KoCommonGEN v2: A Benchmark for Navigating Korean Commonsense Reasoning Challenges in Large Language Models

Jaehyung Seo[1], **Jaewook Lee**[1], **Chanjun Park**[2], **Seongtae Hong**[1]
Seungjun Lee[1] **and Heuiseok Lim**[1†]
[1]Department of Computer Science and Engineering, Korea University
[2]Upstage
[1]{seojae777,jaewook133,ghdchlwls123,dzzy6505,limhseok}@korea.ac.kr
[2]chanjun.park@upstage.ai

Abstract

The evolution of large language models (LLMs) has culminated in a multitask model paradigm where prompts drive the generation of user-specific outputs. However, this advancement has revealed a challenge: LLMs frequently produce outputs against socially acceptable commonsense standards in various scenarios. To address this gap in commonsense reasoning, we present KoCommonGEN v2, a fine-grained benchmark dataset focused on Korean commonsense reasoning. This dataset, enriched with human annotations, comprises multiple-choice questions across seven error categories. These categories include commonsense memorization, numerical commonsense, toxic speech, and more, which are vulnerable to undermining the reliability of LLMs' commonsense reasoning capabilities. The empirical results present that LLMs struggle with Korean commonsense reasoning. With human accuracy benchmarked at approximately 85%, GPT-4's performance lags at about 74%, and other LLMs demonstrate an average accuracy of around 42%. Our findings emphasize the need for targeted improvements in Korean commonsense reasoning within LLMs, paving the way for more socially and contextually sensitive AI models. KoCommonGEN v2 is one of the benchmark datasets for the Open Ko-LLM Leaderboard.

1 Introduction

In the field of natural language processing (NLP) research, the significance of benchmark datasets has grown notably with the advent of Transformer-based language models (Radford et al., 2018; Devlin et al., 2019) and the feasibility of pre-training on large corpora. Language models have evolved to engage in multitask learning, emphasizing the need for overall linguistic skills and language understanding (Wang et al., 2018, 2019). In particular, attaining performance akin to human cognition remains a long-standing challenge, and language models are striving to grasp commonsense reasoning. This pursuit has led to the introduction of prominent commonsense benchmarks, such as CommonsenseQA (Talmor et al., 2019), CommonGEN (Lin et al., 2020b), PiQA (Bisk et al., 2020), and SWAG (Zellers et al., 2018). However, the efficacy of these commonsense benchmark datasets has begun to diminish with the innovative improvements of large language models (LLMs), such as GPT-3 (Brown et al., 2020), FLAN (Wei et al., 2021), InstructGPT (Ouyang et al., 2022), LLaMA2 (Touvron et al., 2023b), Mistral (Jiang et al., 2023), and GPT-4 (OpenAI, 2023). LLMs can outperform upper-bound even with few-shot prompt-based tuning or zero-shot settings.

However, as a wide array of tasks and scenarios are consolidated into a single LLM, the new concern is the heightened risk of these models generating outputs that violate commonsense in social and cultural contexts. Recently, there has been a notable incident referred to as '*the MacBook throwing incident of King Sejong (1397-1450)*[1]' in the Korean community where the media negatively described LLMs as lying to make hallucinations even in commonsense knowledge.

This incident shows the increasing need to recognize the complexity of commonsense required from LLMs based on sociocultural differences and to adopt broad-ranging approaches in commonsense reasoning that extend beyond inherent everyday life or elementary school-level knowledge. It also demonstrates the importance of including commonsense knowledge shared among speakers of the same language community. Therefore, we contend that evaluating LLMs for commonsense reasoning, if solely reliant on universally applicable commonsense knowledge and English translated without considering sociocultural knowledge, along with

† Corresponding author

[1] https://english.hani.co.kr/arti/english_edition/e_international/1095956

approaches limited in scope to explaining commonsense, are grounded in outdated concerns.

To address the challenges posed by the new generation of LLMs, we introduce **KoCommonGEN v2**, a dataset aimed at fine-grained evaluation of Korean commonsense reasoning. We adopt the Korean CommonGEN (Seo et al., 2022)[2] design and enhance outdated settings for evaluating the performance of LLMs. Firstly, we reconstruct the dataset from scratch to associate more closely with Korean commonsense knowledge, moving away from depending on the translated sources and universal commonsense. Secondly, we expand the scope beyond elementary-level commonsense reasoning. This benchmark dataset includes seven error types to precisely identify where LLMs are missing commonsense reasoning. Thirdly, we alter from a natural language generation-based evaluation to a multiple-choice task format. This change standardizes the evaluation process, ensuring that language models are tested on their reasoning ability to produce results aligned with Korean commonsense based on unified instructions (Sai et al., 2022). Our contributions are as follows:

(i) **Introduction to KoCommonGEN v2**: We introduce a fine-grained version of a benchmark for Korean commonsense reasoning. This benchmark dataset is designed to evaluate LLMs' challenge with socioculturally sensitive criteria, addressing the limitations of existing benchmarks.

(ii) **Classification of seven error types**: We extend the scope to include seven error categories. These categories are designed to probe areas where LLMs struggle, including toxicity, commonsense memorization, and numerical inaccuracies.

(iii) **Comparative analysis of LLM performance**: Our comprehensive analysis focuses on the performance of LLMs, both open-source and commercial, in Korean commonsense reasoning. Our findings reveal that these models have not yet reached human proficiency in understanding and applying Korean commonsense, highlighting the weak error types that require further improvement.

(iv) **Commonsense in multilingualism**: Our research examines the effects of language changes on the same Korean commonsense knowledge. We investigate how LLMs' understanding of commonsense can vary when dealing with different languages and code-switching scenarios.

The following task involves combining nouns and verbs from the concept set to create a sentence that is consistent with commonsense. Choose the option that contains the most logically valid sentence among the four examples created by combining nouns and verbs from the concept set.

Concept set: { I#moral#give#topic#lecture }

1. I moral topic give a lecture.
2. I did not give a lecture on a moral topic.
3. I lecture gave because of a moral topic.
4. Moral topic makes me lecture.

Answer: 2. I did not give a lecture on a moral topic.

Table 1: An example of unified instruction and multiple-choice question answering in the plausibility type of **KoCommonGEN v2**. This figure is translated into English for the convenience of non-Korean speakers.

2 KoCommonGEN v2

As illustrated in Table 1, KoCommonGEN v2 is comprised of multiple-choice question answering. Based on the instructions of the task prompt (blue box) and the concept set (yellow box), the model is required to correctly infer the relations among concepts and select the most commonsense-acceptable answer choice from four choices.

2.1 Task Definition

CommonGEN (Lin et al., 2020b) and Korean CommonGEN (Seo et al., 2022) are recognized as generative commonsense reasoning tasks. These tasks utilize given sets of concepts to assess the capability of generating plausible sentences. Drawing inspiration from these benchmark datasets, KoCommonGEN v2 introduces a task focused on interpreting task prompts and selecting an answer choice that effectively transforms an input concept set $\mathcal{C} = c_1, \ldots, c_n$ into a target sentence $\mathcal{T}$. This transformation is guided by commonsense reasoning applied within the bounds of $\mathcal{C}$, aiming to avoid violations against commonsense.

2.2 Error Categorization

We define seven types of errors that can impact the reliability of LLMs' commonsense reasoning. Each type is constituted to navigate shared societal norms, expectations, and cultural contexts, along with commonsense knowledge. (i) **Commonsense Distortion**: Evaluates the causal validity and temporal suitability of historical events considered commonsense by Korean community (Yin et al., 2022; Keleg and Magdy, 2023). (ii) **Commonsense Memorization**: The frequency of co-occurrence during pre-training significantly influences com-

[2] For details about this dataset, refer to the Appendix M

monsense knowledge. This reveals the potential risk of given concepts forming expressions that do not align with commonsense based solely on superficial information (Tirumala et al., 2022; Du et al., 2023). (iii) **Toxic Speech**: The ambiguous boundary between sociocultural commonsense and prejudice poses a risk of leading to aggressive expressions directed at specific groups or individuals. (Sap et al., 2020; Bauer et al., 2023). (iv) **Grammaticality**: Evaluates the task of constructing sentences that comply with Korean morphological rules according to compositional generalization, considering the generative commonsense reasoning task with a given concept set (Keysers et al., 2019; Lin et al., 2020b; Seo et al., 2022). (v) **Plausibility**: Encourages forming the make-sense sentence through implicit and fuzzy reasoning, rather than retrieving knowledge (Bhagavatula et al., 2019; Liu et al., 2023). (vi) **Numerical Commonsense**: Measures the inclusion of hallucinations or the occurrence of recall failure regarding universal or sociocultural numerical commonsense (Lin et al., 2020a; Liu et al., 2022). (vii) **Proverb**: Assesses the ability to understand elusive and idiomatic meanings based on non-compositional concepts. (Zeng et al., 2023). Each multiple-choice question embodies the error corresponding to one of the seven commonsense error categorizations.

Commonsense Distortion The commonsense distortion type determines whether models can accurately interpret content that aligns with established events recognized in commonsense knowledge. In communities sharing the same language and sociocultural commonsense knowledge, there is an ongoing expectation for language models to precisely understand historical causality or chronological sequences that are regarded as commonsense. Commonsense distortions can significantly undermine confidence in language models, emphasizing the importance of this evaluation category. For example, for the given concept set: {*King Sejong#Hangul#Create*}, a correct answer is "*King Sejong created Hangul for his people*," but incorrect candidates are {*"King Sejong did not create Hangul"*, *"Hangul created King Sejong"*, *"Hangul was created before King Sejong"*}.

Commonsense Memorization The commonsense memorization type focuses on LLMs' ability to overcome inherent hallucinations in the given concepts and to derive reliable conclusions through commonsense reasoning. Commonsense knowledge cannot be entirely represented through given prompts or seen data. Hence, it is imperative for models to demonstrate the capacity to transcend the co-occurrence hallucination, employing commonsense reasoning that goes beyond the provided information. For instance, with the given concept set: {*Cold#Recover#Cough*}, a correct choice "*Do coughing make hard to recover from a cold*" includes the unseen concept ***hard*** based on the relational reasoning, thus navigating beyond the limitations of the commonsense memorization. However, incorrect candidates {*"Do Coughing can recover from a cold"*, *"A cold recovers if you do coughing"*, *"Coughing hards to recover as it may lead to tuberculosis"*} present superficial interpretations of the given concepts.

Toxic Speech The toxic speech type examines whether models trained on large corpora mitigate the risks of biased or hateful expressions they may encompass. There is an increased risk of toxic speech disguised as commonsense on corpora. To illustrate, with the given concept set: {*Age#Woman#Man#Difference*}, one incorrect answer "*There should be an age difference between men and women*" implies a biased perspective on age-appropriate relationships between genders, suggesting that men should inherently be older than women, or vice versa. This statement is anchored in stereotypical beliefs rather than derives from commonsense reasoning.

Grammaticality The grammaticality type evaluates whether models can comprehend grammatical nuances resulting from the Korean language's system of combining roots and affixes. This represents a typical issue in multilingual models that have been trained on corpora translated into Korean. An example of this is the 'ㅅ' irregular conjugation phenomenon, where the 'ㅅ' (final consonant) should be omitted before a vowel-starting ending or a linking vowel. However, GPT-3.5 (Brown et al., 2020) often generates awkward sentences by misapplying regular conjugation rules in contexts that require irregular conjugation. GPT-3.5 usually includes an error where 'ㅅ' is not correctly dropped, indicating a regular conjugation instead of the correct irregular change. To demonstrate this issue with an English analogy, consider the concept set: {*Grandma#Cauldron#Take#Pour#Water*}. An incorrect choice is *Grandma took the water and* *pourred X (→ poured O) it into the cauldron*, showing a grammatical error analogous to the incorrect application of conjugation rules in Korean.

Plausibility The plausibility type estimates whether the model can discern and prefer sentences that appear sufficiently plausible to Korean speakers. Machine translation, data augmentation, and informal expressions prevalent in online sources included in training corpora can lead to awkward or unnatural outcomes. For example, with the given `concept set`: {*I#Moral#Give#Topic#Lecture*}, a correct answer is "*I did not give a lecture on a moral topic*", even in considering the issue of models reducing their preference for negations in commonsense (Chen et al., 2023). However, incorrect candidates {*"I moral topic to give a lecture", "I lecture gave because of a moral topic", "Moral topic makes me lecture"*} exhibit unnecessary repetition of concepts, unnatural word order in Korean, and missing concept.

Numerical Commonsense The numerical commonsense type encompasses subjects such as science, math, history, and society, aligned with the level of compulsory education in Korea. To evaluate whether models can grasp numeric relations between given concepts, we provide the concept set to ensure that the model can sufficiently infer numeric commonsense knowledge based on the given information. For instance, with the given `concept set`: {*South Korea#Government#Liberation Day#Designate*}, incorrect choices "*The South Korea government has designated July 17th/March 1st/May 18th (→ August 15th) as Liberation Day*" represent the confusion with dates of other significant national holidays. Such misrepresentations or inaccuracies in numerical commonsense clearly reveal the presence of correction, highlighting the importance of numerical commonsense understanding in models.

Proverb The proverb type discerns whether models can accurately interpret and convey the meanings of metaphorical expressions that are prevalent in Korean culture. Models should grasp and express the underlying meanings of metaphorical expressions deeply rooted in Korean commonsense. As an example, given `concept set`: {*Whale#Fight#Shrimp#Back#Burst*}, a correct answer "*When a shrimp's back bursts, it's a whale fight*" metaphorically means that a bystander can suffer harm from a conflict between powerful parties. Misunderstanding these proverbs without considering the sociocultural context can lead to awkward or illogical interpretations.

2.3 Data Collection

We crafted a benchmark for Korean commonsense reasoning by sourcing commonsense sentences from diverse resources, including Korean CommonGEN (Seo et al., 2022), AI-Hub Korean commonsense dataset[3], and Korean Wikipedia[4]. Our approach involves extracting noun and verb concepts from these sentences, leveraging human-annotated concept extraction results in the AI-Hub Korean commonsense datasets. We used these human-annotated concept sets as example samples to prompt the extraction of appropriate verbs and nouns from given sentences. The concept extraction process begins with GPT-4 (OpenAI, 2023) and is further refined by the authors. We conducted corrections for unintended concept omissions, incorrect tagging of verbs or nouns, and the inclusion of concepts not present in the sentences. Furthermore, if there was an overlap with the human-annotated concept set from the original source, we randomly rearranged the order of concepts and applied synonym replacements for some concepts within the sentences. We assembled 847 multiple-choice questions, each representing one of the seven types of Korean commonsense reasoning challenges. The number of samples for each commonsense error type in KoCommonGEN v2 is shown in Table 2.

Error Types	# Samples
Commonsense Distortion	100
Commonsense Memorization	100
Toxic Speech	98
Grammaticality	211
Plausibility	183
Numerical Commonsense	99
Proverb	56
Total	847

Table 2: Number of samples in each category

2.4 Answer Creation

KoCommonGEN v2 is structured as a multiple-choice question format consisting of three incorrect answers and one correct answer. At least one of the three incorrect answers includes an error type from each category. The correct answer contains all

[3] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71311

[4] https://ko.wikipedia.org/wiki

provided concept information[5] and adheres most logically to commonsense. All incorrect answers are crafted by the authors, modifying the correct answer according to the guidelines below to apply specific types of errors:

1. Errors involving the omission of concept information are accompanied by appropriate paraphrasing of the sentence (e.g., omitting concept information can distort historical facts).
2. In cases where errors can be applied through causal and sequential relationships, the concept information is utilized to its fullest.
3. For the numerical commonsense type, construct incorrect answer choices by replacing only the numerical information.
4. Incorrect answers for toxic speech are crafted to include one of the following: social bias, profanity, sarcasm, or aggressive expressions toward specific targets.
5. When considering linguistic characteristics, errors are based on one of the following: translational style, redundancy, morphological combinations, or sentence order mistakes.

To ensure an equitable distribution among all possible answers, each correct answer is evenly distributed from 1 to 4 with a 25% occurrence.

3 Experimental Settings

We introduce models used in our experiment, prompts, n-shot examples, evaluation metrics, and human evaluation. More details are in Appendix A.

Models To enhance the robustness and diversity of our experiments, we select models that have recorded high performance and are widely used as backbones across various models in the Open-LLM leaderboard. Among these, we determine models that have undergone additional training based on Korean or are frequently used within the Korean community of users as our experimental subjects. We also consider models with a capacity of up to 13B parameters and detailed information publicly disclosed. For our experiments, we employ language models pre-trained in Korean, including **KoGPT2**[6](Radford et al., 2019), and **Polyglot-ko** (Ko et al., 2023). For models based on English, we use **OPT** (Zhang et al., 2022), **LLaMA** (Touvron et al., 2023a), **LLaMA2** (Touvron et al., 2023b), and **Mistral** (Jiang et al., 2023). For a model based on English-Chinese, we employ **QWEN** (Bai et al., 2023). Additionally, we utilize **LLaMA-ko-en**[7] and **LLaMA2-ko-en**[8], adaptation of the LLaMA/LLaMA2 model with an expanded vocabulary, and further pre-training with additional Korean and English corpora. Models based on Polyglot-ko with **instruction tuning** applied include **KULLM**[9] and **KoAlpaca**[10], while for **LLaMA2-ko-en**, we use a model with **instruction tuning** and **DPO** (Rafailov et al., 2023) applied. For commercial APIs, we apply **GPT-3.5** (Brown et al., 2020), **GPT-4** (OpenAI, 2023), and **HyperCLOVA** (Kim et al., 2021).

Prompt We feed the LLMs prompts as shown in Table 1 and Appendix I. The prompt describes the generative commonsense reasoning task and instructs to generate the correct answer number and sentence according to multiple-choice questions.

n-shot Examples To prevent overstated results for specific examples in commonsense reasoning, we initially adopt a 0-shot setting as the baseline framework. Consequently, analyzing performance variations in KoCommonGEN v2 becomes necessary according to the impact of adding knowledge sources and categories. We create a few-shot development set consisting of representative examples from each category and conduct additional experiments in 2-shot, 5-shot, and 10-shot settings. The 10-shot includes one example from each type, with numerical commonsense divided into four parts according to the subject. The 5-shot uses examples excluding the types with the lowest performance in the 0-shot experiments, specifically numerical commonsense and proverb categories. For the 2-shot, we select examples from the categories that rank within the top two regarding the number of samples, namely plausibility and grammaticality.

Evaluation Metric We adopt utilizing generation probabilities as the evaluation method in lm-evaluation-harness. We denote $x_{0:p}$ as the prompt and $x_{p:s}$ as the continuation sentence with a token

[5] We allow morphological changes according to the attachment of functional morphemes to the given concepts.

[6] https://github.com/SKT-AI/KoGPT2

[7] https://huggingface.co/beomi/kollama-13b

[8] https://huggingface.co/beomi/llama-2-koen-13b

[9] https://github.com/nlpai-lab/KULLM

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

length of $|s - p|$. The continuation sentence comprises an 'answer number + sentence'. We calculate the log probability for every candidate sentence, and the answer is chosen based on the option with the highest probability. The probability for each candidate sentence is calculated as follows:

$$\sum_{j=p}^{s} \log \mathbb{P}(x_j|x_{0:j})/(|s-p|) \tag{1}$$

This approach attempts to normalize for length by computing the average log probability per token.

Human Evaluation To measure human-level performance on our benchmark, we engaged native Korean-speaking volunteers as evaluators, compensating them at a rate of $0.8 per question. We conducted the test with 22 native Korean speakers using the web-based test referenced in Appendix L. To ensure an even distribution of the total 847 questions across different types, each volunteer is assigned to answer 45 to 55 multiple-choice questions. As shown in Table 3, The average score for the total is 83.95. The evaluation performed by two volunteers per sample shows a strong positive correlation, as evidenced by Cohen's kappa (Cohen, 1960) is 0.7693 and Krippendorff's alpha (Krippendorff, 2011) value of 0.77. These values are indicators of high inter-annotator reliability.

Human Performance	Total benchmark samples (1-17)
Average score	83.95
Cohen's kappa	0.7693
Krippendorff's alpha	0.7706

Table 3: Human performance in KoCommonGEN v2

4 Experimental Results

***n*-shot Accuracy** We compare the performance of LLMs in n-shot settings. As described in Table 4, the average accuracy of the language models used in the experiment is 42.13%. In the 5-shot setting, the model with instruction tuning applied to LLaMA2-ko-en shows the highest performance at 62.22%, while OPT 6.7B, which lacks training in Korean, exhibits the lowest at 24.79%. The accuracy difference between the highest 0-shot and the lowest 10-shot is just 2.02%. These outcomes imply that significant performance disparities exist among models; however, the differences in performance across n-shot settings are not pronounced. An increase in n-shots does not necessarily guarantee improved performance; certain models exhibit decreased performance with more shots. These results also demonstrate that KoCommonGEN v2 is not excessively overstated by specific knowledge sources or adaptation to categories and prove its robustness against performance fluctuations from few-shot samples. Furthermore, models enhanced with instruction tuning or DPO applied to LLaMA2-ko-en do not consistently surpass the performance of their backbone models. We also observe that QWEN 7B and Mistral 7B, which do not heavily incorporate Korean, outperform Korean-based LLMs. This shows the need for advancements in training approaches for Korean-based LLMs.

Model	0-shot	2-shot	5-shot	10-shot
KoGPT2	0.5455	0.5419	0.5336	0.5325
Polyglot-ko 5.8B	0.4215	0.3967	0.4203	0.4132
Polyglot-ko 12.8B	0.4510	0.3943	0.3400	0.3259
LLaMA-ko-en 7B	0.4451	0.4026	0.3872	0.3943
LLaMA-ko-en 13B	0.4357	0.3991	0.3743	0.3778
LLaMA2-ko-en 13B	0.5266	0.5525	0.5360	0.5159
OPT 6.7B	0.3046	0.2704	0.2621	0.2715
OPT 13B	0.3554	0.2774	0.2479	0.2680
QWEN 7B	0.4841	0.4416	0.4900	0.4994
Mistral 7B	0.5561	0.5407	0.6068	0.5643
LLaMA 7B	0.3802	0.3790	0.3636	0.3695
LLaMA 13B	0.4191	0.3483	0.3235	0.3306
LLaMA2 13B	0.4581	0.3849	0.3672	0.3601
KULLM 13B	0.3754	0.4050	0.4109	0.3754
KoAlpaca 5.8B	0.3908	0.3518	0.3613	0.3447
KoAlpaca 13B	0.3932	0.3200	0.3270	0.3117
LLaMA2-ko-en 13B+INST	0.5289	**0.5974**	**0.6222**	**0.5880**
LLaMA2-ko-en 13B+INST+DPO	**0.5584**	0.5431	0.5325	0.5065

Table 4: Comparative performance of LLMs in 0, 2, 5, and 10-shot settings: INST indicates the model with instruction tuning, and DPO represents the model with direct preference optimization applied. The highest performances in each shot setting are **bolded**.

Model Size We analyze performance differences based on model size from 5.8B, 7B to 13B parameters in the n-shot setting. Table 4 shows that a larger model size does not necessarily guarantee better performance. Limitations in computational resources leading to uneven amounts of training data (Ko et al., 2023), empirical-dependent hyperparameter settings, and potential violations and toxicity increasing with model size (Touvron et al., 2023a) contribute to these results.

Error Type Analysis Figure 1 shows the average n-shot performance of models for each type. The commonsense memorization type consistently scores the highest, with an average of approximately 50.44%. The numerical commonsense type presents the lowest scores, averaging around

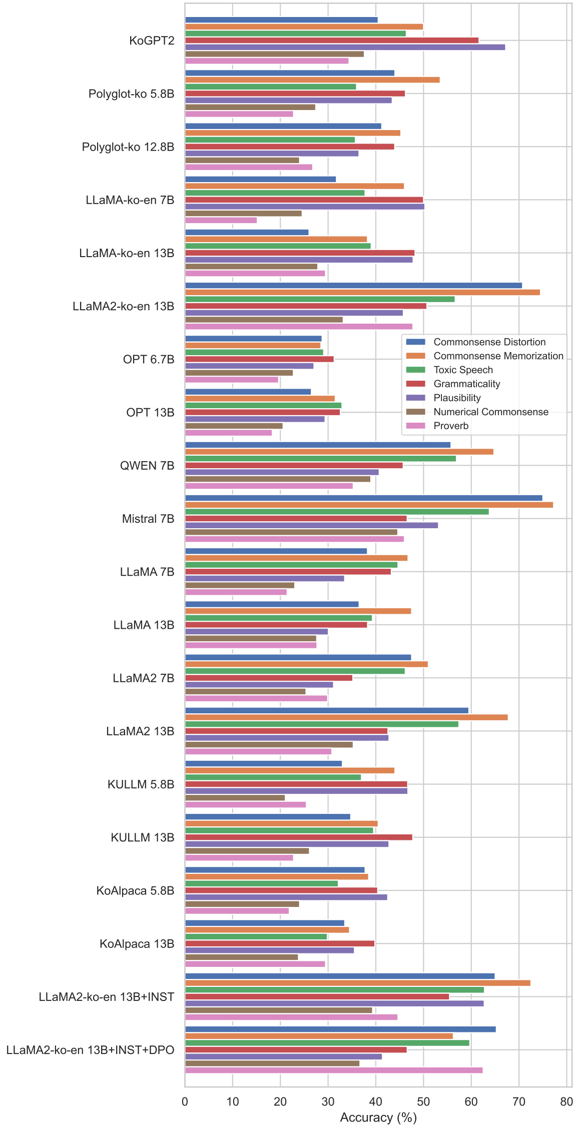
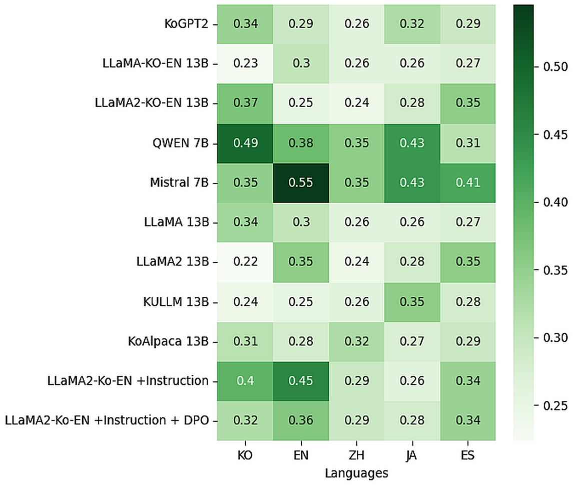


Figure 1: Comparative analysis of average n-shot performance by seven categories. For detailed results of each n-shot setting, please refer to Appendix G



	KO	EN	ZH	JA	ES
KoGPT2	0.34	0.29	0.26	0.32	0.29
LLaMA-KO-EN 13B	0.23	0.3	0.26	0.26	0.27
LLaMA2-KO-EN 13B	0.37	0.25	0.24	0.28	0.35
QWEN 7B	0.49	0.38	0.35	0.43	0.31
Mistral 7B	0.35	0.55	0.35	0.43	0.41
LLaMA 13B	0.34	0.3	0.26	0.26	0.27
LLaMA2 13B	0.22	0.35	0.24	0.28	0.35
KULLM 13B	0.24	0.25	0.26	0.35	0.28
KoAlpaca 13B	0.31	0.28	0.32	0.27	0.29
LLaMA2-Ko-EN +Instruction	0.4	0.45	0.29	0.26	0.34
LLaMA2-Ko-EN +Instruction + DPO	0.32	0.36	0.29	0.28	0.34

Figure 2: Performance for the understanding of Korean commonsense translated into other languages

29.19%. The performance difference in each type varies significantly across models. Korean-based LLMs tend to perform relatively better in types closely aligned with the linguistic intricacies of Korean, such as grammaticality and plausibility types. QWEN, Mistral, and LLaMA2-ko-en demonstrate robust capabilities in addressing commonsense distortion, memorization, and toxic speech. Models using LLaMA2-ko-en as their backbone demonstrate over 15% superior performance in understanding metaphorical expressions compared to other models. We concentrate on struggles for most models in distortion, toxic speech, and proverb type, where accuracy often falls below 45%. These results show the limitation of relying on mimicking acquired knowledge and expressing it based on superficial similarities rather than considering social interaction and cultural context.

Commonsense in Multilingual Contexts We evaluate the ability of models to maintain their judgment capabilities for the same commonsense questions when instructions and multiple-choice question answers, originally written in Korean (ko), are translated into English (en), Chinese (zh), Japanese (ja), and Spanish (es)[11]. We use the numerical commonsense type for this evaluation, as its clearly defined numerical values allow for consistent translations across different languages[12]. Figure 2 presents the 0-shot performance of models in multilingual settings on numerical commonsense reasoning. The multilingual pre-trained language model QWEN shows the highest performance for Korean, Chinese, and Japanese, while Mistral excels in English, Chinese, Japanese, and Spanish. Most models tend to exhibit high performance in languages that occupy a significant portion of their training sources. However, some models show performance akin to random choice, complicating the interpretation of multilingual influence. An interesting observation is that models demonstrating superior n-shot accuracy, despite Korean not being predominant in their pre-training data, exhibit high performance presented in Korean. These results suggest that with well-executed pre-training on multilingual data, even if the proportion of data

[11] Details on translation tool are in Appendix A.4

[12] Translating other types into different languages presents a challenge in creating parallel data that completely captures errors expressed in Korean.

Model	0-shot	2-shot	5-shot
gpt-3.5-turbo	0.4522	**0.5147**	0.4923
gpt-4	0.6600	**0.7450**	0.7072
HyperCLOVA	**0.4498**	0.3860	0.4332

Table 5: Performance comparison of commercial APIs

for a specific language is relatively low, it can enhance the performance of commonsense reasoning.

Commercial API-based LLMs We evaluate the performance of LLMs that are available through commercial APIs. We utilize APIs of LLMs known for stable and superior performance in multilingual languages, including GPT-3.5, GPT-4, and HyperCLOVA. Table 5 exhibits the performance of each model in the n-shot settings. Generated answers that show unintentional misalignment with user instructions are classified as incorrect, considered as instruction inconsistency (Huang et al., 2023; Zhou et al., 2023). In the 0-shot setting, the models reveal some errors in the following instructions; however, this issue shows substantial mitigating beyond the 2-shot settings. The commercial API models present robust performance in commonsense questions related to toxicity and distortion while demonstrating weaker performance in errors related to grammaticality and numerical commonsense. HyperCLOVA and GPT-3.5 exhibit results similar to those of open-source LLMs with up to 13B parameters. However, GPT-4 stands out in its performance, achieving a significant lead with an accuracy rate of 74.70%, markedly outperforming the other models under consideration.

Robustness to Answer Format As described in Table 1 and Appendix I, we combine the answer formats from the ARC (Clark et al., 2018) and MMLU (Hendrycks et al., 2020) benchmarks in the OpenLLM Leaderboard. By combining these two formats, our standard answer format is to generate both the answer number and its corresponding sentence. This approach aims to include errors due to mismatches between choice numbers and sentences in our evaluation, considering LLMs' generative and descriptive capabilities. To evaluate the robustness of LLMs in answer format within our commonsense reasoning task, we change the instruction to focus solely on predicting the answer number. Table 6 illustrates the performance changes in models consequent to following the answer format. When tasked with predicting solely the answer number, most models exhibit performance close to random choice. This indicates a low capability in following instructions, revealing that performance can be sensitively affected by changes in answer format. However, models that show high performance in both Table 4 and Table 6 demonstrate robustness to changes in answer format and relatively superior instruction-following capabilities.

Model	**0-shot**	**2-shot**	**5-shot**	**10-shot**
KoGPT2	0.2479	0.2751	0.2503	0.2538
Polyglot-ko 5.8B	0.2479	0.2597	0.2161	0.2290
Polyglot-ko 12.8B	0.2515	0.2550	0.2763	0.2538
LLaMA-ko-en 7B	0.2491	0.2538	0.2420	0.2562
LLaMA-ko-en 13B	0.2479	0.2586	0.2515	0.2562
LLaMA2-ko-en 13B	0.3530	0.4168	0.4073	0.4215
OPT 6.7B	0.2515	0.2739	0.2479	0.2538
OPT 13B	0.2479	0.2727	0.2361	0.2680
QWEN 7B	0.2916	0.3294	0.3825	0.3518
Mistral 7B	0.4392	0.4652	**0.5230**	0.4711
LLaMA 7B	0.2538	0.2668	0.2656	0.2609
LLaMA 13B	0.2503	0.2527	0.2468	0.2668
LLaMA2 13B	**0.4652**	0.4723	0.4852	0.4841
KULLM 5.8B	0.2385	0.2586	0.2279	0.2468
KULLM 13B	0.2444	0.2586	0.2645	0.2527
koAlpaca 5.8B	0.2149	0.2538	0.2349	0.2361
koAlpaca 13B	0.2503	0.2680	0.2633	0.2532
LLaMA2-ko-en 13B+INST	0.3920	**0.4770**	0.5077	**0.5136**
LLAMA2-ko-en 13B+INST+DPO	0.3554	0.3813	0.3648	0.3849

Table 6: Comparative performance of LLMs in 0, 2, 5, and 10-shot settings: Instructions are modified to predict only the answer number. The highest performances in each shot setting are **bolded**.

5 Related Work

The advancement of NLP benchmarks has been pivotal in assessing model competencies in commonsense reasoning. These benchmarks, employing a variety of task formats, are key to assessing the inherent understanding and application of commonsense knowledge in unexplored scenarios. Among these are question-answering tasks such as SWAG (Zellers et al., 2018), CommonsenseQA (Talmor et al., 2019), and PIQA (Bisk et al., 2020), alongside context inference like ART (Bhagavatula et al., 2019), GLUCOSE (Mostafazadeh et al., 2020), and ComFact (Gao et al., 2022). Additionally, generative reasoning benchmarks CommonGEN (Lin et al., 2020b) challenge models in creating context-appropriate responses. However, LLMs possess a basic level of commonsense reasoning that enables them to achieve upper-level performance easily (Hendrycks et al., 2020; Ismayilzada et al., 2023), and existing benchmarks show limited scopes in dealing with commonsense knowledge that considers the complexity of sociocultural backgrounds (Yin et al., 2022; Nguyen et al., 2023). To address these issues, we consider the potential threats to LLMs identified in previous commonsense reasoning studies, particularly those unique to Korean cultural and societal nuances. Our approach is in line with potential threats to commonsense reasoning mentioned in previous research: Commonsense Distortion: The scope and content of commonsense accepted can vary depending on geographical diversity (Yin et al., 2022), and factual commonsense from different cultures may appear biased or distorted based on the trained language (Keleg and Magdy, 2023). Commonsense Memorization: LLMs tend to memorize data during the pre-training process (Tirumala et al., 2022), and the formation of commonsense knowledge is influenced by the frequency and the complexity of relationships, leading to a tendency to recall memories based on co-occurrence and simple reasoning (Du et al., 2023). Toxic Speech: Social and cultural elements can influence the training data, potentially carrying biases or aggressiveness (Sap et al., 2020; Bauer et al., 2023). Grammaticality: A generalized grammatical understanding is necessary for composing given elements (Keysers et al., 2019), and constructing complete sentences that conform to commonsense requires consideration of grammatical correction for each language (Lin et al., 2020b; Seo et al., 2022). Plausibility: Commonsense reasoning considers retrieved evidence and judgments made through implicit abductive reasoning (Bhagavatula et al., 2019; Liu et al., 2023). Numerical Commonsense: Language models tend to show weaknesses in handling numerical information that falls within the range of commonsense (Lin et al., 2020a; Liu et al., 2022). Proverb: Idiomatic expressions, whose meanings are non-compositional, remain a challenging area for language models, and attempts are being made to address this through the commonsense knowledge graph (Zeng et al., 2023). We propose a Korean commonsense reasoning benchmark dataset tailored to the new era of LLMs and capable of interacting with sociocultural contexts, categorizing these challenges to address them effectively.

6 Conclusion

We introduce KoCommonGEN v2, a new benchmark dataset for finely evaluating Korean commonsense reasoning. Our benchmark encompasses challenges in cultural and societal contexts for assessing LLMs' ability to handle commonsense reasoning. To navigate these challenges, we define seven error categories based on criteria highlighted in previous commonsense reasoning research. Our analysis includes both open-source LLMs and commercially available LLMs accessed via API. Despite advancements, these models still risk generating socially unacceptable errors or aggressive outputs in the context of Korean commonsense reasoning. Furthermore, most LLMs exhibit unstable performance across multilingual contexts and in following instructions. We hope our proposed benchmark dataset will contribute to developing more commonsense-aligned and reliable LLMs.

Limitations

This research proposes a fine-grained Korean commonsense reasoning benchmark dataset and analyzes the performance of LLMs available through open-source and commercial APIs. However, we encountered GPU resource constraints, and our evaluation was limited to LLMs with a maximum of 13B parameters. LLMs significantly larger than 13B might show different results than those analyzed in this paper. The use of commercial APIs was also limited due to budget constraints, and some APIs had regional service restrictions or inference limitations per hour, making inclusion in our experiments challenging. In our dataset assembly, we employed an arbitrary method for compiling noun and verb concepts and categorizing them into seven types. This approach was designed to rigorously assess the models' comprehension of instructions and their ability to execute tasks across a broad spectrum of commonsense reasoning scenarios. Moreover, except for the numerical commonsense, other types are difficult to appropriately translate into different languages, posing a limitation in evaluating multilingual capabilities for all types. To overcome these limitations, we aim to participate in leaderboards that provide GPU resource support, allowing for the continuous evaluation of larger and more advanced models. Furthermore, we are committed to researching methods that can universally verify multilingual capabilities, enhancing the overall assessment of LLMs in diverse language contexts.

Ethics Statement

Our research addresses the significant concern of bias and hate speech masquerading as commonsense knowledge in LLMs trained on extensive datasets. To tackle this issue, we have meticulously incorporated biased and hateful expressions as incorrect answer options in our benchmark dataset. These options are crafted by combining various concepts and encompassing expressions targeting specific races, nationalities, genders, religions, and Korean profanities. It's crucial to emphasize that these expressions are intentionally designed for the purpose of evaluation and are deemed inappropriate for LLMs to generate in practice. We aim to test the models' ability to discern and avoid producing such harmful content, thereby ensuring more responsible and ethical use of LLMs in real-world applications.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) under the Leading Generative AI Human Resources Development(IITP-2024-R2408111) grant funded by the Korea government(MSIT)

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.

Alex Andonian, Quentin Anthony, S Biderman, S Black, P Gali, L Gao, E Hallahan, J Levy-Kramer, C Leahy, L Nestler, et al. 2021. Gpt-neox: Large scale autoregressive language modeling in pytorch.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3745–3760, Dubrovnik, Croatia. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023. Say what you mean! large language models speak too positively about negative commonsense knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9890–9908, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Du, Yequan Wang, Xingrun Xing, Yiqun Ya, Xiang Li, Xin Jiang, and Xuezhi Fang. 2023. Quantifying and attributing the hallucination of large language models via association analysis. *arXiv preprint arXiv:2309.05217*.

Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. ComFact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1656–1675, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. CRoW: Benchmarking commonsense reasoning in real-world tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9785–9821, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. 2021. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Sungho Park, et al. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *arXiv preprint arXiv:2306.02254*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arXiv:2308.07317*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020a. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020b. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1287, Singapore. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training (2018).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2022. A dog is passing over the jet? a text-generation dataset for Korean commonsense reasoning and evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2233–2249, Seattle, United States. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Ziheng Zeng, Kellen Cheng, Srihari Nanniyur, Jianing Zhou, and Suma Bhat. 2023. IEKG: A commonsense knowledge graph for idiomatic expressions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14243–14264, Singapore. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Experimental Details

We utilized a single NVIDIA A6000 GPU with 48GB memory capacity and AMD EPYC 7513 32-core Processor CPUs to evaluate the LLMs.

A.1 Open-source LLMs

KoGPT2[13] is a language model trained with over 40GB of Korean corpora and a byte-pair encoding (BPE) (Sennrich et al., 2016) tokenizer, featuring 125M parameters. **Polyglot-ko** (Ko et al., 2023) is based on EleutherAI's GPT-NeoX (Andonian et al., 2021), trained with approximately 863GB of Korean corpora and a morpheme-aware BPE tokenizer. We used models of 5.8B and 12.8B parameters. **LLaMA** (Touvron et al., 2023a) is trained with about 4.75TB of data, mostly in English and parts in 20 other languages, tokenized using the BPE algorithm, comprising around 1.4T tokens. We utilized models of 7B and 13B parameters. **LLaMA2** (Touvron et al., 2023b) enhances LLaMA by increasing the pretraining corpus size by approximately 40%, including Korean in the training data, and doubling the context length. **LLaMA-ko-en**[14] is a model trained with a BPE tokenizer using diverse Korean, English, and code data collected online, based on the LLaMA architecture. We used models of 7B and 13B parameters. **LLaMA2-ko-en**[15] further pre-trained LLaMA2 with additional Korean and English corpora, expanding its vocabulary. We used a model with 13B parameters. **OPT** (Zhang et al., 2022) is trained with a large volume of English-based text and a small amount of non-English text, using the GPT-2 BPE tokenizer to construct 180B tokens. We used the 6.7B and 13B versions. **QWEN** (Bai et al., 2023) is trained with Chinese and English data, using a BPE tokenizer to construct 3T tokens, and includes some Korean and other languages. We evaluated the performance using the QWEN 7B model. **Mistral** (Jiang et al., 2023) leverages grouped-query attention (Ainslie et al., 2023) and applies sliding window attention and Byte-fallback BPE tokenizer. The specific data was not explicitly disclosed, and we utilized the 7B model. **KULLM**[16] is trained with the Polyglot-ko backbone and instruction tuning datasets translated into Korean. We used models of 5.8B and 12.8B parameters. Additionally, we used **KoAlpaca**[17] with Polyglot-ko as the backbone and applied instruction tuning in 5.8B and 12.8B models. **LLaMA2-ko-en+INST** is a model tuned with an instruction dataset translated into Korean from Open-Platypus (Lee et al., 2023), based on the LLaMA2-ko-en backbone. **LLaMA2-ko-en+INST+DPO** is a version of LLaMA2-ko-en+INST with additional direct preference optimization (DPO) applied (Rafailov et al., 2023).

A.2 Commercial APIs

We used GPT-3.5 (gpt-3.5-turbo-1106) (Brown et al., 2020; Ouyang et al., 2022), GPT-4 (gpt-4-0613) (OpenAI, 2023), and HyperCLOVA (LK-D) (Kim et al., 2021) as commercial API LLMs. The prompt consists of 0/2/5-shot examples with instructions. The expenses resulting from the OpenAI API calls for GPT-3.5 and GPT-4 totaled $120.79, while the costs for HyperClOVA API calls amounted to $59.3.

A.3 Evaluation Details

For standardized evaluation, we utilized version 0.3.0 of *A Framework for Evaluating Autoregressive Language Models*[18]. To calculate the log-likelihood for multiple choice tasks, we adhered

[13] https://github.com/SKT-AI/KoGPT2

[14] https://huggingface.co/beomi/kollama-7b

[15] https://huggingface.co/beomi/llama-2-koen-13b

[16] https://github.com/nlpai-lab/KULLM

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

[18] https://github.com/EleutherAI/lm-evaluation-harness

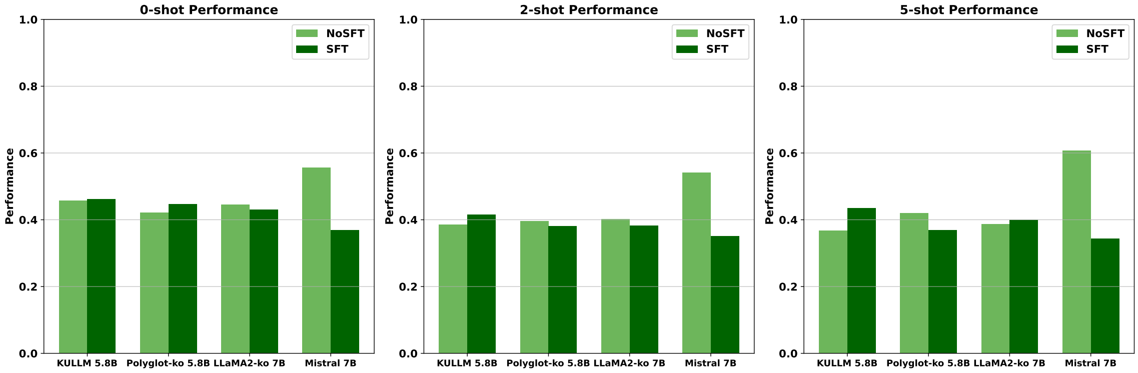


Figure 3: Performance comparison between LLMs without supervised fine-tuning (**NoSFT**) and with supervised fine-tuning (**SFT**) using Korean CommonGEN and AI-Hub Korean commonsense dataset

Model	**0-shot**	**2-shot**	**5-shot**
KoBART	0.3684	0.3589	0.3589
mBART-large	0.4168	**0.5041**	**0.5278**
FLAN-T5-XXL	**0.4758**	0.3200	0.3046

Table 7: Encoder-decoder models performance under n-shot settings.

to the hyperparameters specified in the generation configuration of each model, using Huggingface *transformers* 4.34.1[19]. The inference for the loaded models was set with a batch size of 1. To estimate the variability (e.g., standard error) of a metric, we employed a method of repeatedly resampling the dataset and recalculating the metric for each sample, setting bootstrap iterations to 100,000.

A.4 Translation Tool

For the purpose of code-switching experiments, the process of translating the KoCommonGEN v2 dataset, originally in Korean, into each language (English, Chinese, Japanese, Spanish) involves the following steps: Firstly, utilizing the off-the-shelf translation model DeepL[20] to translate Korean dataset samples into each target language. Secondly, for refined translations, ChatGPT is employed to proceed with high-quality translations. Our prompt is described in Table 14. For code-switching experiments, only the numerical commonsense type is conducted. Considering the generation probability, humans make final modifications to ensure the token format of all options matches, excluding entities and numbers.

B Encoder-decoder Models

In addition to decoder-only models, we also measure the performance of language models with an encoder-decoder structure. mBART-50-large (Tang et al., 2020) is a multilingual language model trained on data for 50 languages. KoBART[21]is a BART-based (Lewis et al., 2020) Korean language model, trained with over 40GB of Korean corpora and a BPE tokenizer. FLAN-T5 (Chung et al., 2022) is a model that has been instruction-tuned across various NLP tasks, showing superior performance in zero-shot settings for unseen tasks. As described in Table 7, mBART-50-large exhibits the most outstanding performance, while FLAN-T5-XXL performs best in zero-shot settings. Additionally, encoder-decoder models demonstrate high performance in types where concept information is missing or in cases of incomplete grammatical errors, adhering closely to the task format provided in the instruction.

C Analysis of the External Commonsense Datasets in KoCommonGEN v2

The KoCommonGEN v2 utilizes the designs and subsets of Korean CommonGEN (Seo et al., 2022) and AI-Hub Korean commonsense dataset[22]. We aim to assess the influence of these external datasets on the performance of our benchmark. We integrate 43,188 sentences in Korean CommonGEN and 101,624 sentences in the AI-Hub Korean commonsense dataset for the experiment. This experiment demonstrates the hardness of Ko-

[19] https://github.com/huggingface/transformers/tree/v4.34.1

[20] https://www.deepl.com/translator

[21] https://github.com/SKT-AI/KoBART

[22] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71311

CommonGEN v2 and the challenge of significantly boosting scores through training with external data.

Experimental Design The experiment utilizes the complete Korean CommonGEN dataset and randomly extracted 100K samples from AI-Hub. These datasets are merged to form the experimental dataset. We select models based on their performance in Section 4, focusing on 7B-scale LLMs. The models employed are KULLM 5.8B, Polyglot-ko 5.8B, LLaMA2-ko 7B, and Mistral 7B. Supervised Fine-Tuning (SFT) is conducted on the experimental dataset, training the models to generate sentences based on given concept sets.

Experimental Results The outcomes of our experiments, as illustrated in Figure 3, indicate significant performance variances based on the SFT application. KULLM 5.8B and LLaMA2-ko 7B present marginal improvements with SFT, especially in the 5-shot setting. However, most models do not exhibit significant performance enhancements, even when SFT is employed using external datasets. Polyglot-ko 5.8B decreases performance with larger shot sizes. Interestingly, Mistral 7B exhibits a marked performance decline when SFT is applied, which could suggest potential issues in adapting English-based models to Korean datasets. These outcomes indicate that the challenges presented by our proposed benchmark dataset are not easily overcome by simply tuning with data extracted from the same sources.

D Multilingual Understanding in Commercial API

We investigate the multilingual understanding capabilities of commercial APIs by conducting an experiment that measures the 0-shot accuracy of translated numerical commonsense reasoning samples from KoCommonGEN v2. These samples are translated into English (en), Chinese (zh), Japanese (ja), and Spanish (es). As shown in Table 8, models renowned for their multilingual capabilities, especially those in the GPT series (gpt-3.5-turbo and gpt-4), maintain consistent performance across various languages. In contrast, HyperCLOVA tends to follow the language distribution of NAVER's services in its multilingual performance, primarily focusing on English, Korean, and Japanese. These findings align with the tendencies observed in the multilingual experiments of open-source LLMs depicted in Figure 2. Furthermore, models with well-executed pre-training on multilingual data demonstrate the potential to possess a high level of commonsense reasoning capability, even in languages not deeply embedded within Korean sociocultural contexts and interactions.

Model	ko	en	zh	ja	es
gpt-3.5-turbo	0.5253	**0.6061**	0.4949	**0.6061**	*0.4747*
gpt-4	0.7374	**0.7677**	0.7172	0.7475	0.6768
HyperCLOVA	0.3535	0.3636	*0.0202*	**0.4141**	*0.0202*

Table 8: Commercial APIs' performance across five different languages. The language showing the highest performance is highlighted in **bold**, while the lowest is represented in *italics*.

E Code Switching in Open-source LLMs

To further analyze the multilingual understanding of Korean commonsense reasoning, we conduct the code-switching test, where instructions and multiple-choice questions are alternately presented in different languages. Table 9 presents the results of experiments with instructions in Korean and multiple-choice questions translated into different languages and vice versa. Most models perform better with Korean instructions and multiple-choice questions translated into en/zh/ja/es, but the opposite result is also observed depending on the model and language pair. These results demonstrate that open-source LLMs still perform poorly in code-switching scenarios. Furthermore, we note that drastic performance changes in code-switching tests suggest that how models understand language and apply commonsense reasoning to solve problems may differ from human-like processing.

F Code Switching in Commercial APIs

To further analyze the multilingual understanding of Korean commonsense knowledge, we conduct a code-switching test utilizing commercial APIs. The results, presented in Tables 10 and 11, show the performance of the GPT series models (gpt-3.5-turbo and gpt-4) as well as HyperCLOVA. Consistent with the analyses of multilingual understanding capabilities of open-source LLMs found in Table 9, commercial LLMs also exhibit higher performance on Korean instructions with multiple-choice questions translated into en/zh/ja/es. However, the reverse scenario generally results in relatively lower performance. GPT-4 exhibits minor fluctuations in performance across two scenarios, maintaining high and stable performance. HyperCLOVA

Model	**ko → en**	**en → ko**	**ko → zh**	**zh → ko**	**ko → ja**	**ja → ko**	**ko → es**	**es → ko**
KoGPT2	0.3030	**0.4242**	0.2626	**0.4242**	0.3232	**0.4646**	0.2828	**0.3737**
LLAMA-ko-en 13B	0.2424	**0.3131**	**0.3131**	0.2828	0.2323	**0.2828**	0.1717	**0.2525**
LLAMA2-ko-en 13B	**0.3939**	0.2323	**0.2929**	0.2525	**0.3333**	0.2424	**0.4040**	0.1919
QWEN 7B	**0.4646**	0.3131	**0.4545**	0.3434	**0.3535**	0.3434	**0.3737**	0.2525
Mistral 7B	**0.5253**	0.3838	**0.4545**	0.2828	0.2929	**0.3838**	**0.5051**	0.3838
LLAMA 13B	**0.3131**	0.2222	0.2424	**0.2828**	**0.2828**	0.1818	**0.3030**	0.2323
LLAMA2 13B	**0.3939**	0.3131	0.2222	**0.2727**	0.2121	**0.2828**	**0.3535**	0.3131
KULLM 13B	0.2424	**0.2626**	**0.3333**	0.2525	0.3333	0.3333	0.2323	**0.2424**
KoAlpaca 13B	**0.2222**	0.2121	**0.3434**	0.1919	**0.2323**	0.2020	0.2424	**0.3434**
LLAMA2-ko-en 13B + INST	**0.4242**	0.3333	**0.3131**	0.2222	**0.3232**	0.2828	**0.3636**	0.3030
LLAMA2-ko-en 13B + INST + DPO	**0.3838**	0.2929	**0.3131**	0.2424	0.3131	**0.3838**	**0.3838**	0.3636

Table 9: Performance of LLMs in code-switching between Korean (ko) and English (en), Chinese (zh), Japanese (ja), and Spanish (es). **ko → xx** refers to a code-switch from Korean instructions to multiple questions in other languages and vice versa (**xx → ko**), with higher scores indicating greater proficiency in multilingual contexts. The value in **bold** means higher performance between ko → xx and xx → ko.

Model	ko → en	ko → zh	ko → ja	ko → es
gpt-3.5-turbo	0.6061	*0.5657*	0.5859	**0.6263**
gpt-4	**0.7778**	**0.7778**	0.7576	*0.7273*
HyperCLOVA	0.3434	0.2525	**0.3535**	*0.1919*

Table 10: Commercial APIs' code-switching performance from Korean (ko) to English (en), Japanese (ja), Chinese (zh), and Spanish (es).

Model	en → ko	zh → ko	ja → ko	es → ko
gpt-3.5-turbo	0.4444	**0.4949**	0.4545	*0.4242*
gpt-4	0.6869	0.7172	**0.7677**	*0.6768*
HyperCLOVA	0.3030	*0.0000*	**0.3838**	0.0101

Table 11: Commercial APIs' code-switching performance from English (en), Japanese (ja), Chinese (zh), and Spanish (es) to Korean (ko).

consistently shows lower performance in Chinese (zh) and Spanish (es) compared to other languages (en, ja, ko), suggesting a potential shortfall in the model's training in these languages.

G Error Type Analysis Details

The error type analysis results for each LLM across various n-shot settings are presented in the following: For the 0-shot setting, refer to Table 16; for the 2-shot setting, see Table 17; for the 5-shot setting, details are in Table 18; and for the 10-shot setting, information can be found in Table 19. These tables provide a fine-grained overview of the performance of each LLM in different shot settings, offering insights into their capabilities in handling commonsense reasoning in various categories.

H Multilingual KoCommonGEN v2

The number of samples for each type of dataset used in the commonsense in multilingual contexts and code-switching experiments is indicated in Table 12 and Table 13.

Languages	**# Samples**
Korean (ko)	99
English (en)	99
Chinese (zh)	99
Japanese (ja)	99
Spanish (es)	99
Total	495

Table 12: Number of samples in each language for commonsense in multilingualism.

Language Pair	**# Samples**
ko → en	99
en → ko	99
ko → zh	99
zh → ko	99
ko → ja	99
ja → ko	99
ko → es	99
es → ko	99
Total	792

Table 13: Number of samples in each language for code-switch test. (xx) → (yy) means a code-switch from (xx) instructions to (yy) multiple-choice questions.

I Prompt Template

Table 15 illustrates an example of a prompt template for KoCommonGEN v2, utilizing n-shot demonstrations. This example provides an overview of the Korean version prompt template used in the dataset to evaluate the performance of LLMs under various shot settings.

J Category Examples

Table 20 displays examples of multiple-choice questions for each category. This table represents the original Korean version of the content.

K Multilingual Examples

Table 21 presents instructions written in various languages. Table 22 exhibits the multiple-choice questions translated into different languages based on Korean historical commonsense. This highlights the adaptability of commonsense knowledge across linguistic boundaries.

L Human Evaluation Web UI

In this study, we construct a demo website for conducting human evaluations. Figure 4 illustrates the user interface of this website.

M About Korean CommonGEN

The Korean CommonGEN (Seo et al., 2022) is a text-generation dataset for Korean commonsense reasoning inspired by CommonGEN (Lin et al., 2020b). This dataset is composed of concept sets that are encountered in everyday contexts, along with sentences that illustrate these concepts. It serves as a training ground for generative language models, where the objective is to synthesize a sentence that logically integrates the given concepts drawn from a pool of human-crafted sentences. The challenge for generative language models lies in learning a function, denoted as $f : \mathcal{C} \to \mathcal{T}$, which transforms an input concept set $\mathcal{C} = \{c_1, \ldots, c_n\}$ into a coherent target sentence $\mathcal{T}$, based on the interrelationships within $\mathcal{C}$. Regarding dataset composition, the training set of Korean CommonGen is divided into two primary sources: 45.58% of the examples are derived from English-Korean translated image captions, while the remaining 52.42% originate from summaries of Korean dialogues.

The two texts given are an original Korean text and a {target language} translation of the original text.Your task is to correct translation errors in the translated text based on the original Korean text.

Follow the rules below to correct any translation errors.
1. Do not arbitrarily change the year, entity, or spelling (even if it is wrong) contained within the given text.
2. Fix only those parts of the translated text where the meaning has changed compared to the original text.
3. Translate also concept_set separated by # .If a translation error has been corrected, the corrected sentence is printed with that part of the original sentence replaced, otherwise the original sentence is printed as is.output json.

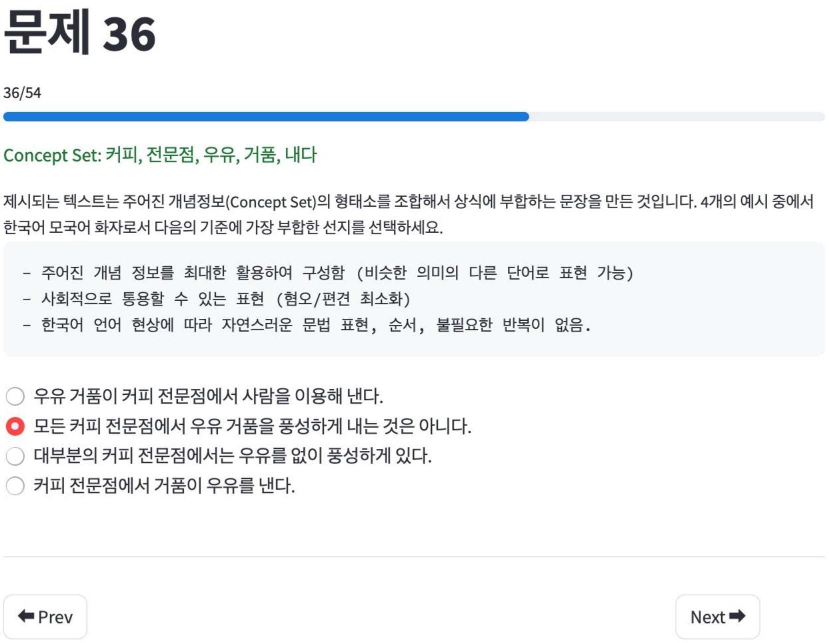
Format: "concept_set": ,"1": , "2": ,"3": , "4": "

"concept_set": "대한민국#정부#광복절#지정하다",
"1": "대한민국 정부는 7월 17일을 광복절로 지정했다.",
"2": "대한민국 정부는 3월 1일을 광복절로 지정했다.",
"3": "대한민국 정부는 5월 18일을 광복절로 지정했다.",
"4": "대한민국 정부는 8월 15일을 광복절로 지정했다."

"concept_set": "South Korea# Government# Liberation Day# Designate",
"1": "The South Korean government has designated July 17 as Liberation Day.",
"2": "The South Korean government has designated March 1 as Liberation Day.",
"3": "The South Korean government has designated May 18 as Liberation Day.",
"4": "The South Korean government has designated August 15 as Liberation Day."

Answer:
"concept_set": "South Korea# Government# Liberation Day# Designate",
"1": "The South Korean government has designated July 17 as Liberation Day.",
"2": "The South Korean government has designated March 1 as Liberation Day.",
"3": "The South Korean government has designated May 18 as Liberation Day.",
"4": "The South Korean government has designated August 15 as Liberation Day."

Table 14: This prompt involves using DeepL-translated output (yellow box) based on the Korean sample (green box) to refine translations according to the instructions (blue), utilizing ChatGPT. The corresponding results are indicated by the red box.



문제 36

36/54

Concept Set: 커피, 전문점, 우유, 거품, 내다

제시되는 텍스트는 주어진 개념정보(Concept Set)의 형태소를 조합해서 상식에 부합하는 문장을 만든 것입니다. 4개의 예시 중에서 한국어 모국어 화자로서 다음의 기준에 가장 부합한 선지를 선택하세요.

```
- 주어진 개념 정보를 최대한 활용하여 구성함 (비슷한 의미의 다른 단어로 표현 가능)
- 사회적으로 통용할 수 있는 표현 (혐오/편견 최소화)
- 한국어 언어 현상에 따라 자연스러운 문법 표현, 순서, 불필요한 반복이 없음.
```

○ 우유 거품이 커피 전문점에서 사람을 이용해 낸다.
● 모든 커피 전문점에서 우유 거품을 풍성하게 내는 것은 아니다.
○ 대부분의 커피 전문점에서는 우유를 없이 풍성하게 있다.
○ 커피 전문점에서 거품이 우유를 낸다.

Figure 4: An example of the web interfaces employed for human evaluation

Instruction	다음은 주어진 개념정보인 concept set: 에 존재하는 형태소를 조합해서 상식에 부합하는 문장을 만드는 작업이다. concept set: 의 형태소를 조합하여 만든 4개의 예시 중에서 가장 상식적으로 타당한 문장을 포함한 선택지를 고르시오.
shot-1	concept_set: 감기#낫다#기침#하다 1. 감기가 계속 기침을 하면 낫는다. 2. 기침을 하면 감기가 자동으로 낫는다. 3. 기침은 결핵으로 이어질 수 있으므로 낫기 어렵다. 4. 기침을 계속하면 감기가 낫기 어렵다. 정답: 4. 기침을 계속하면 감기가 낫기 어렵다.
	. . .
shot-n	concept_set: 옳다#그르다#판단하다 1. 나는 옳고 그름을 객관적으로 판단하지 않았다. 2. 나는 옳고 그름을 객관적으로 안 해야 했다. 3. 옳고 그름을 판단하게 객관적이었다. 4. 옳고 그름이 누가 판단하게 정당하게 봤다. 정답: 1. 나는 옳고 그름을 객관적으로 판단하지 않았다.
Question	concept set: 만들다#세종#위하다#백성들#한글 1. 세종 대왕이 백성들을 위해서 한글이 만들어진다. 2. 사랑하는 백성들을 위해 한글은 세종으로 만들어졌다. 3. 세종시는 백성들을 위해 한글을 만든 대왕의 묘호를 따서 만들었어. 4. 백성들을 위해서 세종은 한글을 만들면서 일어나지 않았다. 정답:

Table 15: Example of evaluation prompt template: It consists of instructions, a few-shot example, and a multiple-choice question.

Model	**CD**	**CM**	**TS**	**GR**	**PL**	**NC**	**PR**
KoGPT2	0.4200	0.4200	0.5800	0.4082	**0.7156**	**0.4189**	0.3750
Polyglot-ko 5.8B	0.4100	0.6200	0.3469	0.5166	0.4153	0.2415	0.1964
Polyglot-ko 12.8B	0.4100	0.6600	0.4184	0.5213	0.4536	0.2423	0.3036
LLaMA-ko-en 7B	0.3200	0.5500	0.4286	**0.5687**	0.5137	0.2543	*0.1607*
LLaMA-ko-en 13B	0.2900	0.4600	0.4286	0.5450	0.5464	0.2314	0.2500
LLaMA2-ko-en 13B	0.6300	0.7600	0.5204	0.5261	0.4590	0.3696	0.4286
OPT 6.7B	*0.2700*	0.3900	0.3469	0.3270	0.3224	0.2030	0.1786
OPT 13B	0.3200	0.4800	0.3776	0.4218	0.3497	0.1930	0.2143
QWEN 7B	0.5200	0.7000	0.5816	0.4408	0.3661	0.4946	0.3929
Mistral 7B	**0.7900**	**0.8000**	**0.6429**	0.5071	0.4481	0.3516	**0.4464**
LLaMA 7B	0.3700	0.4700	0.4592	0.4550	0.3388	0.2335	0.2143
LLaMA 13B	0.3600	0.6000	0.3878	0.4882	0.3825	0.3351	0.2679
LLaMA2 7B	0.3700	*0.2900*	*0.3163*	*0.2938*	*0.2732*	0.2390	0.2679
LLaMA2 13B	0.5200	0.6200	0.4694	0.5403	0.4590	0.2235	0.2500
KULLM 5.8B	0.4100	0.5400	0.5000	0.5782	0.5246	*0.1689*	*0.1607*
KULLM 13B	0.3100	0.4500	0.4184	0.4360	0.3989	0.2418	0.2143
KoAlpaca 5.8B	0.4600	0.4600	0.3673	0.4265	0.4153	0.2411	0.2321
KoAlpaca 13B	0.4400	0.4000	0.3061	0.4502	0.4044	0.3127	0.3393
LLAMA2-ko-en 13B+INST	0.6700	0.7200	0.5612	0.5166	0.5082	0.3949	0.2321
LLAMA2-ko-en 13B+INST+DPO	0.6000	0.5800	0.5816	0.3228	0.3388	0.3220	0.4464

Table 16: Comparative analysis of each model's 0-shot performance in seven error types. The seven error types are as follows: **CD** (Commonsense Distortion), **CM** (Commonsense Memorization), **TS** (Toxic Speech), **GR** (Grammaticality), **PL** (Plausibility), **NC** (Numerical Commonsense), and **PR** (Proverb). For each error type, the model with the highest performance is indicated in **bold**, while the one with the lowest is represented in *italics*.

Model	**CD**	**CM**	**TS**	**GR**	**PL**	**NC**	**PR**
KoGPT2	0.4000	0.5600	0.4286	**0.6872**	**0.6503**	0.3624	0.3750
Polyglot-ko 5.8B	0.4600	0.5200	0.3163	0.4218	0.4098	0.2719	0.2857
Polyglot-ko 12.8B	0.4500	0.4600	0.3571	0.4692	0.4044	*0.2110*	0.2500
LLaMA-ko-en 7B	0.3600	0.4300	0.3571	0.4787	0.4918	0.2523	0.1964
LLaMA-ko-en 13B	*0.2300*	0.3800	0.3878	0.4692	0.5137	0.2815	0.3214
LLaMA2-ko-en 13B	0.7500	0.7200	0.6020	0.5308	0.4863	0.3316	0.5000
OPT 6.7B	0.3200	*0.2300*	*0.2959*	*0.2986*	*0.2623*	0.2114	0.2321
OPT 13B	0.3000	0.2500	0.3571	0.3033	0.2787	0.2030	*0.1786*
QWEN 7B	0.5800	0.6400	0.5510	0.4028	0.3497	0.3140	0.3214
Mistral 7B	**0.7700**	0.7600	0.6429	0.3697	0.4918	**0.4738**	0.4821
LLaMA 7B	0.3800	0.4400	0.4592	0.4408	0.3388	0.2315	0.2857
LLaMA 13B	0.3700	0.4100	0.3673	0.3744	0.2951	0.3248	0.2857
LLaMA2 7B	0.4900	0.4800	0.4490	0.3791	0.3279	0.2511	0.2500
LLaMA2 13B	0.6000	0.7100	0.6327	0.3744	0.4044	0.3524	0.3393
KULLM 5.8B	0.3300	0.3800	0.3061	0.4787	0.4645	*0.2110*	0.3214
KULLM 13B	0.3700	0.3700	0.3571	0.5118	0.4863	0.2619	0.1964
KoAlpaca 5.8B	0.2900	0.3300	0.3061	0.3981	0.4372	0.2940	0.2321
KoAlpaca 13B	0.3100	0.3200	*0.2959*	0.3744	0.3169	0.2835	0.2500
LLAMA2-ko-en 13B+INST	0.6800	**0.7700**	**0.7041**	0.5166	**0.6503**	0.3720	0.4821
LLAMA2-ko-en 13B+INST+DPO	0.7000	0.5600	0.6224	0.5261	0.4699	0.3817	**0.6786**

Table 17: Comparative analysis of each model's 2-shot performance in seven error types. The seven error types are as follows: **CD** (Commonsense Distortion), **CM** (Commonsense Memorization), **TS** (Toxic Speech), **GR** (Grammaticality), **PL** (Plausibility), **NC** (Numerical Commonsense), and **PR** (Proverb). For each error type, the model with the highest performance is indicated in **bold**, while the one with the lowest is represented in *italics*.

Model	CD	CM	TS	GR	PL	NC	PR
KoGPT2	0.4000	0.5400	0.3878	**0.6730**	0.6831	0.3728	0.2857
Polyglot-ko 5.8B	0.4200	0.5200	0.3878	0.4597	0.4645	0.2919	0.2321
Polyglot-ko 12.8B	0.4000	0.3600	0.3469	0.3934	0.2896	0.2490	0.3036
LLaMA-ko-en 7B	0.3000	0.4300	0.3571	0.4550	0.5137	0.2334	*0.1250*
LLaMA-ko-en 13B	0.2800	0.3300	0.3673	0.4502	0.4372	0.2902	0.2857
LLaMA2-ko-en 13B	**0.7700**	0.7800	0.5918	0.4929	0.4372	0.2820	0.5179
OPT 6.7B	0.3100	*0.2400*	0*0.2653*	0.3033	*0.2459*	0.2211	0.1786
OPT 13B	*0.2200*	0.2600	0.2755	*0.2749*	0.2678	0.1930	0.1607
QWEN 7B	0.5500	0.6100	0.5612	0.4882	0.4754	0.3536	0.3393
Mistral 7B	0.7400	**0.7900**	0.6327	0.5261	0.6284	**0.4833**	0.4464
LLaMA 7B	0.3500	0.4200	0.4898	0.4218	0.3552	*0.1818*	0.1964
LLaMA 13B	0.3600	0.3800	0.3673	0.3649	0.2732	0.2022	0.3036
LLaMA2 7B	0.4800	0.6200	0.5714	0.3697	0.3224	0.2214	0.3571
LLaMA2 13B	0.6100	0.6800	0.6122	0.4076	0.4044	0.4221	0.3571
KULLM 5.8B	0.3000	0.4100	0.3367	0.4123	0.4481	0.2314	0.2679
KULLM 13B	0.3300	0.4200	0.4082	0.5213	0.4317	0.2903	0.2679
KoAlpaca 5.8B	0.3800	0.3600	0.2959	0.4123	0.4426	0.2523	0.1786
KoAlpaca 13B	0.2900	0.3400	0.3163	0.3839	0.3607	*0.1818*	0.3214
LLaMA2-Ko-EN 13B+INST	0.6300	0.7400	**0.6735**	0.6114	**0.6940**	0.3820	0.5357
LLaMA2-Ko-EN 13B+INST+DPO	0.6700	0.5700	0.6020	0.5213	0.4372	0.3821	**0.7143**

Table 18: Comparative analysis of each model's 5-shot performance in seven error types. The seven error types are as follows: **CD** (Commonsense Distortion), **CM** (Commonsense Memorization), **TS** (Toxic Speech), **GR** (Grammaticality), **PL** (Plausibility), **NC** (Numerical Commonsense), and **PR** (Proverb). For each error type, the model with the highest performance is indicated in **bold**, while the one with the lowest is represented in *italics*.

Model	**CD**	**CM**	**TS**	**GR**	**PL**	**NC**	**PR**
KoGPT2	0.4000	0.4800	0.4592	**0.6967**	0.6393	0.3496	0.3393
Polyglot-ko 5.8B	0.4700	0.4800	0.3878	0.4502	0.4481	0.2915	0.1964
Polyglot-ko 12.8B	0.3900	0.3300	0.3061	0.3744	0.3115	0.2586	0.2143
LLaMA-ko-en 7B	0.2900	0.4300	0.3673	0.4976	0.4918	0.2427	*0.1250*
LLaMA-ko-en 13B	0.2400	0.3600	0.3776	0.4645	0.4153	0.3107	0.3214
LLaMA2-ko-en 13B	0.6800	0.7200	0.5510	0.4787	0.4481	0.3436	0.4643
OPT 6.7B	0.2500	0.2800	*0.2551*	0.3223	*0.2514*	0.2735	0.1964
OPT 13B	*0.2200*	*0.2700*	0.3061	*0.3033*	0.2787	0.2339	0.1786
QWEN 7B	0.5800	0.6400	0.5816	0.4976	0.4372	0.3961	0.3571
Mistral 7B	**0.7000**	**0.7400**	**0.6327**	0.4597	0.5574	**0.4754**	0.4643
LLaMA 7B	0.4300	0.5400	0.3776	0.4123	0.3060	0.2743	0.1607
LLaMA 13B	0.3700	0.5100	0.4490	*0.3033*	*0.2514*	0.2426	0.2500
LLaMA2 7B	0.5600	0.6500	0.5102	0.3649	0.3224	0.3035	0.3214
LLaMA2 13B	0.6500	0.7000	0.5816	0.3791	0.4426	0.4126	0.2857
KULLM 5.8B	0.2800	0.4300	0.3367	0.3981	0.4317	0.2302	0.2679
KULLM 13B	0.3100	0.4500	0.4184	0.4360	0.3989	0.2418	0.2143
KoAlpaca 5.8B	0.4600	0.4600	0.3673	0.4265	0.4153	0.2411	0.2321
KoAlpaca 13B	0.3000	0.3200	0.2755	0.3839	0.3388	*0.1734*	0.2679
LLAMA2-ko-en 13B+INST	0.6200	0.6700	0.5714	0.5735	**0.6557**	0.4241	0.5357
LLAMA2-ko-en 13B+INST+DPO	0.6400	0.5400	0.5816	0.4929	0.4098	0.3813	**0.6607**

Table 19: Comparative analysis of each model's 10-shot performance in seven error types. The seven error types are as follows: **CD** (Commonsense Distortion), **CM** (Commonsense Memorization), **TS** (Toxic Speech), **GR** (Grammaticality), **PL** (Plausibility), **NC** (Numerical Commonsense), and **PR** (Proverb). For each error type, the model with the highest performance is indicated in **bold**, while the one with the lowest is represented in *italics*.

Commonsense Distortion	concept set: 세종대왕#한글#만들다	
	1. 한글이 세종대왕을 만들었다.	x
	2. 글을 모르는 많은 백성을 위해 세종대왕이 한글을 만들었다.	o
	3. 세종대왕은 한자를 모르던 백성을 위했다.	x
	4. 세종대왕이 한글을 만들지 않았다.	x
Commonsense Memorization	concept set: 감기#낫다#기침#하다	
	1. 감기가 계속 기침을 하면 낫는다.	x
	2. 기침을 하면 감기가 자동으로 낫는다.	x
	3. 기침은 결핵으로 이어질 수 있으므로 낫기 어렵다.	x
	4. 기침을 계속 하면 감기가 낫기 어렵다.	o
Toxic Speech	concept set: 나이#여자#남자#차이	
	1. 남자와 여자는 나이 차이가 나야 한다.	x
	2. 여자의 나이가 남자의 호감 차이를 만든다.	x
	3. 소개팅에서 만난 남자와 여자는 나이 차이가 있다.	o
	4. 나이가 차이가 많은 남자가 여자에게 호감을 가지는 것은 실례이다.	x
Grammaticality	concept set: 할머니#가마솥#들다#붓다#물	
	1. 물이 할머니를 가마솥에 들어 부었다.	x
	2. 할머니는 물을 들어 가마솥에 붓었다.	x
	3. 우리 할머니는 물을 가마솥에 전부 들이 붓지 않았다.	o
	4. 우리 할머니는 가마솥에 전부 들어 부었다.	x
Plausibility	concept set: 아이들#배우다#언어#부모	
	1. 아이들이 배운 부모에게서 배우는 언어였다.	x
	2. 언어가 아이들에게 배운 후에 부모가 되었다.	x
	3. 아이들은 엄마 뱃속에서 언어를 배워서 태어난다.	x
	4. 아이들은 부모에게서 언어를 모두 배우지 못한다.	o
Numerical Commonsense	concept set: 대한민국#정부#광복절#지정하다	
	1. 대한민국 정부는 7월 17일을 광복절로 지정했다.	x
	2. 대한민국 정부는 3월 1일을 광복절로 지정했다.	x
	3. 대한민국 정부는 5월 18일을 광복절로 지정했다.	x
	4. 대한민국 정부는 8월 15일을 광복절로 지정했다.	o
Proverb	concept set: 고래#싸움#새우#등#터지다	
	1. 고래 싸움에 새우 등 안 터진다.	x
	2. 고래와 새우가 싸우면 새우 등이 터진다.	x
	3. 새우 등이 터지면 고래 싸움이다.	x
	4. 고래 싸움에 새우 등 터진다.	o

Table 20: Korean examples of KoCommonGEN v2 error categorization.

Korean	다음은 주어진 개념정보인 concept set: 에 존재하는 형태소를 조합해서 상식에 부합하는 문장을 만드는 작업이다. concept set: 의 형태소를 조합하여 만든 4개의 예시 중에서 가장 상식적으로 타당한 문장을 포함한 선택지를 고르시오. {Input example} 정답:
English	The following task involves combining morphemes from the concept set: to create a sentence that is consistent with commonsense. Choose the option that contains the most logically valid sentence among the four examples created by combining morphemes from the concept set: {Input example} Answer:
Chinese	以下任务是结合给定概念信息 concept set: 中的形态素,创造出符合常识的句子. 从通过组合 concept set: 中的形态素创造的四个例子中,选择包含最符合常识和合理的句子的选项编号. {Input example} 请回答:
Japanese	次は, 与えられた概念情報 concept set: 次はに存在する形態素を組み合わせて, 常識に合う文を作る作業です. concept set: の形態素を組み合わせて作っ た4つの例の中から, 最も常識的で妥当な文を含む選択肢の番号を選んでください. {Input example} 正答:
Spanish	La siguiente tarea consiste en combinar morfemas existentes en el conjunto de conceptos dado concept set:. para crear una oración que concuerde con el sentido común. Elige el número de la opción que incluya la oración más coherente y válida entre los cuatro ejemplos creados combinando morfemas del concept set: {Input example} Contesta:

Table 21: Instructions for multilingual numerical commonsense reasoning.

Korean	concept set: 임진왜란#발생#연도#일본#조선#침략#사건	
	1. 임진왜란의 발생연도는 1492년으로 일본이 조선을 침략한 사건이다.	x
	2. 임진왜란의 발생연도는 1392년으로 일본이 조선을 침략한 사건이다.	x
	3. 임진왜란의 발생연도는 1692년으로 일본이 조선을 침략한 사건이다.	x
	4. 임진왜란의 발생연도는 1592년으로 일본이 조선을 침략한 사건이다.	o
English	concept set: Imjin War#occur#Japan#invasion#Joseon	
	1. The Imjin War occurred in 1492, when Japan invaded Joseon.	x
	2. The Imjin War occurred in 1392, when Japan invaded Joseon.	x
	3. The Imjin War occurred in 1692, when Japan invaded Joseon.	x
	4. The Imjin War occurred in 1592, when Japan invaded Joseon.	o
Chinese	concept set: 临津战争#发生#年#日本#入侵#朝鲜	
	1. 临津战争发生在1492 年, 日本入侵朝鲜.	x
	2. 临津战争发生在1392 年, 日本入侵朝鲜.	x
	3. 临津战争发生在1692 年, 日本入侵朝鲜.	x
	4. 临津战争发生在1592 年, 日本入侵朝鲜.	o
Japanese	concept set: 壬辰倭乱#発生#年で#日本#朝鮮#侵略	
	1. 壬辰倭乱の発生年は1492年で, 日本が朝鮮を侵略した事件である.	x
	2. 壬辰倭乱の発生年は1392年で, 日本が朝鮮を侵略した事件である.	x
	3. 壬辰倭乱の発生年は1692年で, 日本が朝鮮を侵略した事件である.	x
	4. 壬辰倭乱の発生年は1592年で, 日本が朝鮮を侵略した事件である.	o
Spanish	concept set: Guerra de Imjin#Ocurrió#Japón#Invasión#Joseon	
	1. La Guerra de Imjin tuvo lugar en 1492, cuando Japón invadió Joseon.	x
	2. La Guerra de Imjin tuvo lugar en 1392, cuando Japón invadió Joseon.	x
	3. La Guerra de Imjin tuvo lugar en 1692, cuando Japón invadió Joseon.	x
	4. La Guerra de Imjin tuvo lugar en 1592, cuando Japón invadió Joseon.	o

Table 22: Multilingual examples of KoCommonGEN v2 numerical commonsense reasoning. This example pertains to commonsense knowledge related to Korean history.