On the Effect of Pretraining Corpora on In-context Few-shot Learning by a Large-scale Language Model

Anonymous ACL submission

Abstract

Many recent studies on large-scale language models have reported successful in-context few-shot learning ability. However, the in-depth analysis of when in-context learning occurs is still lacking. For example, it is unknown how in-context learning performance changes as the training corpus varies. Here, we investigate the effect of the source and size of the pretraining corpus in a Korean GPT-3. From our in-depth investigation, we introduce the following observations: (1) in-context learning performance heavily depends on the corpus domain source, and the size of pretraining corpus does not necessarily predict the emergence of in-context learning, (2) in-context learning ability can emerge when a language model is trained on a combination of multiple corpora, even when each corpus does not result in in-context learning on its own, (3) pretraining with a corpus related to a downstream task does not always guarantee the competitive in-context learning performance of the downstream task, and (4) the relationship between language modeling (measured in perplexity) and in-context learning does not always correlate: e.g., low perplexity does not always imply high in-context learning performance.

1 Introduction

NLP community has surprised that zero/few-shot in-context learning ability emerges in a large-scale language model such as GPT-3 (Brown et al., 2020) despite no duplication between downstream task data and the pretraining corpus. Indeed, in-context learning uses a natural language description and a few examples to prime a language model. Then the language model can predict the answer of a new example without updating the language model's parameters. Since the release of GPT-3, various large-scale in-context language models have been proposed, including GPT-neo (Black et al., 2021), HyperCLOVA (Kim et al., 2021), and Pangu-α (Zeng et al., 2021).

There still remain many questions on language models' in-context learning capability despite these successful reports. For example, the relationship between the choice of a pretraining corpus and downstream in-context learning task accuracy is unknown. Previous studies argue pretraining with the corpus similar to the downstream task improves the downstream performance, but these observations are often limited to the case where a pretrained language model is finetuned for the downstream task (Gururangan et al., 2020; Lee et al., 2020; Micheli et al., 2020).

In addition, analysis on the relation between the validation perplexity of a language model and in-context learning performance is still less investigated. While previous research on in-context learning implicitly assumes that perplexity is predictive of in-context learning performance, by showing scaling law property of their model (Kaplan et al., 2020; Brown et al., 2020; Kim et al., 2021), their explicit correlations are less discovered.

Motivated by this lack of in-depth analysis on the relationship between in-context learning and corpus properties, we vary the sources and sizes of pretraining corpora and analyze their impact on in-context learning, using a Korean GPT-3 (Kim et al., 2021). We use HyperCLOVA corpus, which is a large-scale pretraining corpus mainly in Korean collected by Kim et al. (2021), as a base corpus from which we derive pretraining corpora for our experiments. Our major findings include:

- **Corpus Source:** The in-context learning performance depends heavily on corpus sources, and with some sources, in-context learning does not work effectively. For example, the model trained only on `Blog` corpus achieves competitive in-context learning performance, but training on `Cafe` or `News` corpus hardly yields in-context learning ability.

- **Corpus Combination:** In-context learning

ability can emerge by fusing two corpora, even when each on its own does not result in in-context learning. For example, while training only on `KiN` corpus, which consists of QnA websites, or `Ency` corpus, which consists of Encyclopedia websites, in-context learning ability was not observed, but training on both corpora makes the emergence of in-context learning.

- **Domain Relevance:** Pretraining with a corpus related to a downstream task is not indicative of the competitive in-context learning performance. For example, training on only `News` corpus does not even yield in-context learning ability on a News-related downstream task, e.g., news topic classification based on its title, KLUE-YNAT (Park et al., 2021).

- **Perplexity:** Although perplexity and in-context learning accuracies correlate well when training a single model, perplexity alone does not reflect the difference in in-context learning accuracies across different language models, particularly when they were trained using different pretraining corpora. For example, `Cafe` model, the model trained with `Cafe` corpus, has the second lowest validation perplexity on various domain sources after `Blog` model, but fails to emerge in-context learning.

2 Related Work

2.1 In-context Learning

Brown et al. (2020) demonstrate the concept of in-context learning, where a few training examples and/or task descriptions are provided together with a new input to a large-scale LM for it to produce a target of this input, without requiring any parameter update. Since then, a few follow-up studies have tried to improve the in-context learning ability (Zhao et al., 2021; Holtzman et al., 2021). Compared to a more conventional approach based on finetuning, such an attempt with in-context learning has a benefit of reduced cost in AI production, as discussed by Bommasani et al. (2021); Kim et al. (2021).

2.2 Domain Relevance on Pretraining Corpus

Previous studies argue a better downstream accuracy is observed with a pretraining corpus more similar to the downstream task corpus (Gururangan et al., 2020; Lee et al., 2020; Micheli et al., 2020). For example, Gururangan et al. (2020) conduct an intensive study and show that further pretraining with the corpus related to the downstream task results in higher downstream task performance. However, these observations are limited to the case where a pretrained language model is finetuned for the downstream task.

There are a few studies on the relationship between effects of different corpus on the relationship between pretraining and in-context learning. A notable example is Codex, where GPT-3 is trained on Github corpus so that the model can generate code from comments (Chen et al., 2021a). However, the corpus used for Codex is limited to code comments and the corresponding code. We study the effect of pretraining corpus on in-context learning performance using various domains.

2.3 Quantity and Quality of Pretraining Corpus

There have been several studies on the quantity and quality of pretraining data. Raffel et al. (2020) conduct an ablation study on different pretraining corpus on T5, and their filtered C4 corpus makes T5 perform better in downstream tasks. As with GPT-3, researchers generally improve the quality of their language model through data filtering (Brown et al., 2020; Kim et al., 2021). Our research differs from the existing work in that we focus on an in-depth analysis of how the amount of data and the corpus source affect in-context learning.

2.4 Multi-task Learning

Multi-task learning or meta-learning approaches, which explicitly finetune on the in-context learning objective by using numerous NLP tasks, are proposed recently to tackle zero/few-shot transfer to the unseen task at test time (Wei et al., 2021; Sanh et al., 2021; Chen et al., 2021b; Min et al., 2021). Unlike the studies in finetuning paradigm, many properties of the in-context learning related to pretraining corpus are still unknown.

3 Task Definition

3.1 Model

We use variants of Korean GPT-3 with different sizes and pretraining corpus. We mainly experiment with the models with 1.3B parameters, but we also include the result for 6.9B-sized models.

Name	Description	Tokens
Blog	Blog corpus	273.6B
Cafe	Online community corpus	83.3B
News	News corpus	73.8B
Comments	Crawled comments corpus	41.1B
KiN	Korean QnA website corpus	27.3B
Modu	Collection of five datasets	6.0B
Ency	Encyclopedia corpus	1.7B
Others	Other corpus	55.0B
Total		561.8B

Table 1: Descriptions of HyperCLOVA corpus (Kim et al., 2021).

All models have a maximum sequence length of 2,048, and use the same vocabulary for all the experiments. We use the morpheme-aware byte-level BPE tokenizer trained with HyperCLOVA corpus (Kim et al., 2021) for all models. We train multiple models with different portions of HyperCLOVA corpus to investigate the effects of the source and size of corpus on in-context learning ability.

3.2 Pretraining with Different Corpus

We analyze the effect of seven subcorpora in the HyperCLOVA corpus: Blog, Cafe, News, Comments, KiN, Modu, and Ency. Table 1 summarizes characteristics of the subcorpora. Blog, Cafe, and News are taken from blog, community sites, and online news articles of a Korean web portal service, respectively. Comments is the comment threads related to the three subcorpora mentioned above. KiN comes from a Korean online community QnA service similar to Quora. Ency is a collection of encyclopedic texts including Korean Wikipedia. Modu consists of five public datasets constructed by National Institute of the Korean Language[1], including 3.2B of news, 2.1B written language, 0.4B of spoken language, 0.2B of web corpus, and 0.02B tokens of messenger. Others was excluded to investigate the explicit effects of domain corpus sources on in-context learning because Others is the corpus where various subcorpora are taken from multiple heterogeneous sources. Table 7 and 8 in Appendix show example instances of seven pretraining corpus in Korean and English, respectively. ALL denotes the original HyperCLOVA corpus including Others.

For corpora with less than 150B tokens, We assign 99% of each corpus to the pretraining corpus and randomly extract 10,000 examples from the remaining 1% to the validation corpus for measuring validation perplexity. For corpora with more than 150B tokens, we make the training corpora 150B tokens via random sampling and construct a validation set with 10,000 examples randomly sampled from the remaining part. As a result, the maximum training set size of each corpus is 150B tokens and the entire validation set contains 70,000 examples for seven domains.

3.3 Downstream Tasks

We evaluate in-context learning performance of each corpus-specific model on four Korean downstream tasks used in Kim et al. (2021): NSMC[2], KorQuAD (Lim et al., 2019), AI Hub translation[3], and YNAT (Park et al., 2021). NSMC is a binary sentiment classification dataset on movie review. KorQuAD is a machine reading comprehension dataset similar to SQuAD 1.0 (Rajpurkar et al., 2016). AI Hub translation dataset consists of Korean-English parallel sentences from news, government websites, legal documents, etc. YNAT is a topic classification problem with seven classes.

It is noticeable that three downstream tasks are closely related to the HyperCLOVA corpus. Passages to construct KorQuAD are taken from Korean Wikipedia, which is also a part of the Ency. YNAT is a topic classification task of news headlines, so the downstream task is deeply related to the News corpus. A significant portion of parallel sentences for AI Hub translation dataset also comes from news articles. KiN corpus is also related to the translation task. About 2.5% of QnA data in KiN include Korean questions on the English language, as a foreign language. These question-style passages often include Korean-English sentence pairs in the passage. Vocabulary overlap between downstream tasks and HyperCLOVA corpus is depicted in Figure 1.

3.4 Experimental Details

We try our best to make the same Hyperparameter of Kim et al. (2021), including global batch size, training step, maximum sequence length, learning rate, and so on. In our experiments, the models are trained for 72K steps with a global batch size of 1,024. We note that under this setting, the number of tokens that were actually used in pretraining is 150B. Therefore, we set the maximum size of

[1] https://corpus.korean.go.kr/

[2] https://github.com/e9t/nsmc

[3] https://aihub.or.kr/aidata/87

training corpus to 150B as mentioned in Section 3.2.

In most experiments, validation perplexity decreases monotonically as training goes on. Thus, we use the checkpoint at 72K step. The only exception is the `Ency` model. The `Ency` model has a minimum validation loss at 12K steps, which is likely to be caused by overfitting to pretraining data due to a small size of the data. Therefore, we use early-stopping checkpoints at the 12K steps for the report.

For optimization, AdamW (Loshchilov and Hutter, 2019) with the learning rate of 2.0e-4 and cosine learning rate scheduling was used. We use mixed precision training. Models are trained on the Nvidia Superpod which consists of 1,024 A100 GPUs spread across 128 nodes. Using Superpod, it spend around 18 hours to train 1.3B model with 72K steps.

For classification tasks such as NSMC and YNAT, we use a rank classification approach (Wei et al., 2021), where we compare pre-defined outputs ("positive" and "negative") and take the higher probability one. KorQuAD and AI Hub are free-form completion tasks, where we directly generate output tokens using greedy decoding.

The number of shots in the few-shot experiments is set to fit the maximum sequence length. As a result, 70, 4, 4, and 70 shots are used for NSMC, KorQuAD, AI Hub, and YNAT, respectively. Downstream tasks are performed 12, 1, 3, and 6 times with different random seeds for NSMC, KorQuAD, AI Hub, and YNAT, respectively. We report the average performance. Random seed influences the sampling of shots from training data and their order. The reason KorQuAD has only one random seed is described in Appendix C. These all experimental setting from the number of shots and number of random trials basically comes from the experimental setting of Kim et al. (2021). However, we change the number of trials of YNAT from 3 to 6, because we found standard derivation of YNAT is relatively high.

Tables 9, 11, 13, and 14 in Appendix C show the examples of the few-shot prompts of NSMC, KorQuAD, AI Hub, and YNAT, respectively. Tables 10, 12, and 15 in Appendix C present the translated version for NSMC, KorQuAD, and YNAT.

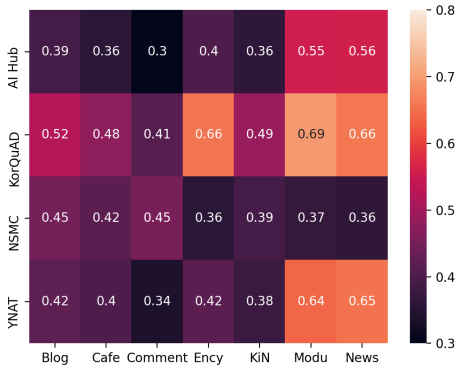

Figure 1: Vocabulary overlap ratio between pretraining corpus and downstream task. Top 1,000 nouns are used to calculate the ratio. Nouns are extracted using in-house part-of-speech tagger.

3.5 Measuring Validation Perplexity

We report validation perplexity in various tables and figures to validate our argument. We use the term "PPL" to denote validation perplexities from the validation set. The validation set consists of 70,000 examples from seven corpus sources, as described in 3.2. We emphasize that, for calculating PPL, all experiments use **the same vocabulary and the same validation set**.

In Tables 2 and 3, we use *Italic* font for the results from a multi-domain model, which is pretrained with two or more corpora. Since multi-domain model trains more domains than a single-domain model, the PPLs of multi-domain models are generally lower than those on single-domain models. To keep readers from directly comparing PPLs between a single-domain and a multi-domain models, we use italic font for the results from a multi-domain model.

4 Experimental Results

We perform intensive experiments to answer these main questions:

1. How large do the source and the size of pretraining corpora have effects on emerging in-context learning ability? (Section 4.1 and 4.2)

2. What is the effect of combining various corpora? (Section 4.3)

3. How large does domain relevance of corpus influence on model performances of the downstream task? (Section 4.4)

Model	Corpus Train	PPL	NSMC (Acc)	KorQuAD (EM)	KorQuAD (F1)	AI Hub (BLEU) Ko→En	AI Hub (BLEU) En→Ko	YNAT (F1)
Majority	-	-	50.35	0.0	0.0	0.0	0.0	8.26
ALL	150B	*119.99*	84.59(1.25)	56.17	73.47	6.15(0.16)	23.36(0.33)	59.57(4.30)
ALL w/o Others	150B	*119.66*	84.59(1.25)	56.49	74.20	6.14(0.21)	23.21(0.45)	50.76(11.81)
Blog	150B	152.40	**83.50**(2.45)	**50.74**	**69.34**	**3.82**(0.10)	**20.11**(0.79)	**60.68**(5.75)
Cafe	82.5B	170.85	57.77(10.28)	3.12	14.26	2.83(0.19)	**16.53**(0.63)	11.04(4.43)
News	73.1B	234.78	50.72(0.56)	0.14	9.96	1.10(0.78)	**15.88**(0.84)	14.36(2.38)
Comments	40.7B	225.39	**79.78** (2.38)	14.69	33.33	0.79(0.01)	5.06(0.16)	**36.17**(3.31)
KiN	27.0B	187.80	54.73(4.26)	4.85	18.99	**6.81**(0.88)	**18.16**(0.71)	9.23(1.96)
Modu	5.9B	226.01	**69.91**(8.41)	**30.20**	**49.29**	1.21(0.06)	6.13(0.39)	**43.27**(6.72)
Ency	1.7B	549.40	53.81(2.22)	0.71	11.88	0.58(0.16)	0.69(0.49)	27.99(2.38)
Blog 54B	54.0B	155.69	**83.06**(2.26)	**49.13**	**68.10**	**3.93**(0.20)	**21.12**(0.19)	**57.97**(5.72)
Blog 27B	27.0B	165.60	**80.27**(2.28)	10.91	23.41	**5.35**(1.95)	**12.32**(5.68)	**48.19**(6.71)
Cafe 27B	27.0B	169.81	49.91(0.29)	1.37	13.98	**4.25**(0.35)	**20.74**(1.57)	8.60(2.79)
News 27B	27.0B	239.79	50.64(3.26)	0.80	8.02	2.42(1.41)	**15.78**(4.21)	27.20(5.86)
Comments 27B	27.0B	229.65	**80.50**(1.44)	13.02	31.53	1.70(0.03)	3.28(0.16)	25.79(7.27)

Table 2: In-context few-shot learning performance with different pretraining corpus. Models with 1.3B parameters are used. Standard derviation exists in parentheses. Majority means classifying each label with the primary class, and its score is 0 for KorQuAD and AI Hub. Purple-underline denotes the score is below the mean performance value of ALL and Majority baseline, and **Teal-bold** denotes the score is above. We use this mean value of Majority and ALL as the performance basis to prevent the in-context learning performance of each model from being distorted by the high basis performance of two classification tasks such as NSMC and YNAT.

4. How strong is the correlation between validation perplexity and in-context learning of language models? (Section 4.5)

4.1 Effect of Corpus Source

Table 2 shows the experimental results on various pretraining corpus sources.

It is noticeable that in-context learning ability emerges differently depending on pretraining corpus sources. For example, Blog model makes competitive in-context learning performance to ALL model, while each of Cafe and News models hardly shows in-context learning ability from Table 2. It is also noticeable that Modu model performs better than Cafe and News model although the size of Modu corpus is less than 1/10 of Cafe or News corpus, showing the corpus size is not the only factor to predict in-context learning performance. Likewise, it is also interesting that Cafe+News model also shows poor performance despite the same size to Blog and ALL, as shown in Table 3.

4.2 Effect of Corpus Size

Table 4 shows the experimental results of ALL model on various pretraining corpus sizes. 56B and 6B correspond to the 1/10 and 1/100 of the original HyperCLOVA corpus with 560B tokens, respectively. The 56B tokens and 6B tokens models are trained with around 3 and 25 epochs, respectively, so that both models can be trained with 72K training steps.

Results show that reducing the corpus size from 150B to 56B does not decrease the performance severely despite training with 1/10 of corpus. However, the performance degradation of 6B tokens model is remarkable comparing to ALL model. Nevertheless, it is noticeable that 6B tokens model still performs much better than Cafe+News model, which trains 150B tokens of Cafe and News corpus.

We can also see the case of three sizes of Blog corpus in Table 2. Blog and Blog 54B show achieves similar performance. However, like in ALL 6B, Blog 27B performs quite worse than Blog 54B.

Figure 3 shows the comparison between 1.3B-sized model and 6.9B-sized model. While a larger model presents better performance than a smaller model on KorQuAD and YNAT, there is no meaningful accuracy difference on NSMC.

4.3 Effect of Combining Corpora

One of our main goals is to investigate the effects of combining multiple corpora from various sources on in-context learning performance. Table 3 includes the experiments on models trained from a corpus merged with multiple source corpora. The

Corpus Type	Corpus Train	PPL	NSMC (Acc)	KorQuAD (EM)	KorQuAD (F1)	AI Hub (BLEU) Ko→En	AI Hub (BLEU) En→Ko	YNAT (F1)
ALL	150B	*119.99*	84.59(1.25)	56.17	73.47	6.15(0.00)	23.36(0.00)	59.57(4.30)
The Case where In-context learning Emerges by Combining Two Poor Corpora								
KiN+Ency	28.7B	*164.69*	59.17(9.34)	**42.09**	**61.00**	**8.99**(0.31)	**23.12**(0.40)	**42.84**(9.01)
Cafe+KiN	109.5B	*141.92*	**76.42**(4.68)	**38.45**	**59.00**	**8.41**(0.68)	**23.41**(0.38)	**56.96**(5.79)
The Case where In-context learning Does Not Emerge by Combining Two Poor Corpora								
Cafe+News	150B	*154.20*	54.15(3.90)	22.86	22.72	**4.45**(0.12)	**17.77**(2.19)	8.19(5.32)
The Case of Combining In-context Emerging Corpora								
Blog+Comments+Modu	150B	*144.67*	**82.82**(1.93)	**54.94**	**72.27**	**4.09**(0.16)	**21.17**(0.43)	**65.01**(2.90)
The Case of Adding News into KiN+Ency to Try to Enhance the Performance of YNAT								
News+KiN+Ency	150B	*142.13*	**75.96**(5.94)	**35.42**	**55.60**	**8.70**(0.45)	**23.38**(0.18)	27.54(7.46)

Table 3: In-context learning performance with different corpus combination. Purple-underline denotes the score is below the mean performance value of ALL and Majority baseline in Table 2, and **Teal-bold** denotes the score is above. 1.3B-size models are used.

# of tokens	NSMC (Acc)	KorQuAD (EM)	AI Hub (BLEU) Ko→En	AI Hub (BLEU) En→Ko	YNAT (F1)
150B	84.59	56.17	6.15	23.36	59.57
56B	84.35	55.13	5.47	22.98	51.89
6B	74.70	36.72	3.97	17.81	30.24

Table 4: In-context learning performance of ALL with different size of the pretraining data. The dataset is randomly sampled from the original corpus.

result shows that in-context learning ability can be emerged by combining two corpora, even if each of both corpora cannot provide in-context learning ability. For example, KiN+Ency model succeeds to make in-context learning ability in most tasks, while each of KiN and Ency fails in most tasks. Likewise, Cafe+KiN model succeeds to make in-context learning ability, while each of Cafe and KiN fails in most tasks.

This phenomenon is related to the argument that in-context learning emerges by multi-task learning. According to the argument, as the language modeling objective function requires a language model to learn variety of next word prediction tasks, the generalization pushes in-context learning ability on unseen tasks. In the example of KiN+Ency model, KiN+Ency learns in-context learning ability of MRC task, by learning next word prediction tasks of both Ency (Wikipedia) and KiN (QnA).

Unlike these positive cases, we observe that combining corpora does not assure the emergence of competitive in-context learning from the case of Cafe+News. Even if the mixed corpus model shows slightly better performance on KorQuAD than each of two corpora, its performances on NSMC, KorQuAD, and YNAT are still below the basis. Furthermore, the performances on NSMC and YNAT even decrease.

4.4 Effect of Domain Relevance

Table 2 also shows that the close relationship between a pretraining corpus and a downstream task does not always guarantee in-context learning ability on the downstream task. KiN and Ency do not perform well on KorQuAD task, although KorQuAD is an MRC task from Korean Wikipedia, Ency includes Korean Wikipedia, and KiN consists of question answering pair, respectively. Likewise, News does not perform well on YNAT task, although YNAT consists of news headline queries. Table 3 further shows that News+KiN+Ency model shows more degenerated F1 score on YNAT than KiN+Ency, even though a large amount of News corpus is added to News+KiN+Ency model.

For further investigation, we analyze vocabulary statistics of each corpus. Figure 1 shows the vocabulary overlapping ratio between pretraining corpora and downstream tasks. The result shows that high vocabulary overlap between a pretraining corpus and a downstream task does not indicate high downstream task performance. Although the Modu corpus has a large vocabulary overlapping ratio to AI Hub, in-context learning performances of the Modu model on the translation task are much lower than Blog and KiN.

The counter example of above supports is AI

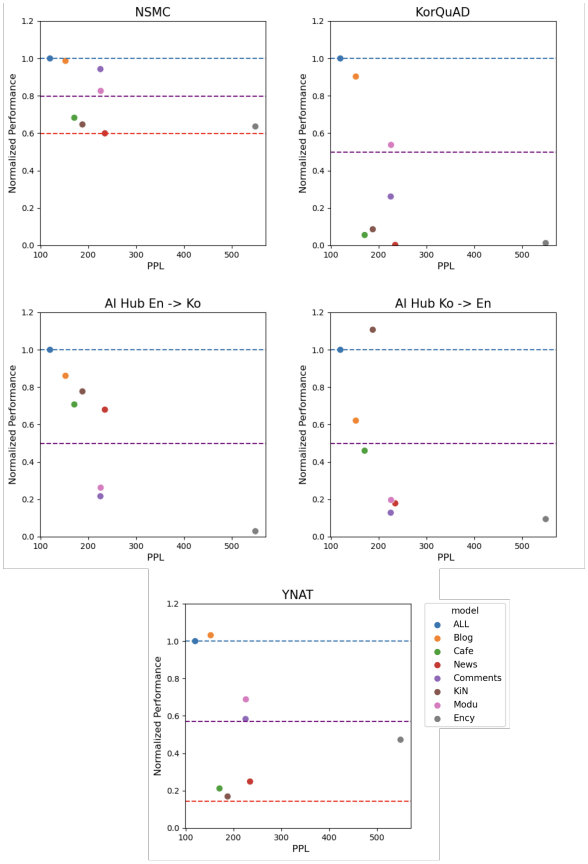

Figure 2: In-context learning performance of various corpus models and their PPL. A score of the model is divided by that of `ALL` to calculate the normalized performance. Blue and red lines denote the performance of `ALL` model and majority baseline, and the purple line represents the average of both defined in the caption of Table 3.

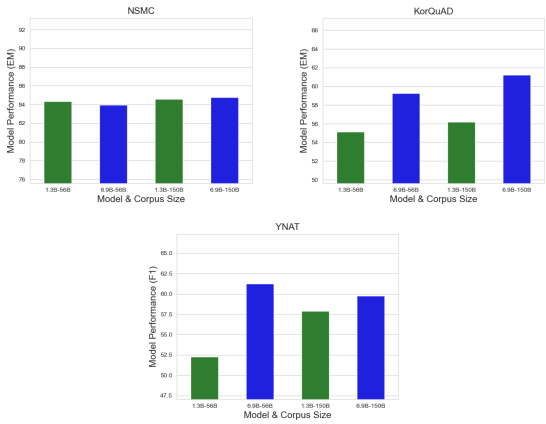

Figure 3: Comparison on two model sizes and two corpus sizes of the original HyperCLOVA corpus such as 1.3B-sized model and 6.9B-sized model, and 56B and 150B tokens. Green rectangle denotes 1.3B-sized model and blue rectangle denotes 6.9B-sized model.

Hub task performance of `KiN` model. `KiN` model learned the pattern of Korean-English sentence pairs, since the corpus includes a lot of Korean questions on English language. While `KiN` model does not work well in other downstream tasks, the performance on AI Hub translation is competitive and makes the best performance in Ko→En among seven pretraining corpora.

We carefully conjecture that pretraining significantly improves in-context learning performance when both query and answer related to the downstream task exist in the pretraining corpus document. Additional analysis is required to clearly discover this hypothesis.

4.5 Perplexity and Downstream Task

Figure 2 presents the scatter plots of PPL (x-axis) and in-context learning performance (y-axis) on five downstream tasks for single corpus models and the `ALL` model. In Figure 2, we normalized in-context learning performance by dividing `ALL` model performance for calibrating various task metrics. Because we observe less explicit tendency of correlation between validation perplexity and in-context performance, we argue that it is difficult to hypothesize better perplexity assures emerging of in-context learning ability.

According Table 2, `Blog` model shows both lowest PPL and best in-context learning performance, and `Ency` model shows both highest PPL and worst in-context learning performance. On the contrary, while `Cafe` model and `KiN` model shows the second and third lowest PPL, in-context learning ability was not observed. These results show that perplexity does not serve as a strong predictor of in-context learning performance in comparing models trained using different corpora. Table 2 also shows that corpus size affects in-context learning performance more than PPL. `Blog` 27B performs quite worse than `Blog`, but PPL relatively does not decrease much.

Table 5 shows validation perplexity scores per each subcorpus. One row denotes the model and one column denotes the validation set's subcorpus. Each validation set except All in Table 5 consists of 10,000 instances, and is the part of our main validation sets, consists of 70,000 instances. Figure 5 in Appendix A, supporting Table 5, also shows the validation perplexity of different models from different subcorpora.

On the other hand, Figure 4 shows that PPL and in-context learning performance correlate well

	Blog	Cafe	News	Comments	KiN	Modu	Ency	All
`Blog`	*126.89*	**201.15**	**83.98**	599.68	**138.95**	98.83	**108.17**	**152.40**
`Cafe`	**168.03**	*135.37*	107.78	**596.27**	163.94	124.71	142.15	170.85
`News`	281.33	432.90	*60.21*	1543.03	253.73	**87.65**	156.28	234.78
`Comments`	228.37	242.59	176.28	*390.30*	164.88	196.40	239.40	225.39
`KiN`	232.13	278.89	150.78	689.89	*50.06*	172.78	141.45	187.80
`Modu`	267.59	411.35	84.36	1086.04	243.19	*69.48*	136.02	226.01
`Ency`	841.53	1348.38	213.87	5889.79	543.69	266.11	*73.06*	549.40

Table 5: Validation perplexity scores per each subcorpus. All denotes the validation perplexity on our main validation set from seven corpus sources. *Italic* font denotes the validation PPL of their corpus domain, and **Bold** denotes second best after own corpus. Overall, `Blog` has the best overall validation perplexity in most tasks.

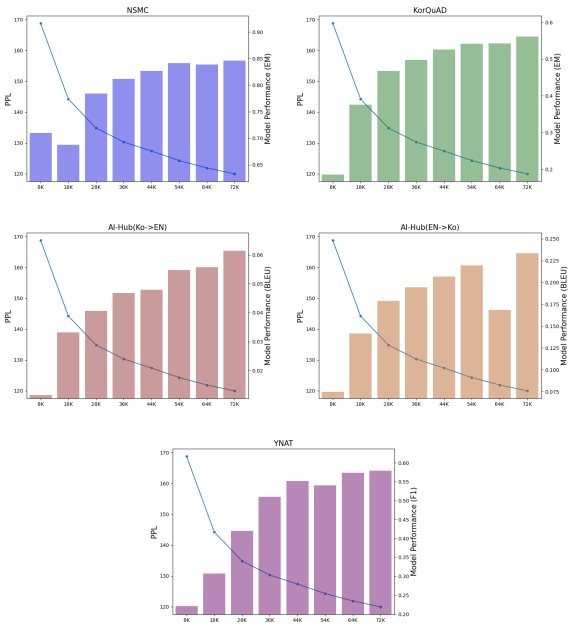

Figure 4: Relation between validation PPL and in-context learning performance for five downstream tasks as pretraining steps proceed.

when training a single model. We can find that the correlation trends between the cases in the training and the cases between the corpus domain are different.

5 Discussion

Our knowledge can be used in terms of increasing the performance of in-context learning where their corpus is small, and demand exists for collecting more corpus. In the case of XGLM (Lin et al., 2021), which is a concurrent work on multilingual GPT-3, achieved better in-context learning performance for many languages. However, it does not reach the performance of a single language model. We hope our observation can give insight into what types of pretraining to be collected more, both for multilingual model and low-resource language model.

Another notable example comes from Gopher (Rae et al., 2021), which is a concurrent work on state-of-the-art in-context learner. Rae et al. (2021) determine the ratio between subcorpora based on the perplexity of the validation corpus. They implicitly claim that this ratio results in better downstream task performance, but do not address explicit evidence for this. On the other hand, we are in a position to doubt the strong correlation between perplexity and in-context learning. We hope our findings contribute to making better in-context learners along with other research.

6 Conclusion

This paper investigates the effects of the source and the size of the training corpus on in-context learning ability, using the HyperCLOVA corpus. Our discoveries include that corpus sources play a crucial role in whether or not in-context learning ability will emerge in a large-scale language model. Interestingly, in-context learning ability emerges when a model was trained with the two corpora that do not incur in-context learning ability when used alone. In addition, our analyses reveal that larger pretraining data and vocabulary overlapping between a pretraining corpus and a downstream corpus do not always lead to a better in-context language model.

One direction for future work is to investigate linguistic properties of corpus sources which make a competitive in-context learning model. For example, quantifying the difference between two corpora can shed light on how to select suitable corpora for NLP practitioners who build large-scale language models. In addition, intensive studies on different corpus sources other than the HyperCLOVA corpus can help understand the properties of in-context learning.

Broader Impact Statement

We present multiple pieces of evidence that models using only a part of the pretraining corpus are comparable with those trained with the entire corpus in terms of in-context performances. Although we leave the validation on larger-scale models, such as tens of billion parameters, to future work, our analysis presents a hint to effectively training LMs with smaller corpora. This approach can contribute to alleviating severe energy consumption issues caused by large-scale LMs.

Meanwhile, our study relates to the misuse and fairness of large-scale LMs. For example, reweighting domain-specific corpus might cause LMs to be biased inherent in the domain corpus. Therefore, alleviating domain corpus bias would be a valuable future direction.

References

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

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

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2021b. Meta-learning via language model in-context tuning.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. 2021. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. In *EMNLP*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1.0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *EMNLP*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits

of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Jason Wei, Maarten Bosma, Vincent Y. 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.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *ArXiv*, abs/2102.09690.

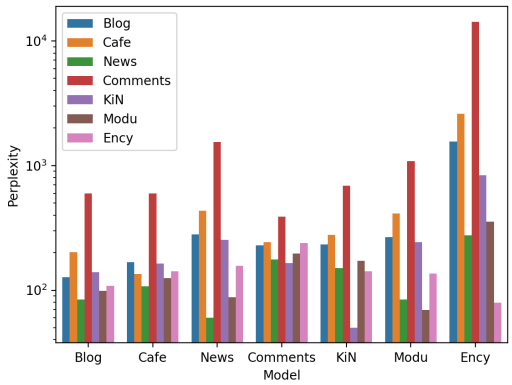

Figure 5: Validation perplexity of different 1.3B-size models in log scale. Color indicates the source of validation corpus.

A Details on Experimental Results

Table 6 shows score difference with `ALL` in addition to in-context learning scores on Table 2. Figure 5 shows the validation perplexity of different model from different corpus.

B Details on Pretraining Corpus

Tables 7 and 8 show example instances of seven pretraining corpus in Korean and English, respectively.

For preprocessing steps of our pretraining corpus, we use HyperCLOVA corpus which is also used in (Kim et al., 2021) as described in Section 3.1. Therefore, we share the preprocessing steps of Kim et al. (2021). Appendix A in (Kim et al., 2021) describes their preprocessing methods on data descriptoin, data clearning, data anonymization, and data postprocessing.

B.1 Deduplication Preprocess

We additionally introduce the deduplication preprocess of HyperCLOVA corpus, which is used in (Kim et al., 2021). The deduplication preprocess was applied to construct HyperCLOVA corpus to prevent explicit duplication within and between subcorpora (Kim et al., 2021). According to the response of Kim et al. (2021), they use an in-house search engine and an in-house engineering trick to detect document pairs that are very similar to each other. There are two pipelined steps: (1) removing duplicates within subparts of the corpus, and then (2) removing duplicates between subparts of the corpus. Therefore, documents with high overlap do not exist throughout the documents. Here, the number of subparts is 29. These 29 subparts are categorized into the eight domains we deal with in the paper (i.e., `Blog`, `News`, `Cafe`, `Comments`, `KiN`, `Modu`, `Ency`, and `Others`). Overall, there is no explicit overlap between each corpus, since very similar documents have already been removed from the corpus. The overlap between eight HyperCLOVA subcorpora is quite small. There were many overlaps within the subpart of the corpus. However, the overlap between subparts of the corpus was only 0.024% of the total, according to the counts in the second pipelined step of deduplication between subparts.

C Examples of Few-shot Prompt

Tables 9, 11, 13, and 14 show the example few-shot prompt of NSMC, KorQuAD, AI Hub, and YNAT, respectively. Tables 10, 12, and 15 show the translated version for NSMC, KorQuAD, and YNAT, respectively.

On the other hand, the number of random seed is one for KorQuAD. We explain why evaluation on KorQuAD with many random seeds is difficult, from the perspective of prompt design. The way we make randomness on trials to change few-shot examples. However, in the case of Kim et al. (2021) and in our case, there are no alternative examples to put into the prompt. The prompt examples of KorQuAD are one document and a few question-answer pairs, and not a few document-question-answer triples. In other words, in the prompt of KorQuAD, the number of the document is one. Thus, the document is used for both few-shot question-answer pairs and a query question for the inference. In KorQuAD, there are five corresponding question-answer pairs in each document. In the experimental setting of ours and Kim et al. (2021), four question-answers are put into the prompt and one question is used for the test. Therefore, there are no other question-answer pairs to replace the four pairs.

D Generalization to Other Languages

Someone can ask whether our results can be extended to other language, including English. We have left experiments on non-Korean language as future work. However, we describe some explanations below to defend our experiments on the Korean language and to discuss why experiments on other languages are practically non-trivial.

First, we think our findings are basically generalizable to other languages. From the perspective

Corpus Type	Corpus Train	NSMC (Acc)	KorQuAD (EM)	KorQuAD (F1)	AI Hub (BLEU) Ko→En	AI Hub (BLEU) En→Ko	YNAT (F1)
ALL	150B	84.59	56.17	73.47	6.15	23.36	59.57
The Case where In-context learning Emerges by Combining Two Poor Corpora							
KiN+Ency	28.7B	59.17(-25.42)	**42.09(-14.08)**	**61.00(-12.47)**	**8.99(+2.84)**	**23.12(-0.24)**	**42.84(-16.73)**
Cafe+KiN	109.5B	**76.42(-8.17)**	**38.45(-17.72)**	**59.00(-14.47)**	**8.41(+2.26)**	**23.41(+0.06)**	**56.96(-2.61)**
The Case where In-context learning Does Not Emerge by Combining Two Poor Corpora							
Cafe+News	150B	54.15(-30.44)	22.86(-33.31)	22.72(-50.75)	**4.45(-1.70)**	**17.77(-5.59)**	8.19(-51.38)
The Case of Combining In-context Emerging Corpora							
Blog+Comments+Modu	150B	**82.82(-1.77)**	**54.94(-1.23)**	**72.27(-1.20)**	**4.09(-2.06)**	**21.17(-2.19)**	**65.01(+5.44)**
The Case of Adding News into KiN+Ency to Try to Enhance the Performance of YNAT							
News+KiN+Ency	150B	**75.96(-8.63)**	**35.42(-20.75)**	**55.60(-17.87)**	**8.70(+2.55)**	**23.38(+0.02)**	27.54(-32.03)

Table 6: Table 3 which includes the difference from ALL in parentheses.

of pretraining and in-context learning, fundamental differences between Korean and English were limitedly reported. For example, XGLM (Lin et al., 2021), a concurrent work on, also does not show critical evidence on language-specific properties.

Second, It is non-trivial to control various aspects of corpora for our purpose. Most corpus for in-context few-shot learners comes from crawled web site which is not easy to distinguish from its original source. For example, 82% of OpenAI GPT-3 Corpus (Brown et al., 2020) is a filtered version of Common Crawl. In this regard, we used relatively a well-refined corpus which consist of several subcorpus from a single web service. (Please see also Section B.1 of this letter.) On the other hand, we have interests to extend our work onto Pile dataset (Gao et al., 2020), by controlling the subcorpora in the direction our study pursuits, in the future.

	Example
Blog	블로그 제목: 촬영하러 온 꼬맹이들^^ 본문: 엄마 회사에 오늘은 모델로 일하러온 꼬맹이들. 신나게 놀고 까불고 뛰어다니다가 책보더니 잠잠해진다. 조용해진 아이들 보고 놀라는 스탭들. 순간 엄마얼굴엔 미소가! ㅋㅋ. 책 잘 읽는 아이들이라 자랑스러움이 잠시 ㅋㅋ 그러다가 촬영하고 생각보다 잘해줘서 고맙네. 엉망진창으로 못할줄 알았는데 카메라를 보다니. ㅋ 어린시절부터 카메라를 본 경험이 빛을 발하긴 하나보다. ㅋㅋ 하여간. 엄마 회사에 모델로 와준 꼬맹이들. 고마워^^. 좋은 추억이 되었길.
Cafe	카페 제목: 탐스우먼 상자채 새신발(직구한것보다 싸게 내놓아요 ^^) 본문: 벼룩시장(중고), 판매중, 가격 1원, 안전거래 미사용, 탐스클래식, 판 매 양 식아이디 이메일싸이,블로그,타카페,타사이트 링크시 삭제 및 강퇴 거주지역도,시,동까지 정확히 기재 판매 제품명 구입시기년,월 기재 희망가격정확히 기재: (3만 4만등의 경매 유도글 삭제) 거래방법 직거래, 택배, 안전거래 상세설명 탐스 공홈에서 직구했는데사이즈가 커서 내놓아요 빨리 팔려구 직구한것 보다 싸게 내놓아요 ^^ 1. 탐스우먼 유니버시티 애쉬 그레이 택포 45,00 2. 탐스우먼 초코캔버스 택포 40,000 많은 문의 부탁드려요 ^^
News	뉴스 제목: 전명환, 이병기 시문학상 수상 본문: '2016 이병기 청년시문학상ㆍ최명희 청년소설문학상' 수상자가 결정됐다. 지난 1일 전북대 총장실에서 시상식을 연 가운데 이병기 청년시문학상 대학 부문에는 '대과거'를 쓴 전명환(중앙대 국어국문 2년), 고등 부문에는 '몽상'을 선보인 황주연(경산여고 2년) 이 선정됐다. 최명희 청년소설문학상 대학 부문에는 '꽃에서부터'를 쓴 윤선미(서울디지털대 문창 3년), 고등 부문에는 '야간비행'을 쓴 윤정은(안양예고 2년)이 수상의 영예를 안았다. 전북대학교(총장 이남호) 신문방송사와 혼불기념사업회ㆍ최명희문학관(대표 장성수)이 공동으로 주관하는 공모전에는 올해 시 부문 167명 669편, 소설 부문 108명 116편이 출품됐다. 시 부문 심사는 최승범 양병호 유인 이승철 위원이, 소설 부문 심사는 서철원 황보윤 장마리 김소윤 최기우 위원이 맡았다. 박준호 문학상 운영위원장 및 신문방송사 주간은 "수준 높았으며 시대를 바라보는 청년들의 녹록치 않은 고민과 생각을 엿볼 수 있었다"고 평했다.
Comments	대화 본문: 하루를 엄청 길게 사용하시네요^^ 점심은 더 많이 드세요 아점입니다ㅎㅎ 저녁을 기다려야죠^^ 이렇게 드시고 무슨 운동까지 하십니까?? 저녁 윗몸일으키기는 빼세요 ㅋㅋㅋ요즘 가끔 빼먹습니다..저 담주 월.화중에 앤더슨님 방문할까합니다..같이가시죠? 다음주 월,화요??^^ 가야죠,,갑니다..시간을 만들어서라도 가야죠^^ 어찌 움직이실건지요?? 화요일날로..저는 전철을타야해서 무찌르자님은 어디서 출발하시는지요 전 서울 성수동에서 출발합니다.. 성수까지만 오시면 제가 모시겠습니다..^^
KiN	질의응답 질문: 독실라게???? 사투리라는데 독실라게 가 뭔뜻인가요? 경상도 쪽이라는데. 본문: 경상도 방언에서는 엄청나게 .억수로 강조하는 부사입니다
Modu	뉴스 본문: 춘분에 눈 내린 부산...강풍까지 불며 피해 속출눈이 잘 오지 않는 부산에 춘분인 21일 0.7cm의 눈이 내려 산간지역 도로가 통제되는 등 피해가 잇따랐다고 연합뉴스가 보도했다. 부산기상청에 따르면 이날 부산의 아침 최저기온은 공식 관측소가 있는 중구 대청동 기준 1도였다. 해발고도가 500m 이상인 중구 서대신동은 영하 1.4도, 영도구는 영하 0.9도를 기록했다. 강한 바람까지 불면서 체감온도는 영하 2.6도까지 떨어졌다. 아침 최저기온이 영하권을 넘나들면서 밤새 내리던 비가 진눈깨비로 변했다. 부산 기장군에 있는 기상청 적설 자동 관측장비에는 적설량이 0.7cm로 기록됐다. 부산 북구와 사상구, 기장군 등지 산간도로에 눈이 쌓이면서 기장군 곰내재와 이곡 고개, 금정구 산성로, 북구 만덕고개길, 연제구 황령산로 등 8곳의 차량통행이 오전 한때 통제됐다. 관할 지자체 등은 작업 차량을 동원해 염화칼슘을 살포하는 등 제설작업을 벌였다.
Ency	문서 제목: 설악면 본문: 설악면(雪岳面)은 대한민국 경기도 가평군의 면이다. 넓이는 141.53 km²이고, 인구는 2016년 12월 말 주민등록 기준으로 8,986 명이다.설악면은 북한강 남쪽에 있어서 본래 양평군에 속했는데, 1942년 가평군에 편입되었다. 강원도에 있는 설악산(雪嶽山)과는 무관하다.

Table 7: Example document from various domains. Note that Modu consists of 5 different subdomains and the example is taken from the news subdomain, which is the largest.

	Example
Blog	Blog Title: Kids who came to shoot ^^ Body: Kids who came to my mom's company as models today. After having fun, playing around, and running around, they read a book and calmed down. Staff members are surprised to see the quiet kids. A smile on my mom's face! Haha. I'm proud of them because they're good at reading books, and then I took phots and thank them for doing better than I thought. I thought they do mess it up, but I can't believe that they are looking at the camera. The experience of looking at the camera since early childhood must have helped. Anyway. The kids who came to my mom's company as models. Thank you ^^. I hope it was a good memory.
Cafe	Cafe Title: Toms women's shoes (It's cheaper than what I bought directly ^^) Text: Flea market (used), selling, price of 1 won, not used for safety transactions, Tom's Classic, sales form ID Email, blog, other cafe, other site link, city, dong, exact date of purchase of sales product, desired price of 30,000 to 40,000 won, direct auction transaction. Urgent sale and sell it cheaper than what I bought directly. ^^1. Tom's Woman University Ash Gray 45,00 2. Tom's Woman Chocolate Canvas 40,000. Please contact us.^^
News	News Title: Jeon Myeonghwan and Lee Byungki won the Poem Literature Award. Body: The winners of the 2016 Young Poetry Literature Award and Choi Myung-hee Young Novel Literature Award have been decided. While the awards ceremony was held at Jeonbuk National University's president's office on the 1st, Jeon Myung-hwan (second year of Chung-Ang University's Korean Language Language) who wrote "the past" in the college category and Hwang Joo-yeon (second year of Gyeongsan Girls' High School) were selected. Yoon Sun-mi (3rd year of Moonchang, Seoul Digital University), who wrote "From Flowers" in the college category of Choi Myung-hee's Youth Novel Literature Award, and Yoon Jung-eun (2nd year of Anyang Arts High School), who wrote "Night Flight" in the high school category, were honored. The contest, co-hosted by Jeonbuk National University (President Lee Nam-ho) newspaper broadcasters, Honbul Memorial Society, and Choi Myung-hee Literature Museum (CEO Jang Sung-soo), featured 669 works of 167 people in the poetry category and 116 works of 108 people in the novel category this year. Choi Seung-beom, Yang Byung-ho, Yoo Seung-chul, a member of the Yoo In, and Seo Cheol-won, Hwang Bo-yoon, Jang Mari, Kim So-yoon, and Choi Ki-woo, a member of the novel division, were in charge of the screening. Park Joon-ho, chairman of the Literature Award's steering committee and weekly newspaper broadcaster, commented, "It was high-quality, and I could get a glimpse of the difficult worries and thoughts of young people looking at the times."
Comments	Conversation Body: You spend a long day.^^ Eat more for lunch. It's brunch. We have to wait for dinner.^^ What kind of exercise do you do after eating like this? Don't do sit-ups in the evening. I've been skipping it from time to time. I'm going to visit Anderson next Monday and Tuesday.Let's go together, right? Next Monday and Tuesday?^^ I have to go, I'm going...I'll make time to go there.^^ How are you going to move? On Tuesday... I have to take the subway, so where will you leave? I'm departing from Seongsu-dong, Seoul. If you come all the way to Seongsu, I'll take you.^^
KiN	QnA Question: Doksilagae? It's a dialect. What does doksilagae mean? It's used near Gyeongsang-do. Text: In Gyeongsang-do dialect, it means tremendously.It's an adverb to emphasize.
Modu	News Text: It snowed in the spring equinox in Busan...Strong winds are blowing and they're avoiding it. Yonhap News Agency reported that 0.7 centimeters of snow fell on the 21st, the spring equinox in Busan, where snow was not easy, and roads in mountainous areas were controlled. According to the Busan Meteorological Administration, the lowest temperature in the morning in Busan was 1 degree in Daecheong-dong, Jung-gu, where the official observation station is located. Seodaemun-dong, Jung-gu, with an altitude of more than 500m above sea level, recorded minus 1.4 degrees Celsius and Yeongdo-gu recorded minus 0.9 degrees Celsius. As strong winds blew, the sensible temperature dropped to minus 2.6 degrees Celsius. As the morning low temperature crossed below zero, the rain that had been falling all night turned into sleet. The automatic snow observation equipment of the Korea Meteorological Administration in Gijang-gun, Busan recorded a snowfall of 0.7cm. As snow piled up on mountain roads in Buk-gu, Sasang-gu, and Gijang-gun, eight vehicles, including Gomnaejae Pass, Igok Pass, Geumjeong-gu Sanseong-ro, Mandeok Pass Road in Buk-gu, and Hwangnyeongsan-ro in Yeonje-gu, were once controlled in the morning. Local governments and others in charge carried out snow removal work by mobilizing working vehicles to spray calcium chloride.
Ency	Document Title: Seorakmyeon. Body: Seorak-myeon is a myeon of Gapyeong-gun, Gyeonggi-do, Korea. The area is 141.53 km², and the population is 8,986 based on resident registration at the end of December 2016. Seorak-myeon was originally part of Yangpyeong-gun in the south of the Bukhangang River, but was incorporated into Gapyeong-gun in 1942. It has nothing to do with Seoraksan Mountain in Gangwon-do.

Table 8: An example document in Table 7, translated in English by a machine translator.

Context ->	아 더빙.. 진짜 짜증나네요 목소리 (부정) 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나 (부정) 너무재밓었다그래서보는것을추천한다 (긍정) 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정 (부정) 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다 (긍정) ... 원작의 긴장감을 제대로 살려내지못했다.
Correct Answer ->	(부정)
Incorrect Answer ->	(긍정)

Table 9: Formatted dataset example for NSMC. (few-shot: 70)

Context ->	Ah dubbing.. It's really annoying. voice (Negative) Hm... I saw the poster and gave elementary school student movie lines...Even overacting isn't light. (negative) It was so much fun, so I recommend watching it (negative) It's about the prison...Honestly, it's not fun.Adjusting the rating (negative) It's a movie where Simon Peg's humorous acting stood out!Kirsten Dunst, who only looked old in Spider-Man, looked so pretty. (positive) ... It did not capture the tension of the original work properly.
Correct Answer ->	(negative)
Incorrect Answer ->	(positive)

Table 10: An example document in Table 9, translated in English by a machine translator

Context ->	제목: 임종석 지문: 1989년 2월 15일 여의도 농민 폭력 시위를 주도한 혐의(폭력행위등처벌에관한법률위반)으로 지명수배되었다. 1989년 3월 12일 서울지방검찰청 공안부는 임종석의 사전구속영장을 발부받았다. 같은 해 6월 30일 평양축전에 임수경을 대표로 파견하여 국가보안법위반 혐의가 추가되었다. 경찰은 12월 18일 20일 사이 서울 경희대학교에서 임종석이 성명 발표를 추진하고 있다는 첩보를 입수했고, 12월 18일 오전 7시 40분 경 가스총과 전자봉으로 무장한 특공조 및 대공과 직원 12명 등 22명의 사복 경찰을 승용차 8대에 나누어 경희대학교에 투입했다. 1989년 12월 18일 오전 8시 15분 경 서울청량리경찰서는 호위 학생 5명과 함께 경희대학교 학생회관 건물 계단을 내려오는 임종석을 발견, 검거해 구속을 집행했다. 임종석은 청량리경찰서에서 약 1시간 동안 조사를 받은 뒤 오전 9시 50분 경 서울 장안동의 서울지방경찰청 공안분실로 인계되었다. 질문: 1989년 6월 30일 평양축전에 대표로 파견 된 인물은? 답변: 임수경 질문: 임종석이 여의도 농민 폭력 시위를 주도한 혐의로 지명수배된 연도는? 답변: 1989년 질문: 임종석을 검거한 장소는 경희대 내 어디인가? 답변: 학생회관 건물 계단 질문: 임종석이 조사를 받은 뒤 인계된 곳은 어딘가? 답변: 서울지방경찰청 공안분실 질문: 1989년 2월 15일 여의도 농민 폭력 시위를 주도한 혐의로 지명수배된 사람의 이름은? 답변:
Target Completion ->	임종석

Table 11: Formatted dataset example for KorQuAD: Machine Reading Comprehension (MRC) (few-shot: 4)

Context ->	Title: Lim Jongseok. Main Text: On February 15, 1989, he was wanted for leading a violent demonstration against farmers in Yeouido (violence of the Punishment of Violence, etc. Act). On March 12, 1989, the Ministry of Public Security of the Seoul District Prosecutors' Office received a preliminary arrest warrant for Lim Jong-seok. On June 30 of the same year, Lim Soo-kyung was dispatched as a representative to the Pyongyang Festival, adding charges of violating the National Security Act. The police obtained information that Lim Jong-seok was pushing for a statement at Kyung Hee University in Seoul between December 18 and December 18, and distributed 22 plainclothes police, including 12 special forces and anti-aircraft staff armed with gas guns and electronic rods, to Kyung Hee University. At around 8:15 a.m. on December 18, 1989, the Seoul Cheongnyangni Police Station found Lim Jong-seok, who came down the stairs of the Kyunghee University Student Center building with five escort students, arrested him and executed his arrest. Lim Jong-seok was investigated by the Cheongnyangni Police Station for about an hour and handed over to the Seoul Metropolitan Police Agency's public security loss office in Jangan-dong, Seoul, at around 9:50 a.m. Question: Who was dispatched as a representative at the Pyongyang Festival on June 30, 1989? Answer : Lim Su-kyung Question: When was Lim Jong-seok wanted to be arrested for leading a violent demonstration against farmers in Yeouido? Answer: 1989 Question: Where in Kyung Hee University did you arrest Lim Jong-seok? Answer: Stairs in the building of the student center. Question: Where was Lim Jongseok handed over after being investigated? Answer: Seoul Metropolitan Police Agency lost public security. Question: What is the name of the person who was wanted for leading the Yeouido peasant violence protest on February 15, 1989? Answer:
Target Completion ->	Lim Jongseok.

Table 12: Example document in Table 11, translated in English by a machine translator

Context ->	스키너가 말한 보상은 대부분 눈으로 볼 수 있는 현물이다.=Skinner's reward is mostly eye-watering. 심지어 어떤 문제가 발생할 건지도 어느 정도 예측이 가능하다.=Even some problems can be predicted. ... 오직 하나님만이 그 이유를 제대로 알 수 있을 겁니다.=
Target Completion ->	Only God will exactly know why.

Table 13: Formatted dataset example for AI-Hub: Translation

Context ->	네이버랩스 3D 지도 기술업체 에피폴라 인수(과학) 野 北 축구 생중계 거부에 대북정책 현주소종합(정치) 즐라탄 행선지 정한 듯…큰 소식 알려드리겠다(스포츠) 페루 아마존 지역서 70대 英 환경운동가 불에 타 숨진 채 발견(세계) 머리 맞댄 경제부총리와 한국은행 총재(경제) 전주 MBC 전북 출>신 故 이용마 기자 추모공간 사흘간 운영(사회) ... 구글 인공지능 다음 도전은 스타크래프트
Correct Answer ->	(과학)
Incorrect Answer ->	(세계)

Table 14: Formatted dataset example for YNAT: Topic Classification (few-shot: 70)

Context ->	NAVER LABS acquires 3D map technology company Epipolar(Science) North Korea's refusal to broadcast live soccer matches the current state of North Korea policy(Politics) It seems that Zlatan's destination has been decided... I'll tell you the big news(Sport) British environmentalist 70-year-old found burnt to death in Peruvian Amazon(World) Deputy Prime Minister of Economy and Bank of Korea Governor(Economy) Jeonju MBC Operates a three-day memorial space for the late reporter Lee Yong-ma from Jeonbuk(Social) ... Google's next AI challenge is Starcraft
Correct Answer ->	(Science)
Incorrect Answer ->	(World)

Table 15: Example document in Table 14, translated in English by a machine translator