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

Seongjin Shin∗,1 Sang-Woo Lee∗,1,2 Hwijeen Ahn1 Sungdong Kim2 HyoungSeok Kim1 Boseop Kim1 Kyunghyun Cho3 Gichang Lee1 Woomyoung Park1 Jung-Woo Ha1,2 Nako Sung1

NAVER CLOVA1 NAVER AI Lab2 NYU3

Abstract

Many recent studies on large-scale language models have reported successful in-context zero- and 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 effects of the source and size of the pretraining corpus on in-context learning in HyperCLOVA, a Korean-centric GPT-3 model. 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 the pretraining corpus does not necessarily determine 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, especially in the few-shot setting, 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 few-shot learning performance.

1 Introduction

NLP community has been surprised by emergence of in-context learning ability of a large-scale language model (LM) 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 (Black et al., 2021; Kim et al., 2021; Zeng et al., 2021; Rae et al., 2021; Hoffmann et al., 2022; Chowdhery et al., 2022).

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. 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). Rae et al. (2021) also use perplexity for the hyperparameter selection on corpus reweighting in the pretraining of their in-context learner. However, 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 HyperCLOVA, which is a Korean-centric large LM. (Kim et al., 2021). We mainly discover in-context few-shot learning as in the previous work (Kim et al., 2021) but also explore in-context zero-shot learning. 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
Our major findings include:

- **Corpus Source:** 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 a subcorpus of blog (Blog) achieves competitive in-context few-shot learning performance, but training on a subcorpus of community website (Cafe) or online news articles (News) hardly yields in-context few-shot 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 few-shot learning ability was not observed, but training on both corpora makes the emergence of in-context few-shot learning.

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

- **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. This is prominent 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 few-shot 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 for a large-scale LM to produce a target of this input, without requiring any parameter update. A few training examples are used in the in-context few-shot learning setting, whereas no training example is used in the in-context zero-shot setting. A few follow-up studies have tried to improve the in-context learning ability (Zhao et al., 2021; Holtzman et al., 2021). On the other hand, another group of papers tries to explain the mechanism of in-context few-shot learning (Min et al., 2022; Xie et al., 2022).

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). 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 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 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 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
Unlike the studies in a finetuning paradigm, many properties of the in-context learning related to pretraining corpus are still unknown. As the previous multi-task studies show that diverse tasks improve the ability of in-context learning, our study shows that diverse pretraining corpora strengthen the ability of in-context learning.

3 Task Definition

3.1 Model

We use the variants of HyperCLOVA with various parameter sizes and pretraining corpus. We mainly experiment with models with 1.3B parameters, but we also include the result for 6.9B-sized models. All models have a maximum sequence length of 2,048.

We emphasize that all models use the same vocabulary across all our 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 the 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 the characteristics of the subcorpora. Blog, Cafe, and News are taken from blog, community sites, and online news articles of NAVER1, a Korean web portal service, respectively. Comments is the comment threads related to the three subcorpora mentioned above. KiN comes from NAVER’s 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 Language2, including 3.2B of news, 2.1B of 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. Tables 12 and 13 in Appendix show the examples 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. As a result, the maximum training set size of each corpus is 150B tokens.

The validation set for each corpus consists of 10,000 examples and is used for the early stopping of models trained with each corpus. However, we combine all validation set to make the entire validation set contains 70,000 examples for seven domains, and the entire validation set is used for calculating perplexity, as described in Section 3.5.

3.3 Downstream Tasks

We evaluate in-context learning performance of each corpus-specific model on four Korean downstream task datasets used in Kim et al. (2021): NSMC3, KorQuAD (Lim et al., 2019), AI Hub translation4, 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.

We think that three datasets for downstream tasks are closely related to the HyperCLOVA cor-

---

1https://www.naver.com/
2https://corpus.korean.go.kr/
3https://github.com/e9t/nsmc
4https://aihub.or.kr/aidata/87

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

| 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  |
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 our in-house part-of-speech tagger.

pus. Passages which 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 includes 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 training corpus to 150B as 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 the cosine learning rate scheduling are used. We use the mixed precision training. Models are trained on the Nvidia Superpod which consists of 1,024 A100 GPUs spread across 128 nodes. Using Superpod, it spends 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 one with higher probability. KorQuAD and AI Hub are free-form completion tasks, where we directly generate output tokens using the greedy decoding.

In the few-shot experiments, the number of shots is set to 70, 4, 4, and 70 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 D. Appendix D also includes the examples of the few-shot prompts used in our experiments. These all experimental settings in the few-shot experiments, from the number of shots to the number of random trials, basically come 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 that the standard derivation of YNAT is relatively high.

3.5 Measuring Validation Perplexity

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

In Tables 2 and 4, we use Italics font for the results from a multi-domain model, which is pretrained with two or more mixed corpora. Since a multi-domain model trains more domains than a single-domain model, the PPLs of multi-domain models are generally lower than those of single-domain models. To keep readers from directly comparing PPLs between a single-domain and a multi-
domain model, we use italic font for the results of a multi-domain model.

4 Experimental Results

We perform intensive experiments to answer these four main questions:

1. How large do the source and the size of pre-training corpora have the effects on emerging in-context learning ability? (Section 4.2 and 4.3)

2. What is the effect of combining various corpora? (Section 4.4)

3. How large does domain relevance of corpus influence on model performances of the downstream task? (Section 4.5)

4. How strong is the correlation between validation perplexity and in-context learning of language models? (Section 4.6)

4.1 Main Results

Tables 2 and 4 show the in-context few-shot results on various pretraining corpus sources and different corpus combination, respectively. Tables 3 and 5 depict the in-context zero-shot results of some models in Tables 2 and 4, respectively. All results in Tables 2, 3, 4, and 5 come from models with 1.3B parameters. Tables 8 and 9 in Appendix A present the standard derivation values on the results of Tables 2 and 4.

In Tables 2, 3, 4, and 5, 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. We use this mean value of Majority and ALL in Table 2 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.

Tables 2 and 6 include in-context few-shot results on various pretraining corpus sizes. In Table 6, for example, 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. On the other hand, Table 2 compares 27B tokens models trained with different corpus sources to show the results in controlled corpus size.

4.2 Effect of Corpus Source

It is noticeable that in-context learning ability emerges differently depending on pretraining corpus sources, as shown in Tables 2, 3, 4, and 5. For example, Blog model makes competitive in-context few-shot learning performance to ALL model, while each of Cafe and News models hardly shows in-context few-shot 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 4.

These differences in in-context learning are dramatic compared to the finetuning results we expect in general. For a comparative experiment between in-context learning and finetuning in our setting, we also finetuned the experimented models with LoRA (Hu et al., 2021). As Table 11 in Appendix C shows, the performance differences in finetuning are much smaller than in the case of in-context learning.

4.3 Effect of Corpus Size

Table 6 shows 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 similar results for three Blog models of different sizes in Table 2. Blog and Blog 54B achieve 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. In the 6.9B-sized models, the in-context few-shot performance with 56B tokens does not decrease significantly compared to 150B tokens, as in the 1.3B-sized models.
### Table 2: In-context few-shot learning performance with different pretraining corpus. Models with 1.3B parameters are used. 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.

| Model | Corpus | PPL | NSMC (Acc) | KorQuAD (EM) | AI Hub (BLEU) Ko→En | En→Ko | YNAT (F1) |
|-------|--------|-----|------------|-------------|----------------------|-------|-----------|
| Majority | -     | -   | 50.35      | 0.0         | 0.0                  | 0.0   | 8.26      |
| ALL    | 150B   | 119.99 | 84.59      | 56.17       | 73.47                | 6.15  | 23.36     | 59.57     |
| ALL w/o Others | 150B | 119.66 | 84.59      | 56.49       | 74.20                | 6.14  | 23.21     | 50.76     |
| Blog   | 150B   | 152.40 | 83.50      | 50.74       | 69.34                | 3.82  | 20.11     | 60.68     |
| Cafe   | 82.5B  | 170.85 | 57.77      | 3.12        | 14.26                | 2.83  | 16.53     | 11.04     |
| News   | 73.1B  | 234.78 | 50.72      | 0.14        | 9.96                 | 1.10  | 15.88     | 14.36     |
| Comments | 40.7B | 225.39 | 79.78      | 14.69       | 33.33                | 0.79  | 5.06      | 36.17     |
| KiN    | 27.0B  | 187.80 | 54.73      | 4.85        | 18.99                | 6.81  | 16.16     | 9.23      |
| Modu   | 5.9B   | 226.01 | 69.91      | 30.20       | 49.29                | 1.21  | 6.13      | 43.27     |
| Ency   | 1.7B   | 549.40 | 53.81      | 0.71        | 11.88                | 0.58  | 0.69      | 27.99     |
| Blog 54B | 54.0B | 155.69 | 83.06      | 49.13       | 68.10                | 3.93  | 21.12     | 57.97     |
| Blog 27B | 27.0B | 165.60 | 80.27      | 10.91       | 23.41                | 5.35  | 12.32     | 48.19     |
| Cafe 27B | 27.0B | 169.81 | 49.91      | 1.37        | 13.98                | 4.25  | 20.74     | 8.60      |
| News 27B | 27.0B | 239.79 | 50.64      | 0.80        | 8.02                 | 2.42  | 15.78     | 27.20     |
| Comments 27B | 27.0B | 229.65 | 79.78      | 14.69       | 33.33                | 0.79  | 3.28      | 25.79     |

Table 3: In-context zero-shot performance with different pretraining corpus.

| Model | Corpus | PPL | NSMC (Acc) | KorQuAD (EM) | AI Hub (BLEU) Ko→En | En→Ko | YNAT (F1) |
|-------|--------|-----|------------|-------------|----------------------|-------|-----------|
| Majority | -     | -   | 50.35      | 0.0         | 0.0                  | 0.0   | 8.26      |
| ALL    | 150B   | 119.99 | 61.68      | 56.17       | 73.47                | 7.43  | 24.81     | 42.79     |
| Blog   | 150B   | 152.40 | 75.28      | 50.74       | 69.34                | 5.44  | 22.88     | 49.34     |
| Cafe   | 82.5B  | 170.85 | 69.38      | 3.12        | 14.26                | 4.34  | 16.44     | 38.12     |
| News   | 73.1B  | 234.78 | 54.96      | 0.14        | 9.96                 | 1.28  | 10.21     | 48.03     |
| Comments | 40.7B | 225.39 | 57.69      | 14.69       | 33.33                | 1.98  | 3.94      | 32.46     |
| KiN    | 27.0B  | 187.80 | 65.43      | 4.85        | 18.99                | 4.64  | 10.42     | 36.06     |
| Modu   | 5.9B   | 226.01 | 72.50      | 30.22       | 49.30                | 2.39  | 7.55      | 35.28     |
| Ency   | 1.7B   | 549.40 | 42.96      | 14.01       | 31.51                | 0.80  | 0.77      | 30.22     |

### 4.4 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 4 shows that in-context few-shot learning ability can be emerged by combining two corpora, even if each of both corpora cannot provide in-context few-shot 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 few-shot learning ability, while each of Cafe and KiN fails in most tasks. In-context zero-shot abilities of these models follow similar patterns as shown in Table 5.

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 may learn 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. For example, from the case of Cafe+News in Table 4, even if the mixed corpus model shows slightly better performance on KorQuAD than each of two corpora, its in-context few-shot performances on NSMC, KorQuAD, and YNAT are still below the basis. Furthermore, the performances on NSMC and YNAT even decrease.
Table 4: In-context few-shot learning performance with different corpus combination.

| Corpus Type | Corpus Train | PPL | NSMC (Acc) | KorQuAD (EM) | AI Hub (BLEU) Ko→En | En→Ko (F1) | YNAT (F1) |
|-------------|--------------|-----|------------|--------------|----------------------|------------|-----------|
| ALL         | 150B         | 119.99 | 84.59 | 56.17  | 73.47 | 6.15 | 23.36 | 59.57 |
| KiN+Ency    | 28.7B        | 164.69 | 59.17  | 42.09 | 61.00 | 8.99 | 23.12 | 42.84 |
| Cafe+KiN    | 109.5B       | 141.92 | 76.42  | 38.45 | 59.00 | 8.41 | 23.41 | 56.96 |
| Cafe+News   | 150B         | 154.20 | 54.13  | 8.95  | 22.72 | 4.45 | 17.77 | 8.19  |
| Blog+Comments+Modu | 150B | 144.67 | 82.82  | 54.94 | 72.27 | 4.09 | 21.17 | 65.01 |
| News+KiN+Ency | 101.8B | 142.13 | 75.96  | 35.42 | 55.60 | 10.18 | 23.38 | 27.54 |

Table 5: In-context zero-shot learning performance with different corpus combination.

| Corpus Type | Corpus Train | PPL | NSMC (Acc) | KorQuAD (EM) | AI Hub (BLEU) Ko→En | En→Ko (F1) | YNAT (F1) |
|-------------|--------------|-----|------------|--------------|----------------------|------------|-----------|
| ALL         | 150B         | 119.99 | 61.88  | 56.17  | 73.47 | 7.43 | 24.81 | 42.79 |
| KiN+Ency    | 28.7B        | 164.69 | 56.78  | 42.09 | 61.00 | 11.51 | 24.93 | 37.71 |
| Cafe+KiN    | 109.5B       | 141.92 | 59.27  | 38.45 | 59.00 | 10.12 | 24.95 | 45.44 |
| Cafe+News   | 150B         | 154.20 | 66.92  | 8.95  | 22.85 | 3.49 | 15.77 | 47.34 |
| Blog+Comments+Modu | 150B | 144.67 | 69.15  | 54.94 | 72.27 | 6.06 | 22.03 | 48.25 |
| News+KiN+Ency | 101.8B | 142.13 | 61.49  | 35.42 | 55.60 | 10.18 | 24.13 | 51.89 |

Table 6: In-context few-shot learning performance of ALL with different size of the pretraining data. The dataset is randomly sampled from the original corpus.

| # of tokens | NSMC (Acc) | KorQuAD (EM) | AI Hub (BLEU) Ko→En | En→Ko (F1) | 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 few-shot learning performance with different corpus combination.

4.5 Effect of Domain Relevance

Speaking of the few-shot results, Table 2 shows that the close relationship between a pretraining corpus and a downstream task does not always guarantee in-context few-shot 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 4 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 tasks are much lower than Blog and KiN.

The counter example of above supports is AI 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.

In the zero-shot setting, on the other hand, do-
main relevance seems to affect more positively. For example, training the News corpus helps in-context zero-shot learning in KLUE-YNAT consistently. As shown in Tables 3 and 5, the models whose training corpus includes the News corpus (i.e., News, Cafe+News, and News+KiN+Ency) even perform better than the model trained whole HyperCLOVA corpus.

In the case of KiN and AI Hub, zero-shot performance increase for AI Hub tasks of the KiN model is less significant than few-shot. However, adding KiN corpus into the pretraining corpus in the experiments of Table 5 (i.e., KiN+Ency, Cafe+KiN, and News+KiN+Ency) makes a consistent performance increase, and the model outperform ALL.

### 4.6 Perplexity and Downstream Task

Figure 2 presents the scatter plots of PPL (x-axis) and in-context few-shot learning performance (y-axis) on five downstream tasks for single corpus models and the ALL model. In Figure 2, we normalized in-context few-shot 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 few-shot learning ability.

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

Speaking of zero-shot results, it seems Table 3 shows that in-context zero-shot learning performances relatively more correlate with perplexity than the few-shot cases. Nevertheless, Modu still has both relatively high perplexity and relatively high in-context zero-shot learning performances.

Table 7 shows validation perplexity scores for
Table 7: 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.

|     | 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|

Figure 4: Relation between validation PPL and in-context few-shot learning performance for five downstream tasks as pretraining steps proceed. The results come from the 1.3B-sized **ALL** model.

Each row corresponds to the model and each column corresponds to the validation set’s subcorpus. Each validation set except All in Table 7 consists of 10,000 instances, and is the part of our main validation sets, consists of 70,000 instances.

On the other hand, Figure 4 shows that PPL and in-context learning performance correlate well in the perspective of 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 to increase the performance of in-context learning when the corpus is small or there exists demand 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, especially in the few-shot setting. 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.

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.

Acknowledgment

The authors thank all the members of CLOVA, AI Lab for devoted supporting and discussion. In particular, they thank Joonsuk Park and Seok Ho Yoon for proofreading.

References

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

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

Leo Gao, Stella Biderman, Sid Black, Laurence Goldberg, 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.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Sharan Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

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 Artestxe, 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.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? arXiv preprint arXiv:2202.12837.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoung 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, Ur mish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Dutta, 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.

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.

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

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.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In ICLR.
A Details on Experimental Results

Tables 8 and 9 show standard derivation value on Tables 2 and 4. Table 10 shows score difference with ALL in addition to in-context learning scores on Table 2. Figure 5, supporting Table 7, shows the validation perplexity of different model from different corpus.

B Details on Pretraining Corpus

Tables 12 and 13 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 description, data cleaning, 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 Experiments on LoRA

Table 11 shows the results of LoRA (Hu et al., 2021) finetuning on some models in Tables 2 and 4.

D Examples of Few-shot Prompt

Tables 14, 16, 18, and 19 show the example few-shot prompt of NSMC, KorQuAD, AI Hub, and YNAT, respectively. Tables 15, 17, and 20 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 is to change few-shot examples in the prompt. 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.
E Generalization to Other Languages

Someone can ask whether our results can be extended to other languages, 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 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.

Table 8: The results of Table 2 with standard deviation in parentheses.

| Corpus Type | NSMC (Acc) | AI Hub (BLEU) | YNAT (F1) |
|-------------|------------|---------------|-----------|
|             | Ko→En | En→Ko | Ko→En | En→Ko | Ko→En | En→Ko | Ko→En | En→Ko | Ko→En | En→Ko |
| ALL         | 84.59(1.25) | 6.15(0.16) | 23.36(0.33) | 59.57(4.30) |
| ALL w/o Others | 84.59(1.25) | 6.14(0.21) | 23.21(0.45) | 30.76(11.81) |

Table 9: The results of Table 4 with standard deviation in parentheses.

| Corpus Type | NSMC (Acc) | AI Hub (BLEU) | YNAT (F1) |
|-------------|------------|---------------|-----------|
|             | Ko→En | En→Ko | Ko→En | En→Ko | Ko→En | En→Ko | Ko→En | En→Ko | Ko→En | En→Ko |
| Blog 54B    | 83.06(2.26) | 3.93(0.29) | 21.12(0.19) | 57.97(7.27) |
| Blog 27B    | 80.27(2.28) | 5.35(1.95) | 12.32(5.68) | 48.19(6.71) |
| Cafe 27B    | 49.91(0.29) | 4.25(0.35) | 20.74(1.57) | 8.60(2.79) |
| News 27B    | 50.64(3.26) | 2.42(1.41) | 15.78(4.21) | 27.20(5.86) |
| Comments 27B | 80.50(1.44) | 1.70(0.03) | 3.38(0.18) | 25.79(7.27) |

Table 10: Table 4 which includes the difference from ALL in parentheses.
Table 11: LoRA finetuning performance on different pretraining corpus and its combination.

| Model       | NSMC (Acc) | YNAT (F1) |
|-------------|------------|-----------|
| ALL         | 91.83      | 86.47     |
| Comments    | 92.02      | 84.07     |
| Blog        | 91.93      | 86.21     |
| Cafe        | 91.57      | 85.45     |
| News        | 90.62      | 86.57     |
| KiN         | 90.89      | 84.46     |
| Modu        | 90.60      | 86.31     |
| Ency        | 86.93      | 82.37     |
| KiN+Ency    | 90.92      | 84.00     |
| Cafe+KiN    | 90.99      | 87.52     |
| Cafe+News   | 91.37      | 86.37     |
| Blog+Comment+Modu | 88.83 | 87.07 |
| News+KiN+Ency | 91.13 | 86.62 |

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 website 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    |
| 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 photos 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     |
| 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     |
| 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 |
| 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 Gyeongbuk-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 Daejeon-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. |

| 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 13: An example document in Table 12, translated into English by a machine translator.
Table 14: Formatted dataset example for NSMC. (few-shot: 70)

| Context | Correct Answer | Incorrect Answer |
|---------|----------------|------------------|
| Incorrect Answer -> | (부정) | (긍정) |

Table 15: An example document in Table 14, translated into English by a machine translator

| Context | Correct Answer | Incorrect Answer |
|---------|----------------|------------------|
| Incorrect Answer -> | (negative) | (positive) |

Table 16: Formatted dataset example for KorQuAD: Machine Reading Comprehension (MRC) (few-shot: 4)

| Context | Target Completion |
|---------|-------------------|
| Incorrect Answer -> | 임종석 |
Title: Lim Jong-seok.
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 Jong-seok 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: Lim Jong-seok.
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 20: Example document in Table 19, translated into English by a machine translator