table A.30: Detection of English translationese artifacts in Korean corpora and Korean sentences generated by language models in the few-shot task.

Model	Zero-shot Count			Percentage (%)			Few-shot Count			Percentage (%)		
	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg
<i>Corpora</i>												
corpus spoken	6506	1033	2461	65.1	10.3	24.6	6506	1033	2461	65.1	10.3	24.6
corpus written	2533	4379	3088	25.3	43.8	30.9	2533	4379	3088	25.3	43.8	30.9
<i>Korean Monolingual Models</i>												
kogpt2-base-v2-125M	3113	3752	2531	33.1	39.9	26.9	2892	3887	3582	27.9	37.5	34.6
ko-gpt-trinity-1.2B-v0.5	2248	2254	1979	34.7	34.8	30.5	2138	1725	2452	33.9	27.3	38.8
polyglot-ko-1.3b	1566	1358	1346	36.7	31.8	31.5	1650	1004	1643	38.4	23.4	38.2
polyglot-ko-3.8b	1597	1506	1231	36.8	34.7	28.4	1743	981	1476	41.5	23.4	35.1
polyglot-ko-5.8b	1832	1489	1330	39.4	32.0	28.6	1412	743	1144	42.8	22.5	34.7
polyglot-ko-12.8b	1652	1550	1194	37.6	35.3	27.2	1483	751	1253	42.5	21.5	35.9
kogpt-6B	2181	1006	1440	47.1	21.7	31.1	1028	345	865	45.9	15.4	38.7
<i>Multilingual Models</i>												
xglm-564M	1205	678	594	48.6	27.4	24.0	1172	360	649	53.7	16.5	29.8
xglm-1.7B	1195	648	686	47.3	25.6	27.1	1264	404	885	49.5	15.8	34.7
xglm-4.5B	1267	704	841	45.1	25.0	29.9	1156	485	1009	43.6	18.3	38.1
xglm-7.5B	1263	1110	985	37.6	33.1	29.3	1495	732	1499	40.1	19.6	40.2
mGPT-1.3B	1319	815	830	44.5	27.5	28.0	850	319	834	42.4	15.9	41.6
mGPT-13B	1683	1559	1324	36.9	34.1	29.0	1260	900	1460	34.8	24.9	40.3
<i>Korean Continually Pre-trained Models</i>												
Yi-Ko-6B	2322	2485	2103	33.6	36.0	30.4	1074	868	1200	34.2	27.6	38.2
llama-2-ko-7b	2377	2337	2040	35.2	34.6	30.2	2110	1658	2011	36.5	28.7	34.8
open-llama-2-ko-7b	1058	771	936	38.3	27.9	33.9	575	288	551	40.7	20.4	39.0
llama-2-koen-13b	1866	1557	1398	38.7	32.3	29.0	986	669	1072	36.2	24.5	39.3
OPEN-SOLAR-KO-10.7B	1451	1314	1514	33.9	30.7	35.4	810	690	937	33.2	28.3	38.4
SOLAR-KOEN-10.8B	2393	1764	1754	40.5	29.8	29.7	1311	822	1354	37.6	23.6	38.8
<i>Non-Korean-trained Models</i>												
Yi-6B	741	180	328	59.3	14.4	26.3	1258	362	877	50.4	14.5	35.1
Llama-2-7b	731	234	389	54.0	17.3	28.7	1002	492	795	43.8	21.5	34.7
Llama-2-13b	1054	492	588	49.4	23.1	27.6	1411	770	1032	43.9	24.0	32.1
SOLAR-10.7B-v1.0	257	35	120	62.4	8.5	29.1	38	10	24	52.8	13.9	33.3

Table A.31: Sentiment classification distribution in Korean corpora and Korean sentences generated by language models.

Bibliography

- Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. <https://doi.org/10.1145/3461702.3462624>
- Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D. R., Smith, N. A., & Tsvetkov, Y. (2023, May 23). Do all languages cost the same? tokenization in the era of commercial language models. <https://doi.org/10.48550/arXiv.2305.13707>
- Ahn, J., & Oh, A. (2021). Mitigating language-dependent ethnic bias in BERT, 533–549. <https://doi.org/10.18653/v1/2021.emnlp-main.42>
- AI, O., Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., ... Dai, Z. (2024, March 7). Yi: Open foundation models by 01.AI. <https://doi.org/10.48550/arXiv.2403.04652>
- Amatriain, X. (2023, May 25). Transformer models: An introduction and catalog. <https://doi.org/10.48550/arXiv.2302.07730>
- Andonian, A., Anthony, Q., Biderman, S., Black, S., Gali, P., Gao, L., Hallahan, E., Levy-Kramer, J., Leahy, C., Nestler, L., Parker, K., Pieler, M., Phang, J., Purohit, S., Schoelkopf, H., Stander, D., Songz, T., Tigges, C., Thérien, B., ... Weinbach, S. (2023, September). *GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch* (Version 2.0.0). <https://doi.org/10.5281/zenodo.5879544>
- Anthropic. (2023). Model card and evaluations for claude models.
- Arora, A., Kaffee, L.-a., & Augenstein, I. (2023). Probing pre-trained language models for cross-cultural differences in values. *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 114–130. Retrieved June 17, 2023, from <https://aclanthology.org/2023.c3nlp-1.12>
- Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations, 4623–4637. <https://doi.org/10.18653/v1/2020.acl-main.421>
- Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597–610. https://doi.org/10.1162/tacl_a_00288
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023, February 8). *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and*

- interactivity* [arXiv.org]. Retrieved October 10, 2023, from <https://arxiv.org/abs/2302.04023v2>
- Bapna, A., & Firat, O. (2019). Simple, scalable adaptation for neural machine translation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1538–1548. <https://doi.org/10.18653/v1/D19-1165>
- Ben Zaken, E., Goldberg, Y., & Ravfogel, S. (2022). BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1–9. <https://doi.org/10.18653/v1/2022.acl-short.1>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 13.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. https://doi.org/10.1162/tacl_a_00051
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.
- Cabello Piqueras, L., & Søgaaard, A. (2022). Are pretrained multilingual models equally fair across languages? *Proceedings of the 29th International Conference on Computational Linguistics*, 3597–3605. Retrieved June 22, 2023, from <https://aclanthology.org/2022.coling-1.318>
- Cao, S., Kitaev, N., & Klein, D. (2019). Multilingual alignment of contextual word representations. Retrieved October 11, 2023, from <https://openreview.net/forum?id=r1xCMYBtPS>
- Carroll, J. B. (1964). *Language and thought*. Prentice-Hall.
- Chang, T., Tu, Z., & Bergen, B. (2022). The geometry of multilingual language model representations. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 119–136. <https://doi.org/10.18653/v1/2022.emnlp-main.9>
- Chi, E. A., Hewitt, J., & Manning, C. D. (2020). Finding universal grammatical relations in multilingual BERT. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5564–5577. <https://doi.org/10.18653/v1/2020.acl-main.493>
- Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.-L., Huang, H., & Zhou, M. (2021). InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

- Human Language Technologies*, 3576–3588. <https://doi.org/10.18653/v1/2021.naacl-main.280>
- Chi, Z., Huang, S., Dong, L., Ma, S., Zheng, B., Singhal, S., Bajaj, P., Song, X., Mao, X.-L., Huang, H.-Y., & Wei, F. (2022). XLM-e: Cross-lingual language model pre-training via ELECTRA, 6170–6182. <https://doi.org/10.18653/v1/2022.acl-long.427>
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. <https://doi.org/10.3115/v1/D14-1179>
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124. <https://doi.org/10.1109/TIT.1956.1056813>
- Chomsky, N. (1957). *Syntactic structures*. De Gruyter Mouton. <https://doi.org/doi:10.1515/9783112316009>
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022, October 5). PaLM: Scaling language modeling with pathways. <https://doi.org/10.48550/arXiv.2204.02311>
- Chun, J., Han, N.-R., Hwang, J. D., & Choi, J. D. (2018). Building universal dependency treebanks in korean. Retrieved January 18, 2024, from <https://aclanthology.org/L18-1347>
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., ... Wei, J. (2022). Scaling instruction-finetuned language models.
- Church, K. W., & Gale, W. A. (1991). A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech & Language*, 5(1), 19–54. [https://doi.org/10.1016/0885-2308\(91\)90016-J](https://doi.org/10.1016/0885-2308(91)90016-J)
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. <https://doi.org/10.18653/v1/2020.acl-main.747>
- Conneau, A., & Lample, G. (2019, December). Cross-lingual language model pretraining. In *Proceedings of the 33rd international conference on neural information processing systems* (pp. 7059–7069). Curran Associates Inc.
- Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018, October). XNLI: Evaluating cross-lingual sentence representations. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2475–2485). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1269>

- Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type-token ratio (MATTR) [Publisher: Routledge _eprint: <https://doi.org/10.1080/09296171003643098>]. *Journal of Quantitative Linguistics*, 17(2), 94–100. <https://doi.org/10.1080/09296171003643098>
- Del, M., & Fishel, M. (2022, August 14). Similarity of sentence representations in multilingual LMs: Resolving conflicting literature and case study of baltic languages. <https://doi.org/10.48550/arXiv.2109.01207>
- Deshpande, A., Talukdar, P., & Narasimhan, K. (2022). When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3610–3623. <https://doi.org/10.18653/v1/2022.naacl-main.264>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Diaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018). Addressing age-related bias in sentiment analysis. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3173574.3173986>
- Doddapaneni, S., Ramesh, G., Khapra, M. M., Kunchukuttan, A., & Kumar, P. (2021, December 23). A primer on pretrained multilingual language models. <https://doi.org/10.48550/arXiv.2107.00676>
- Dufter, P., & Schütze, H. (2020). Identifying elements essential for BERT’s multilinguality. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4423–4437. <https://doi.org/10.18653/v1/2020.emnlp-main.358>
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211. [https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E)
- Etxaniz, J., Azkune, G., Soroa, A., de Lacalle, O. L., & Artetxe, M. (2023, August 2). Do multilingual language models think better in english? <https://doi.org/10.48550/arXiv.2308.01223>
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2021). Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).
- Field, A., Blodgett, S. L., Waseem, Z., & Tsvetkov, Y. (2021, August). A survey of race, racism, and anti-racism in NLP. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1905–1925). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.149>
- Gellerstam, M. (1986). Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1, 88–95.

- González, A. V., Barrett, M., Hvingelby, R., Webster, K., & Søgaard, A. (2020, November). Type b reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 2637–2648). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.209>
- Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language*, 15(4), 403–434.
- Goyal, N., Du, J., Ott, M., Anantharaman, G., & Conneau, A. (2021). Larger-scale transformers for multilingual masked language modeling. *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 29–33. <https://doi.org/10.18653/v1/2021.repl4nlp-1.4>
- Greenberg, J. H. (1963). Universals of language.
- Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique* [Google-Books-ID: odsrAAAAMAAJ]. Presses universitaires de France.
- Guo, W., & Caliskan, A. (2021). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122–133. <https://doi.org/10.1145/3461702.3462536>
- Hämmerl, K., Deiseroth, B., Schramowski, P., Libovický, J., Fraser, A., & Kersting, K. (2022, March 18). Do multilingual language models capture differing moral norms? Retrieved June 17, 2023, from <http://arxiv.org/abs/2203.09904>
- Hämmerl, K., Deiseroth, B., Schramowski, P., Libovický, J., Rothkopf, C. A., Fraser, A., & Kersting, K. (2023, June 1). Speaking multiple languages affects the moral bias of language models [version: 2]. <https://doi.org/10.48550/arXiv.2211.07733>
- Hao, Y., Dong, L., Wei, F., & Xu, K. (2019). Visualizing and understanding the effectiveness of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4143–4152. <https://doi.org/10.18653/v1/D19-1424>
- Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023, February 17). How good are GPT models at machine translation? a comprehensive evaluation. <https://doi.org/10.48550/arXiv.2302.09210>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Houlsby, N., Giurui, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019, September). Parameter-efficient transfer learning for NLP. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 2790–2799, Vol. 97). PMLR. <https://proceedings.mlr.press/v97/houlsby19a.html>
- Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8), e12432. <https://doi.org/10.1111/lnc3.12432>

- Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339. <https://doi.org/10.18653/v1/P18-1031>
- Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020, September 4). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. <https://doi.org/10.48550/arXiv.2003.11080>
- Huang, H., Tang, T., Zhang, D., Zhao, X., Song, T., Xia, Y., & Wei, F. (2023, December). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 12365–12394). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.826>
- Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities, 5491–5501. <https://doi.org/10.18653/v1/2020.acl-main.487>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023, October 10). Mistral 7b. <https://doi.org/10.48550/arXiv.2310.06825>
- Jiao, W., Wang, W., Huang, J.-t., Wang, X., & Tu, Z. (2023, March 19). Is ChatGPT a good translator? yes with GPT-4 as the engine. <https://doi.org/10.48550/arXiv.2301.08745>
- Johnson, W. (1944). I. a program of research. *Psychological Monographs*, 56(2), 1–15. <https://doi.org/10.1037/h0093508>
- Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.
- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing (3rd ed. draft)*. Pearson Education India.
- K, K., Wang, Z., Mayhew, S., & Roth, D. (2019). Cross-lingual ability of multilingual BERT: An empirical study. Retrieved October 6, 2023, from <https://openreview.net/forum?id=HJeT3yrtDr>
- Kaneko, M., Imankulova, A., Bollegala, D., & Okazaki, N. (2022). Gender bias in masked language models for multiple languages. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2740–2750. <https://doi.org/10.18653/v1/2022.naacl-main.197>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020, January 22). Scaling laws for neural language models. <https://doi.org/10.48550/arXiv.2001.08361>
- Kassner, N., Dufter, P., & Schütze, H. (2021). Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3250–3258. <https://doi.org/10.18653/v1/2021.eacl-main.284>

- Ke, Z., & Liu, B. (2023, May 11). Continual learning of natural language processing tasks: A survey. <https://doi.org/10.48550/arXiv.2211.12701>
- Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., Ahn, C., Yang, S., Lee, S., Park, H., Gim, G., Cha, M., Lee, H., & Kim, S. (2024, April 3). SOLAR 10.7b: Scaling large language models with simple yet effective depth up-scaling. <https://doi.org/10.48550/arXiv.2312.15166>
- Kim, I., Han, G., Ham, J., & Baek, W. (2021). Kogpt: Kakaobrain korean(hangul) generative pre-trained transformer. <https://github.com/kakaobrain/kogpt>
- Ko, H., Yang, K., Ryu, M., Choi, T., Yang, S., jiwung Hyun, & Park, S. (2023). A technical report for polyglot-ko: Open-source large-scale korean language models.
- Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213. https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html
- Koloski, B., Škrlj, B., Robnik-Šikonja, M., & Pollak, S. (2023, September 12). Measuring catastrophic forgetting in cross-lingual transfer paradigms: Exploring tuning strategies. <https://doi.org/10.48550/arXiv.2309.06089>
- Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the dark secrets of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4365–4374. <https://doi.org/10.18653/v1/D19-1445>
- L. Junbum. (2023a). Llama-2-ko-7b (revision 4a9993e). <https://doi.org/10.57967/hf/1098>
- L. Junbum. (2023b). Open-llama-2-ko-7b. <https://huggingface.co/beomi/open-llama-2-ko-7b>
- L. Junbum. (2024a). Solar-ko-10.7b. <https://huggingface.co/beomi/SOLAR-KO-10.7B>
- L. Junbum. (2024b). Yi-ko-6b (revision 205083a). <https://doi.org/10.57967/hf/1708>
- L. Junbum & Choi, T. (2023). Llama-2-koen-13b. <https://doi.org/10.57967/hf/1280>
- L. Junbum & Choi, T. (2024). Solar-koen-10.8b. <https://huggingface.co/beomi/SOLAR-KOEN-10.8B>
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations.
- Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020, May 1). From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. <https://doi.org/10.48550/arXiv.2005.00633>
- Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for french, 2479–2490. Retrieved October 29, 2023, from <https://aclanthology.org/2020.lrec-1.302>
- Lee, D.-G., & Rim, H.-C. (2005). Probabilistic models for korean morphological analysis. *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*. Retrieved May 2, 2024, from <https://aclanthology.org/I05-2034>

- Lee, M. (2022, September). *Kiwi, korean intelligent word identifier* (Version v0.14.0). GitHub. <https://doi.org/10.5281/zenodo.7041425>
- Lee, Y. (2000). Problems of korean-english translation: With respect to passive constructions. *Korean Association for Translation Studies*, 1, 47–76.
- Lester, B., Al-Rfou, R., & Constant, N. (2021, November). The power of scale for parameter-efficient prompt tuning. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 3045–3059). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.243>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
- Li, B., & Callison-Burch, C. (2023, May 23). This land is {your, my} land: Evaluating geopolitical biases in language models. <https://doi.org/10.48550/arXiv.2305.14610>
- Li, H. (2022). Language models: Past, present, and future. *Commun. ACM*, 65(7), 56–63. <https://doi.org/10.1145/3490443>
- Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.-H., Wu, W., ... Zhou, M. (2020, November). XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6008–6018). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.484>
- Libovický, J., Rosa, R., & Fraser, A. (2019, November 8). How language-neutral is multilingual BERT? <https://doi.org/10.48550/arXiv.1911.03310>
- Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., ... Li, X. (2022). Few-shot learning with multilingual generative language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9019–9052. Retrieved July 16, 2023, from <https://aclanthology.org/2022.emnlp-main.616>
- Liu, C.-L., Hsu, T.-Y., Chuang, Y.-S., & Lee, H.-Y. (2020, April 20). A study of cross-lingual ability and language-specific information in multilingual BERT. <https://doi.org/10.48550/arXiv.2004.09205>
- Liu, Q., McCarthy, D., Vulić, I., & Korhonen, A. (2019). Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 33–43. <https://doi.org/10.18653/v1/K19-1004>
- Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., & Tang, J. (2022, May). P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In S.

- Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 61–68). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-short.8>
- Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation [Place: Cambridge, MA Publisher: MIT Press]. *Transactions of the Association for Computational Linguistics*, 8, 726–742. https://doi.org/10.1162/tacl_a_00343
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.
- Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605. Retrieved May 13, 2024, from <http://jmlr.org/papers/v9/vandemaaten08a.html>
- MacDonald, D., & Carroll, S. E. (2018). Second-language processing of english mass-count nouns by native-speakers of korean [Number: 1 Publisher: Open Library of Humanities]. *Glossa: a journal of general linguistics*, 3(1). <https://doi.org/10.5334/gjgl.363>
- Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An introduction to information retrieval*. Cambridge university press.
- Markov, A. A. (1913). Essai d’une recherche statistique sur le texte du roman “eugene onegin” illustrant la liaison des epreuve en chain (‘example of a statistical investigation of the text of “eugene onegin” illustrating the dependence between samples in chain’). *Izvestia Imperatorskoi Akademii Nauk (Bulletin de l’Academie Imp ´eriale ´ des Sciences de St.-Petersbourg)* ´, 7, 153–162.
- Markov, A. A. (2006). An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19(4), 591–600. <https://doi.org/10.1017/S0269889706001074>
- Masoud, R. I., Liu, Z., Ferianc, M., Treleaven, P., & Rodrigues, M. (2023, August 25). Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions. <https://doi.org/10.48550/arXiv.2309.12342>
- McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld)* [Doctoral dissertation, The University of Memphis].
- McInnes, L., Healy, J., & Melville, J. (2020, September 17). UMAP: Uniform manifold approximation and projection for dimension reduction. <https://doi.org/10.48550/arXiv.1802.03426>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.
- Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2), 30:1–30:40. <https://doi.org/10.1145/3605943>

- Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z. X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2023). Crosslingual generalization through multitask finetuning. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15991–16111. <https://doi.org/10.18653/v1/2023.acl-long.891>
- Muller, B., Elazar, Y., Sagot, B., & Seddah, D. (2021). First align, then predict: Understanding the cross-lingual ability of multilingual BERT. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2214–2231. <https://doi.org/10.18653/v1/2021.eacl-main.189>
- Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2023, August 17). Contrasting linguistic patterns in human and LLM-generated text. <https://doi.org/10.48550/arXiv.2308.09067>
- Nadas, A. (1984). Estimation of probabilities in the language model of the ibm speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(4), 859–861. <https://doi.org/10.1109/TASSP.1984.1164378>
- Nadeem, M., Bethke, A., & Reddy, S. (2021, August). StereoSet: Measuring stereotypical bias in pretrained language models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 5356–5371). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.416>
- Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2023, November 15). Having beer after prayer? measuring cultural bias in large language models. Retrieved January 4, 2024, from <http://arxiv.org/abs/2305.14456>
- National Institute of Korean Language. (2021). NIKL dependency-parsed corpus (v.2.0). <https://kli.korean.go.kr/corpus>
- Nicholas, G., & Bhatia, A. (2023, June 12). Lost in translation: Large language models in non-english content analysis. <https://doi.org/10.48550/arXiv.2306.07377>
- Ousidhoum, N., Zhao, X., Fang, T., Song, Y., & Yeung, D.-Y. (2021, August). Probing toxic content in large pre-trained language models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 4262–4274). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.329>
- Papadimitriou, I., Lopez, K., & Jurafsky, D. (2023). Multilingual BERT has an accent: Evaluating english influences on fluency in multilingual models. *Findings of the Association for Computational Linguistics: EACL 2023*, 1194–1200. Retrieved June 17, 2023, from <https://aclanthology.org/2023.findings-eacl.89>
- Park, C. (2012). Statistical approach about ellipsis of korean and english -focused on ellipsis of subject and object-. *Journal of The Society of Korean Language and Literature*, 171–192.

- Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., ... Cho, K. (2021). Klue: Korean language understanding evaluation.
- Park, S., & Shin, H. (2021). Kr-sbert: A pre-trained korean-specific sentence-bert model.
- Park, T. (2004). *English article use by advanced korean efl learners* [Doctoral dissertation, Seoun National University].
- Patil, V., Talukdar, P., & Sarawagi, S. (2022). Overlap-based vocabulary generation improves cross-lingual transfer among related languages. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 219–233. <https://doi.org/10.18653/v1/2022.acl-long.18>
- Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. <https://doi.org/10.3115/v1/D14-1162>
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. <https://doi.org/10.18653/v1/N18-1202>
- Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2021). AdapterFusion: Non-destructive task composition for transfer learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 487–503. <https://doi.org/10.18653/v1/2021.eacl-main.39>
- Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., & Gurevych, I. (2020). AdapterHub: A framework for adapting transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 46–54. <https://doi.org/10.18653/v1/2020.emnlp-demos.7>
- Philippy, F., Guo, S., & Haddadan, S. (2023, May 26). Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. <https://doi.org/10.48550/arXiv.2305.16768>
- Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4996–5001. <https://doi.org/10.18653/v1/P19-1493>
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. *OpenAI*.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Radiya-Dixit, E., & Wang, X. (2020, 26–28 Aug). How fine can fine-tuning be? learning efficient language models. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the twenty third international conference on artificial intelligence and statistics* (pp. 2435–2443, Vol. 108). PMLR. <https://proceedings.mlr.press/v108/radiya-dixit20a.html>
- Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick,

- J., Cassirer, A., Powell, R., Driessche, G. v. d., Hendricks, L. A., Rauh, M., Huang, P.-S., ... Irving, G. (2022, January 21). Scaling language models: Methods, analysis & insights from training gopher. <https://doi.org/10.48550/arXiv.2112.11446>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Ramezani, A., & Xu, Y. (2023, June 2). Knowledge of cultural moral norms in large language models. <https://doi.org/10.48550/arXiv.2306.01857>
- Ranathunga, S., & de Silva, N. (2022). Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 823–848. Retrieved June 21, 2023, from <https://aclanthology.org/2022.aacl-main.62>
- Ravishankar, V., & Nivre, J. (2022). The effects of corpus choice and morphosyntax on multilingual space induction. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4130–4139. <https://doi.org/10.18653/v1/2022.findings-emnlp.304>
- Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. <https://arxiv.org/abs/2004.09813>
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works [Place: Cambridge, MA Publisher: MIT Press]. *Transactions of the Association for Computational Linguistics*, 8, 842–866. https://doi.org/10.1162/tacl_a_00349
- Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., & Johnson, M. (2021). XTREME-r: Towards more challenging and nuanced multilingual evaluation, 10215–10245. <https://doi.org/10.18653/v1/2021.emnlp-main.802>
- Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1), 569–630. <https://doi.org/10.1613/jair.1.11640>
- Samuel, D., Kutuzov, A., Touileb, S., Velldal, E., Øvrelid, L., Rønningstad, E., Sigdel, E., & Palatkina, A. (2023, May). NorBench –a benchmark for norwegian language models. In T. Alumäe & M. Fishel (Eds.), *Proceedings of the 24th nordic conference on computational linguistics (NoDaLiDa)* (pp. 618–633). University of Tartu Library. Retrieved March 31, 2024, from <https://aclanthology.org/2023.nodalida-1.61>
- Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., ... Rush, A. M. (2022). Multitask prompted training enables zero-shot task generalization.

- International Conference on Learning Representations*. <https://openreview.net/forum?id=9Vrb9D0WI4>
- Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Workshop, B., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., ... Wolf, T. (2023, June 27). BLOOM: A 176b-parameter open-access multilingual language model. <https://doi.org/10.48550/arXiv.2211.05100>
- Shanahan, M. (2023). Talking about large language models.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1), 50–64. <https://doi.org/10.1002/j.1538-7305.1951.tb01366.x>
- Shen, L. (2021). Measuring political media slant using text data. <https://www.lucasshen.com/research/media.pdf>
- Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D., & Wei, J. (2022, October 6). Language models are multilingual chain-of-thought reasoners. <https://doi.org/10.48550/arXiv.2210.03057>
- Shliazhko, O., Fenogenova, A., Tikhonova, M., Mikhailov, V., Kozlova, A., & Shavrina, T. (2022, April 15). mGPT: Few-shot learners go multilingual. <https://doi.org/10.48550/arXiv.2204.07580>
- Singh, J., McCann, B., Socher, R., & Xiong, C. (2019). BERT is not an interlingua and the bias of tokenization, 47–55. <https://doi.org/10.18653/v1/D19-6106>
- SKT-AI. (2020). Kogpt2 ver 2.0. <https://github.com/SKT-AI/KoGPT2>
- SKT-AI. (2021). Ko-gpt-trinity 1.2b (v0.5). <https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5>
- Soltan, S., Ananthakrishnan, S., FitzGerald, J., Gupta, R., Hamza, W., Khan, H., Peris, C., Rawls, S., Rosenbaum, A., Rumshisky, A., Prakash, C. S., Sridhar, M., Triefenbach, F., Verma, A., Tur, G., & Natarajan, P. (2022, August 3). AlexaTM 20b: Few-shot learning using a large-scale multilingual seq2seq model. <https://doi.org/10.48550/arXiv.2208.01448>
- Son, S., Park, C., Lee, J., Shim, M., Lee, C., Jang, Y., Seo, J., & Lim, H. (2022, September 14). Language chameleon: Transformation analysis between languages using cross-lingual post-training based on pre-trained language models. <https://doi.org/10.48550/arXiv.2209.06422>
- Steinborn, V., Dufter, P., Jabbar, H., & Schuetze, H. (2022). An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models. *Findings of the Association for Computational Linguistics: NAACL 2022*, 921–932. <https://doi.org/10.18653/v1/2022.findings-naacl.69>
- Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Wei, J., Wang, X., Chung, H. W., Bahri, D., Schuster, T., Zheng, S., et al. (2022). Ul2: Unifying language learning paradigms. *The Eleventh International Conference on Learning Representations*.

- Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., ... Le, Q. (2022, February 10). LaMDA: Language models for dialog applications. <https://doi.org/10.48550/arXiv.2201.08239>
- Torruella, J., & Capsada, R. (2013). Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, 95, 447–454. <https://doi.org/10.1016/j.sbspro.2013.10.668>
- Touileb, S., Øvrelid, L., & Velldal, E. (2022). Occupational biases in norwegian and multilingual language models. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 200–211. <https://doi.org/10.18653/v1/2022.gebnlp-1.21>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023, February 27). LLaMA: Open and efficient foundation language models. <https://doi.org/10.48550/arXiv.2302.13971>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023, July 19). Llama 2: Open foundation and fine-tuned chat models. <https://doi.org/10.48550/arXiv.2307.09288>
- Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.
- Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T.-H. 'K., & Wilson, S. (2023, August 8). Unmasking nationality bias: A study of human perception of nationalities in AI-generated articles. <https://doi.org/10.48550/arXiv.2308.04346>
- Venkit, P. N., Srinath, M., & Wilson, S. (2022, October). A study of implicit bias in pretrained language models against people with disabilities. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 1324–1332). International Committee on Computational Linguistics. Retrieved November 2, 2023, from <https://aclanthology.org/2022.coling-1.113>
- Ventura, M., Ben-David, E., Korhonen, A., & Reichart, R. (2023, October 3). Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models [version: 1]. <https://doi.org/10.48550/arXiv.2310.01929>
- Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019, December 15). Multilingual is not enough: BERT for finnish. <https://doi.org/10.48550/arXiv.1912.07076>

- Vries, W. d., Wieling, M., & Nissim, M. (2022). Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages, 7676–7685. <https://doi.org/10.18653/v1/2022.acl-long.529>
- Wang*, Z., Xie*, J., Xu, R., Yang, Y., Neubig, G., & Carbonell, J. G. (2019). Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. Retrieved October 11, 2023, from <https://openreview.net/forum?id=S1l-C0NtwS>
- Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. *International Conference on Learning Representations*. <https://openreview.net/forum?id=gEZrGCozdqR>
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022, October 26). Emergent abilities of large language models. <https://doi.org/10.48550/arXiv.2206.07682>
- Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024, February 24). Do llamas work in english? on the latent language of multilingual transformers. Retrieved March 17, 2024, from <http://arxiv.org/abs/2402.10588>
- Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4003–4012. <https://aclanthology.org/2020.lrec-1.494>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- Wu, S., & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 833–844. <https://doi.org/10.18653/v1/D19-1077>
- Xia, F., Han, C., Palmer, M., & Joshi, A. (2000). Comparing lexicalized treebank grammars extracted from chinese, korean, and english corpora. *Second Chinese Language Processing Workshop*, 52–59. <https://doi.org/10.3115/1117769.1117778>
- Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10, 291–306. https://doi.org/10.1162/tacl_a_00461
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). MT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter*

- of the Association for Computational Linguistics: Human Language Technologies, 483–498. <https://doi.org/10.18653/v1/2021.naacl-main.41>
- Yang, Y., Zhang, Y., Tar, C., & Baldridge, J. (2019, November). PAWS-x: A cross-lingual adversarial dataset for paraphrase identification. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3687–3692). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1382>
- Yin, D., Bansal, H., Monajatipoor, M., Li, L. H., & Chang, K.-W. (2022). GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2039–2055. Retrieved June 17, 2023, from <https://aclanthology.org/2022.emnlp-main.132>
- Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023, December). Don’t trust ChatGPT when your question is not in english: A study of multilingual abilities and types of LLMs. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 7915–7927). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.491>
- Zhao, J., Mukherjee, S., Hosseini, S., Chang, K.-W., & Hassan Awadallah, A. (2020, July). Gender bias in multilingual embeddings and cross-lingual transfer. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2896–2907). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.260>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023, June 29). A survey of large language models. <https://doi.org/10.48550/arXiv.2303.18223>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023, December 23). Judging LLM-as-a-judge with MT-bench and chatbot arena. <https://doi.org/10.48550/arXiv.2306.05685>
- Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2023, May 1). Multilingual machine translation with large language models: Empirical results and analysis. <https://doi.org/10.48550/arXiv.2304.04675>