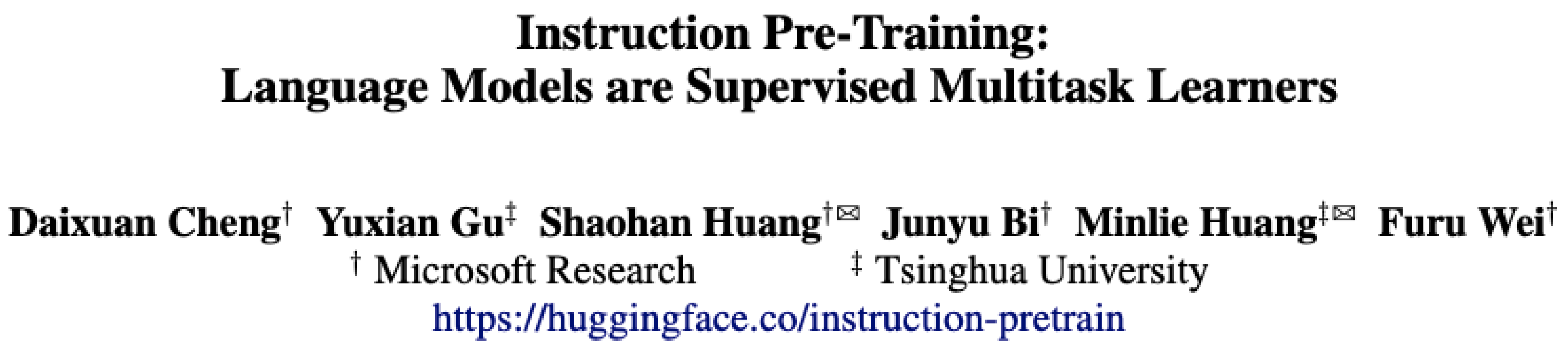


# Instruction Pre-Training: Language Models are Supervised Multitask Learners

**Daixuan Cheng**[†] **Yuxian Gu**[‡] **Shaohan Huang**[†✉] **Junyu Bi**[†] **Minlie Huang**[‡✉] **Furu Wei**[†]

[†] Microsoft Research [‡] Tsinghua University

https://huggingface.co/instruction-pretrain

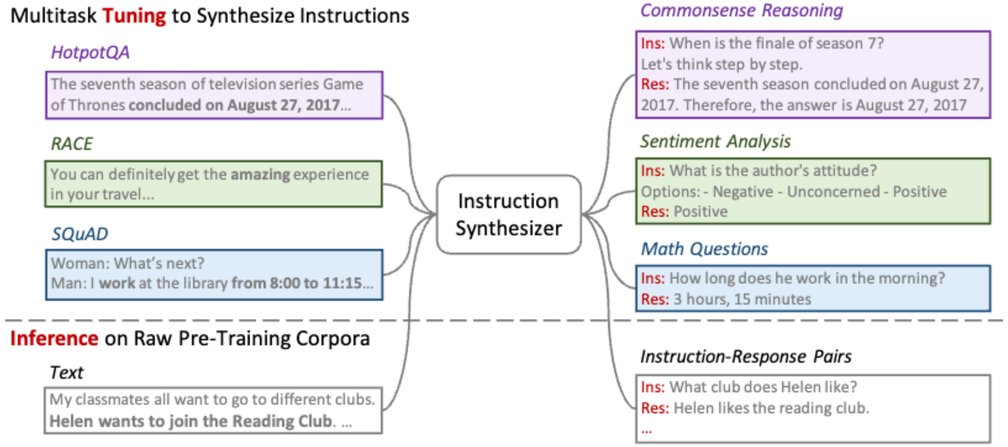
# Instruction Pretraining

## Key Point

- Why Instruction Pretraining?

⇒ Pre-training – Fine Tuning Alignment

⇒ Instruction Format Tuning enhances Task Generalization



Multitask **Tuning** to Synthesize Instructions

*HotpotQA*

The seventh season of television series Game of Thrones **concluded on August 27, 2017**...

*RACE*

You can definitely get the **amazing** experience in your travel...

*SQuAD*

Woman: What's next?
Man: I **work** at the library **from 8:00 to 11:15**...

Instruction Synthesizer

*Commonsense Reasoning*

Ins: When is the finale of season 7?
Let's think step by step.
Res: The seventh season concluded on August 27, 2017. Therefore, the answer is August 27, 2017

*Sentiment Analysis*

Ins: What is the author's attitude?
Options: - Negative - Unconcerned - Positive
Res: Positive

*Math Questions*

Ins: How long does he work in the morning?
Res: 3 hours, 15 minutes

**Inference** on Raw Pre-Training Corpora

*Text*

My classmates all want to go to different clubs. **Helen wants to join the Reading Club**. ...

*Instruction-Response Pairs*

Ins: What club does Helen like?
Res: Helen likes the reading club.
...

- How to make Instruction Pretraining Dataset from Raw data?

⇒ Human? GPT4 API? -> High Cost

⇒ Instruction synthesizer based on Open Source Model

# Instruction Synthesizer

## Training

- Mistral-7B 튜닝
- 다양한 분야에 대한 여러 Fine-tuning Dataset을 활용
- Context를 기반으로 Instruction-Response Pair를 생성하도록 학습
- 서로 다른 Context들을 직렬적으로 이어주어서 학습 및 추론하는 구조

## Inference

- Concatenating the texts and instruction-pairs from M rounds
- Multi-round 다양한 분야에 대한 여러 Fine-tuning Dataset을 활용
- 평균적으로 1 raw text / 약 5개의 Pairs를 생성

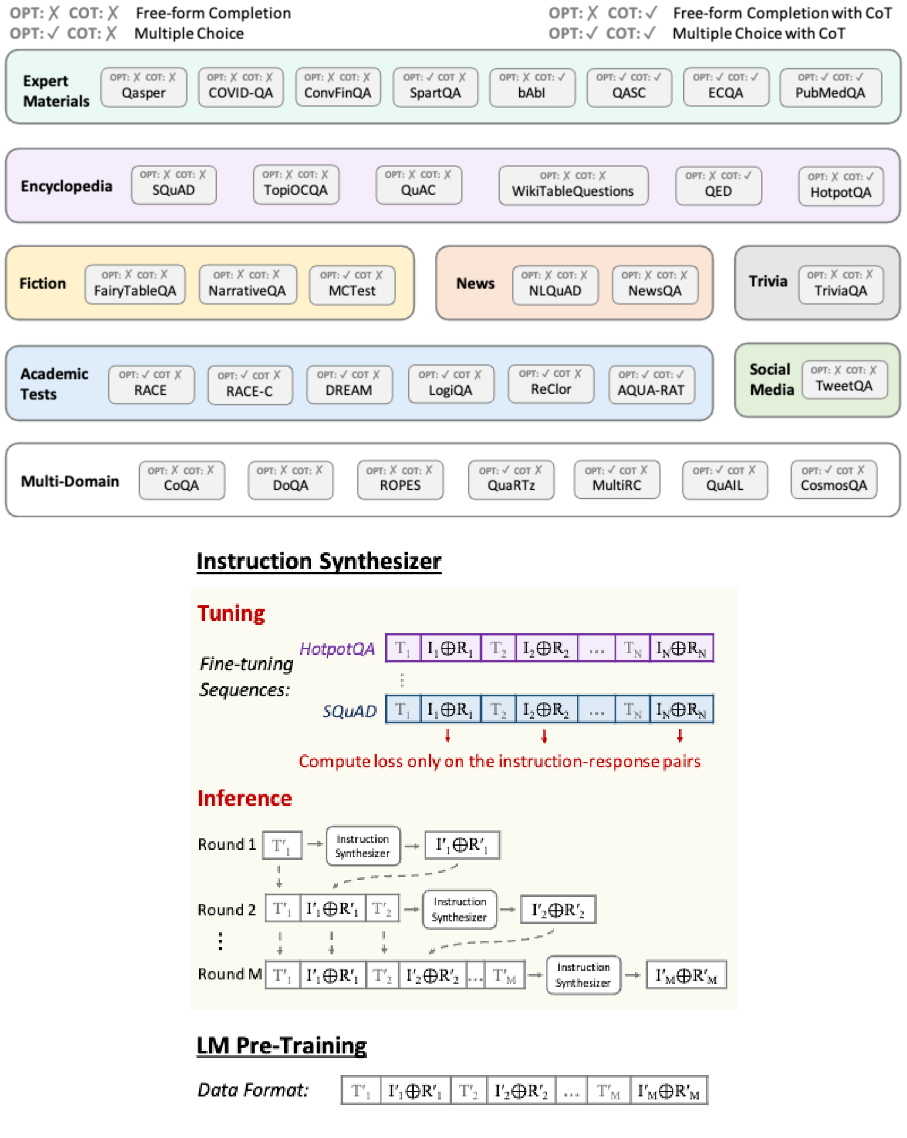


Figure 3: **For instruction synthesizer**, a one-shot ex-

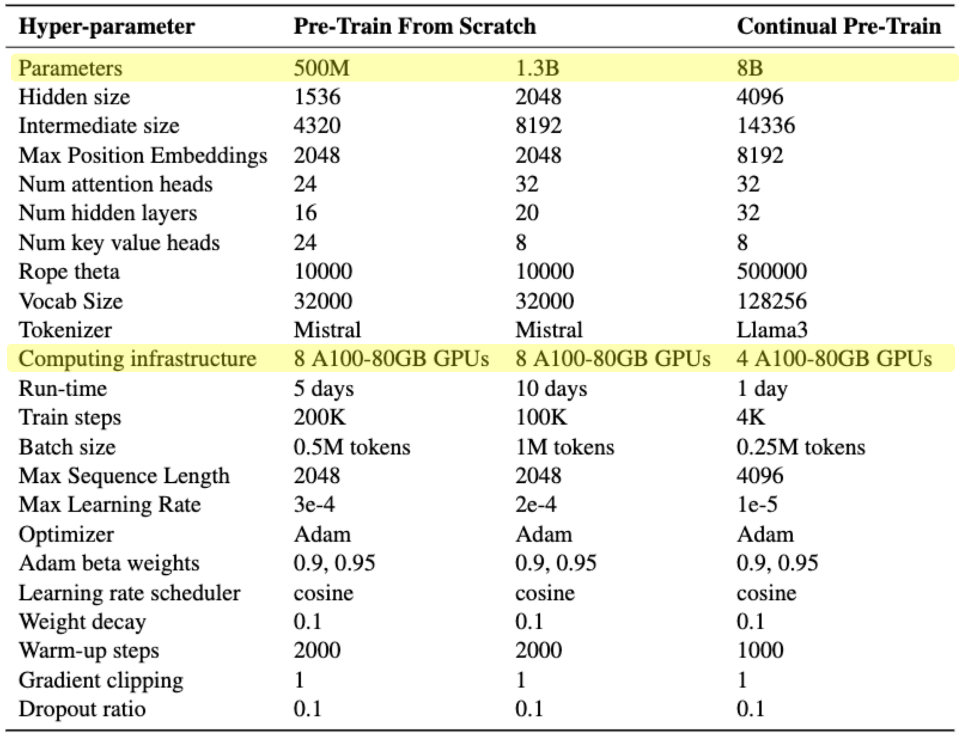
# LM Pre-Training

## 1. General Pre-Training From Scratch

- RefinedWeb: 200M text. 100B tokens
- 전체 Pre-Training Data를 전부 변환 X, 일부만 변환 (20%)
- 총 40M raw text를 기반으로 200M synthesized pairs를 생성
- Mix the fine-tuning data (0.2B) for instruction synthesizer

## 2. Domain-Adaptive Continual Pre-Training

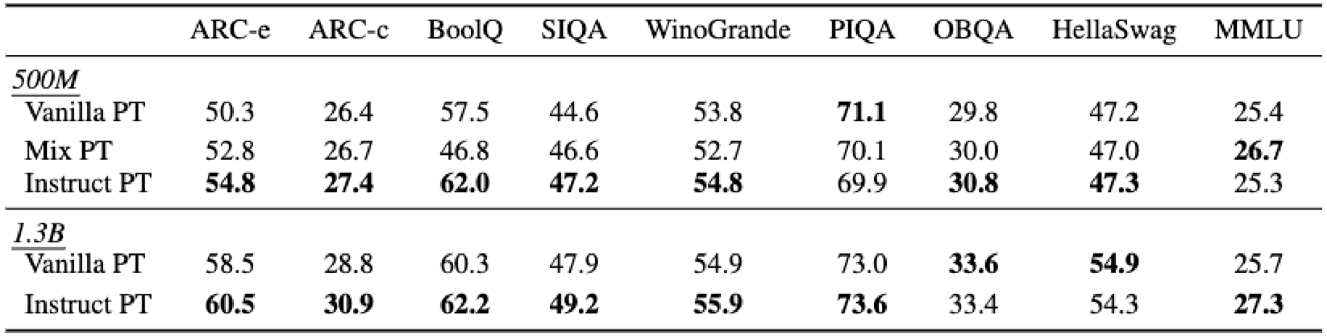
- PubMed Abstracts / Financial news
- Mix the instruction-augmented corpora with general instruction
- Llama3-8B



| Hyper-parameter | Pre-Train From Scratch | | Continual Pre-Train |
|---|---|---|---|
| Parameters | 500M | 1.3B | 8B |
| Hidden size | 1536 | 2048 | 4096 |
| Intermediate size | 4320 | 8192 | 14336 |
| Max Position Embeddings | 2048 | 2048 | 8192 |
| Num attention heads | 24 | 32 | 32 |
| Num hidden layers | 16 | 20 | 32 |
| Num key value heads | 24 | 8 | 8 |
| Rope theta | 10000 | 10000 | 500000 |
| Vocab Size | 32000 | 32000 | 128256 |
| Tokenizer | Mistral | Mistral | Llama3 |
| Computing infrastructure | 8 A100-80GB GPUs | 8 A100-80GB GPUs | 4 A100-80GB GPUs |
| Run-time | 5 days | 10 days | 1 day |
| Train steps | 200K | 100K | 4K |
| Batch size | 0.5M tokens | 1M tokens | 0.25M tokens |
| Max Sequence Length | 2048 | 2048 | 4096 |
| Max Learning Rate | 3e-4 | 2e-4 | 1e-5 |
| Optimizer | Adam | Adam | Adam |
| Adam beta weights | 0.9, 0.95 | 0.9, 0.95 | 0.9, 0.95 |
| Learning rate scheduler | cosine | cosine | cosine |
| Weight decay | 0.1 | 0.1 | 0.1 |
| Warm-up steps | 2000 | 2000 | 1000 |
| Gradient clipping | 1 | 1 | 1 |
| Dropout ratio | 0.1 | 0.1 | 0.1 |

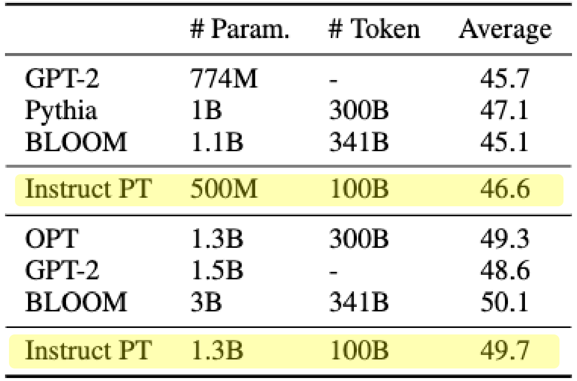
Table 10: **Hyper-Parameters of Pre-Training From Scratch and Continual Pre-Training.**

# Results



| | ARC-e | ARC-c | BoolQ | SIQA | WinoGrande | PIQA | OBQA | HellaSwag | MMLU |
|---|---|---|---|---|---|---|---|---|---|
| *500M* | | | | | | | | | |
| Vanilla PT | 50.3 | 26.4 | 57.5 | 44.6 | 53.8 | **71.1** | 29.8 | 47.2 | 25.4 |
| Mix PT | 52.8 | 26.7 | 46.8 | 46.6 | 52.7 | 70.1 | 30.0 | 47.0 | **26.7** |
| Instruct PT | **54.8** | **27.4** | **62.0** | **47.2** | **54.8** | 69.9 | **30.8** | **47.3** | 25.3 |
| *1.3B* | | | | | | | | | |
| Vanilla PT | 58.5 | 28.8 | 60.3 | 47.9 | 54.9 | 73.0 | **33.6** | **54.9** | 25.7 |
| Instruct PT | **60.5** | **30.9** | **62.2** | **49.2** | **55.9** | **73.6** | 33.4 | 54.3 | **27.3** |

Table 1: **General Performance of the Pre-Trained Base Models** via *Vanilla Pre-Training* (Vanilla PT), mixing raw corpora with fine-tuning data for the instruction synthesizer (Mix PT), and *Instruction Pre-Training* (Instruct PT) in general pre-training from scratch.



| | # Param. | # Token | Average |
|---|---|---|---|
| GPT-2 | 774M | - | 45.7 |
| Pythia | 1B | 300B | 47.1 |
| BLOOM | 1.1B | 341B | 45.1 |
| Instruct PT | 500M | 100B | 46.6 |
| OPT | 1.3B | 300B | 49.3 |
| GPT-2 | 1.5B | - | 48.6 |
| BLOOM | 3B | 341B | 50.1 |
| Instruct PT | 1.3B | 100B | 49.7 |

Table 2: **Comparison between Our Pre-Trained Base Models and Others** on general benchmarks. Detailed results are in Table 11.

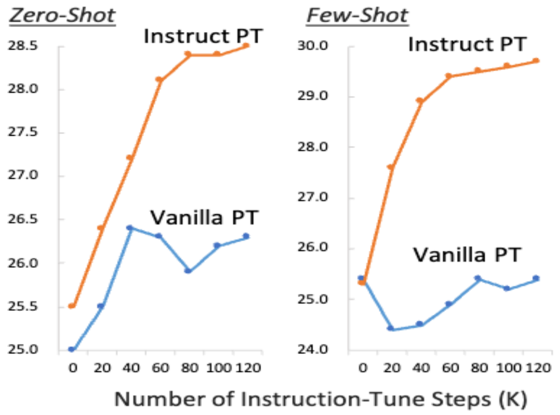


Figure 4: **MMLU Performance during Instruction Tuning** of models pre-trained via *Vanilla Pre-Training* (Vanilla PT) and *Instruction Pre-Training* (Instruct PT).

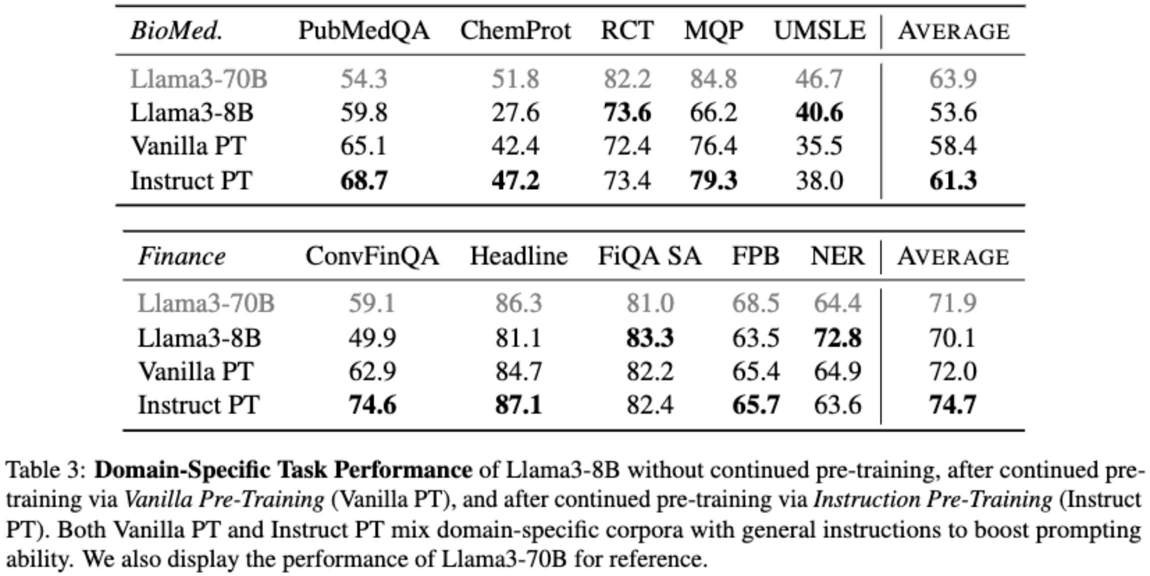
## General Pre-Training From Scratch

- Vanilla < Mix < Instruct

- Instruction PT == Data Efficiency
  => 토큰수가 적어도 매우 높은 성능

- **PT후, Instruction Tuning하면 성능이 급격히 증가**
  => PT와 SFT의 Alignment 향상으로 Downstream Task 에 더욱 빠르게 적응함

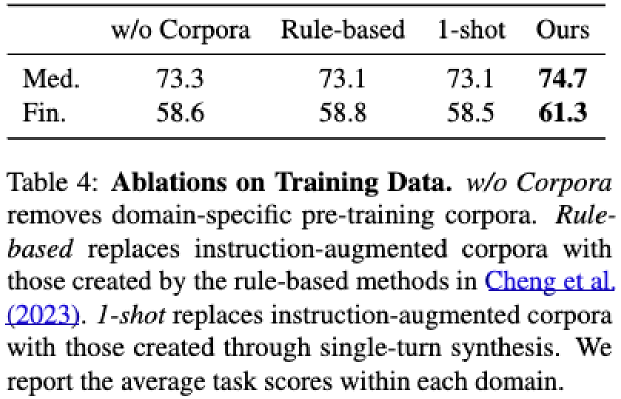
# Results



| *BioMed.* | PubMedQA | ChemProt | RCT | MQP | UMSLE | AVERAGE |
|---|---|---|---|---|---|---|
| Llama3-70B | 54.3 | 51.8 | 82.2 | 84.8 | 46.7 | 63.9 |
| Llama3-8B | 59.8 | 27.6 | **73.6** | 66.2 | **40.6** | 53.6 |
| Vanilla PT | 65.1 | 42.4 | 72.4 | 76.4 | 35.5 | 58.4 |
| Instruct PT | **68.7** | **47.2** | 73.4 | **79.3** | 38.0 | **61.3** |

| *Finance* | ConvFinQA | Headline | FiQA SA | FPB | NER | AVERAGE |
|---|---|---|---|---|---|---|
| Llama3-70B | 59.1 | 86.3 | 81.0 | 68.5 | 64.4 | 71.9 |
| Llama3-8B | 49.9 | 81.1 | **83.3** | 63.5 | **72.8** | 70.1 |
| Vanilla PT | 62.9 | 84.7 | 82.2 | 65.4 | 64.9 | 72.0 |
| Instruct PT | **74.6** | **87.1** | 82.4 | **65.7** | 63.6 | **74.7** |

Table 3: **Domain-Specific Task Performance** of Llama3-8B without continued pre-training, after continued pre-training via *Vanilla Pre-Training* (Vanilla PT), and after continued pre-training via *Instruction Pre-Training* (Instruct PT). Both Vanilla PT and Instruct PT mix domain-specific corpora with general instructions to boost prompting ability. We also display the performance of Llama3-70B for reference.



| | w/o Corpora | Rule-based | 1-shot | Ours |
|---|---|---|---|---|
| Med. | 73.3 | 73.1 | 73.1 | **74.7** |
| Fin. | 58.6 | 58.8 | 58.5 | **61.3** |

Table 4: **Ablations on Training Data.** *w/o Corpora* removes domain-specific pre-training corpora. *Rule-based* replaces instruction-augmented corpora with those created by the rule-based methods in Cheng et al. (2023). *1-shot* replaces instruction-augmented corpora with those created through single-turn synthesis. We report the average task scores within each domain.

## Domain-Adaptive Continual Pre-Training

- 해당 결과들은 기본 모델(Llama3)들을 제외하고 모두 Continual Pre-training을 한 결과
- 마찬가지로 Vanilla < Instruct

## Ablation on PT data

- w/o Corpora: w/o 도메인 특화 Instruction PT data
- Rule-based: 룰기반으로 PT 데이터 단순 제작
- 1-shot: Synthesizer Inference 단계에서 1 round 결과만 사용

# Analysis on Instruction Synthesizer

## Response Generation

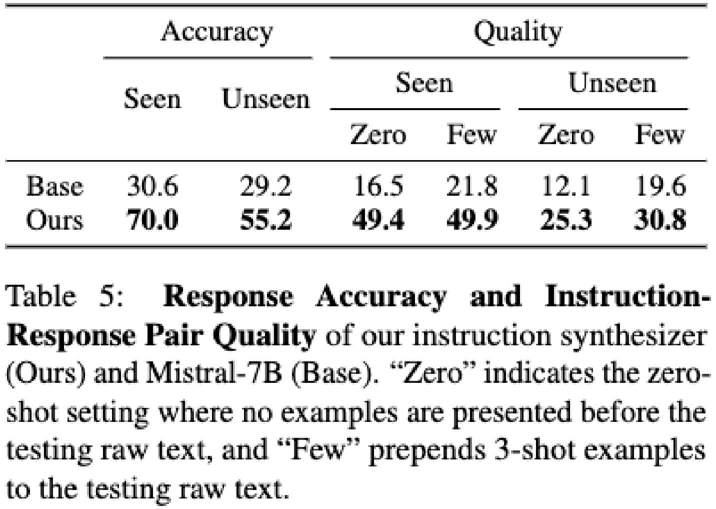
- Response Accuracy: Compute Accuracy
- Instruction Pair Quality:  F1 similarity between generated response and the gold response

1) Zero-shot: Input contains only the raw text

2) Few-shot: Input contains 3-shot gold pairs

- Unseen에 대해서도 높은 성능

⇒ 보지 못한 데이터에 대해서도 Instruction Pair를 만들기에 적합



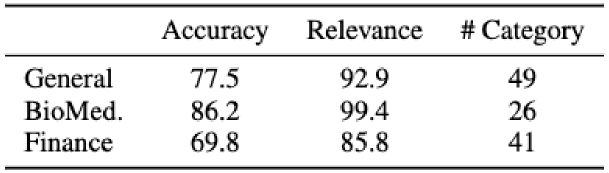
|  | Accuracy |  | Quality |  |  |  |
|---|---|---|---|---|---|---|
|  | Seen | Unseen | Seen |  | Unseen |  |
|  |  |  | Zero | Few | Zero | Few |
| Base | 30.6 | 29.2 | 16.5 | 21.8 | 12.1 | 19.6 |
| Ours | **70.0** | **55.2** | **49.4** | **49.9** | **25.3** | **30.8** |

Table 5: **Response Accuracy and Instruction-Response Pair Quality** of our instruction synthesizer (Ours) and Mistral-7B (Base). "Zero" indicates the zero-shot setting where no examples are presented before the testing raw text, and "Few" prepends 3-shot examples to the testing raw text.

# Analysis on Instruction Corpora

## Appropriateness of Instruction Corpora

- 생성된 Instruction Pair의 Accuracy, Relevance, Diversity 를 평가
- Instruction Pair에서 500개를 샘플링하고 GPT-4를 사용하여 명령어-응답 쌍을 위의 3가지 측면에서 평가 혹은 카테고라이즈하도록 시킴
- 정확한 응답이며, Context와 관련된 pair이면서도 다양한 Instruction Task Category들에 대한 응답을 생성



| | Accuracy | Relevance | # Category |
|---|---|---|---|
| General | 77.5 | 92.9 | 49 |
| BioMed. | 86.2 | 99.4 | 26 |
| Finance | 69.8 | 85.8 | 41 |

Table 6: **Response Accuracy, Context Relevance, and Number of Task Categories** of the instruction-augmented corpora.

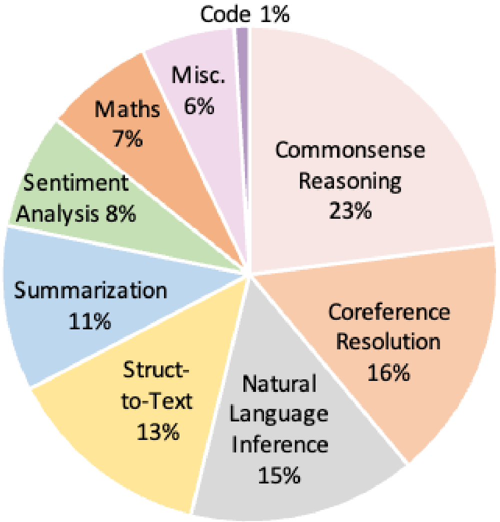
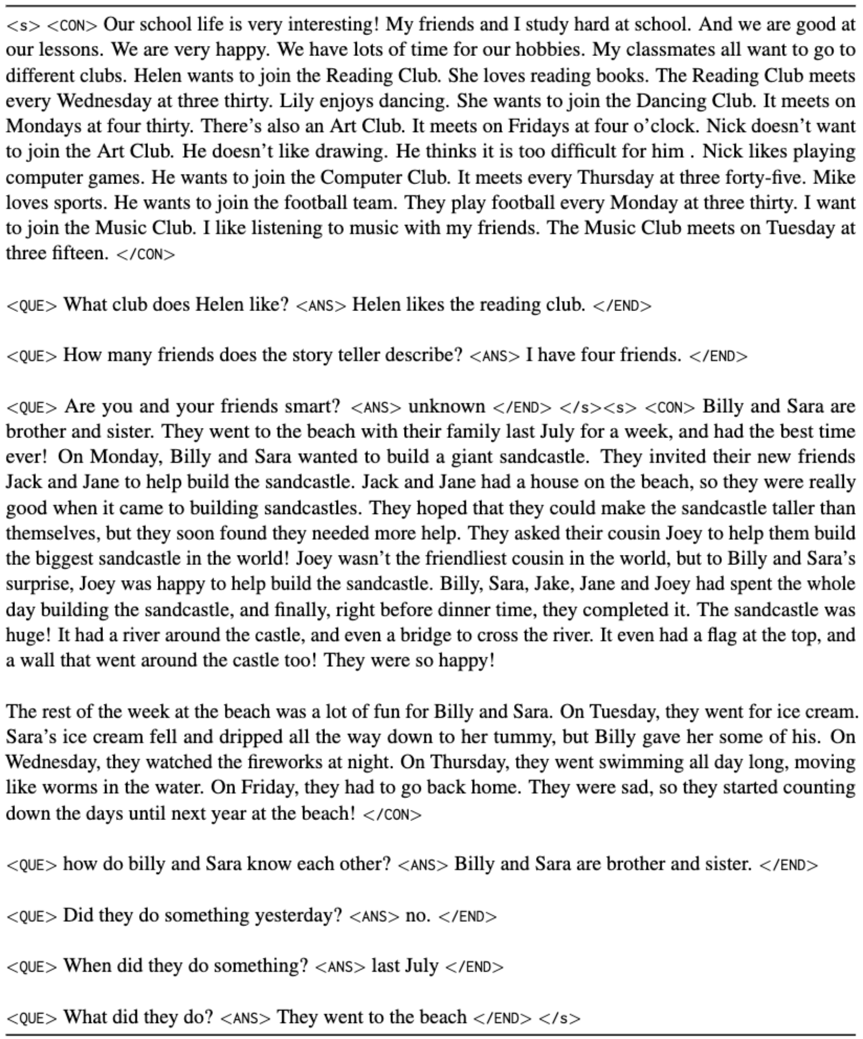


Figure 6: **Distribution of Task Scenarios of Synthesized Instruction-Response Pairs** in the instruction-augmented corpora.

# Examples



<s> <CON> Our school life is very interesting! My friends and I study hard at school. And we are good at our lessons. We are very happy. We have lots of time for our hobbies. My classmates all want to go to different clubs. Helen wants to join the Reading Club. She loves reading books. The Reading Club meets every Wednesday at three thirty. Lily enjoys dancing. She wants to join the Dancing Club. It meets on Mondays at four thirty. There's also an Art Club. It meets on Fridays at four o'clock. Nick doesn't want to join the Art Club. He doesn't like drawing. He thinks it is too difficult for him . Nick likes playing computer games. He wants to join the Computer Club. It meets every Thursday at three forty-five. Mike loves sports. He wants to join the football team. They play football every Monday at three thirty. I want to join the Music Club. I like listening to music with my friends. The Music Club meets on Tuesday at three fifteen. </CON>

<QUE> What club does Helen like? <ANS> Helen likes the reading club. </END>

<QUE> How many friends does the story teller describe? <ANS> I have four friends. </END>

<QUE> Are you and your friends smart? <ANS> unknown </END> </s><s> <CON> Billy and Sara are brother and sister. They went to the beach with their family last July for a week, and had the best time ever! On Monday, Billy and Sara wanted to build a giant sandcastle. They invited their new friends Jack and Jane to help build the sandcastle. Jack and Jane had a house on the beach, so they were really good when it came to building sandcastles. They hoped that they could make the sandcastle taller than themselves, but they soon found they needed more help. They asked their cousin Joey to help them build the biggest sandcastle in the world! Joey wasn't the friendliest cousin in the world, but to Billy and Sara's surprise, Joey was happy to help build the sandcastle. Billy, Sara, Jake, Jane and Joey had spent the whole day building the sandcastle, and finally, right before dinner time, they completed it. The sandcastle was huge! It had a river around the castle, and even a bridge to cross the river. It even had a flag at the top, and a wall that went around the castle too! They were so happy!

The rest of the week at the beach was a lot of fun for Billy and Sara. On Tuesday, they went for ice cream. Sara's ice cream fell and dripped all the way down to her tummy, but Billy gave her some of his. On Wednesday, they watched the fireworks at night. On Thursday, they went swimming all day long, moving like worms in the water. On Friday, they had to go back home. They were sad, so they started counting down the days until next year at the beach! </CON>

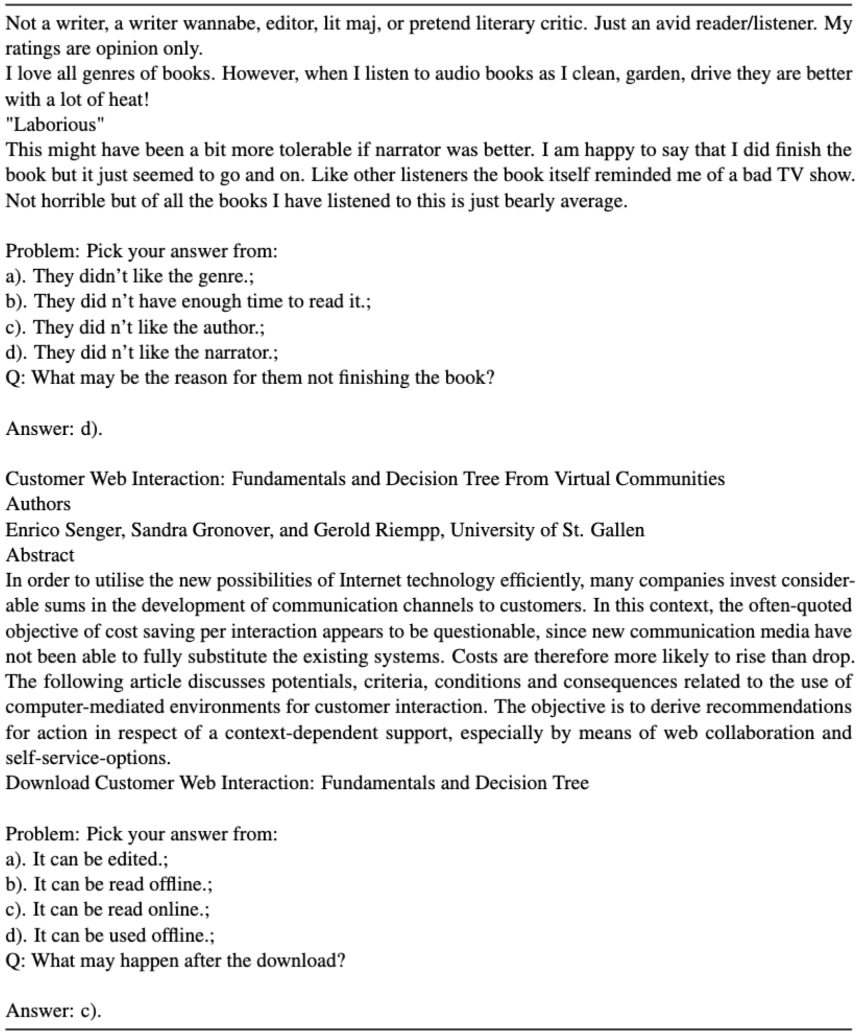
<QUE> how do billy and Sara know each other? <ANS> Billy and Sara are brother and sister. </END>

<QUE> Did they do something yesterday? <ANS> no. </END>

<QUE> When did they do something? <ANS> last July </END>

<QUE> What did they do? <ANS> They went to the beach </END> </s>

Table 9: **An Example of a Sequence for Fine-Tuning the Instruction Synthesizer.** This sequence contains two examples, both from the CoQA dataset (Reddy et al., 2019), constituting a 2-shot example.



Not a writer, a writer wannabe, editor, lit maj, or pretend literary critic. Just an avid reader/listener. My ratings are opinion only.
I love all genres of books. However, when I listen to audio books as I clean, garden, drive they are better with a lot of heat!
"Laborious"
This might have been a bit more tolerable if narrator was better. I am happy to say that I did finish the book but it just seemed to go and on. Like other listeners the book itself reminded me of a bad TV show. Not horrible but of all the books I have listened to this is just bearly average.

Problem: Pick your answer from:
a). They didn't like the genre.;
b). They did n't have enough time to read it.;
c). They did n't like the author.;
d). They did n't like the narrator.;
Q: What may be the reason for them not finishing the book?

Answer: d).

Customer Web Interaction: Fundamentals and Decision Tree From Virtual Communities
Authors
Enrico Senger, Sandra Gronover, and Gerold Riempp, University of St. Gallen
Abstract
In order to utilise the new possibilities of Internet technology efficiently, many companies invest considerable sums in the development of communication channels to customers. In this context, the often-quoted objective of cost saving per interaction appears to be questionable, since new communication media have not been able to fully substitute the existing systems. Costs are therefore more likely to rise than drop. The following article discusses potentials, criteria, conditions and consequences related to the use of computer-mediated environments for customer interaction. The objective is to derive recommendations for action in respect of a context-dependent support, especially by means of web collaboration and self-service-options.
Download Customer Web Interaction: Fundamentals and Decision Tree

Problem: Pick your answer from:
a). It can be edited.;
b). It can be read offline.;
c). It can be read online.;
d). It can be used offline.;
Q: What may happen after the download?

Answer: c).

Table 12: **A Case of a 2-shot Example in the General Instruction-Augmented Corpora.**

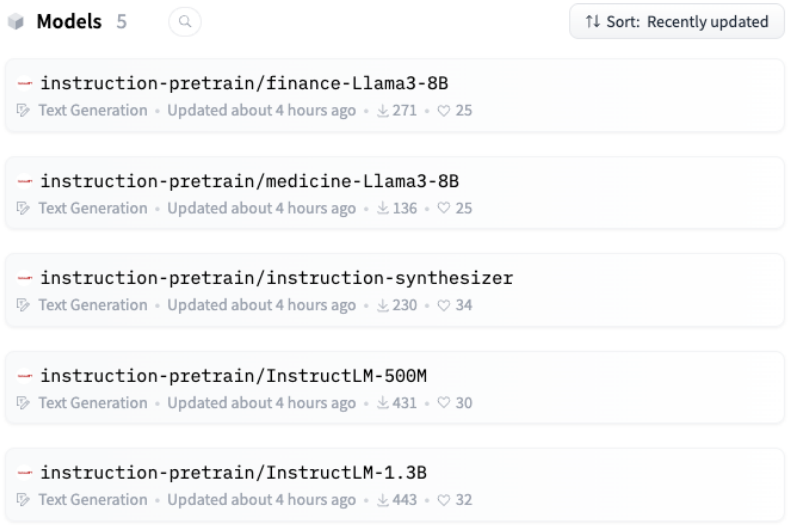
# Conclusion

## SOLAR의 Further PT 방법

- Dept up scaling 한 뒤, PT 할 때 데이터를 Instruction형태로 구성

## Continual Pre-Training의 가능성

- From scratch는 불가능해도, Continual 관점에서 충분히 가능



Models 5

⇅ Sort: Recently updated

instruction-pretrain/finance-Llama3-8B
Text Generation • Updated about 4 hours ago • ⤓271 • ♡ 25

instruction-pretrain/medicine-Llama3-8B
Text Generation • Updated about 4 hours ago • ⤓136 • ♡ 25

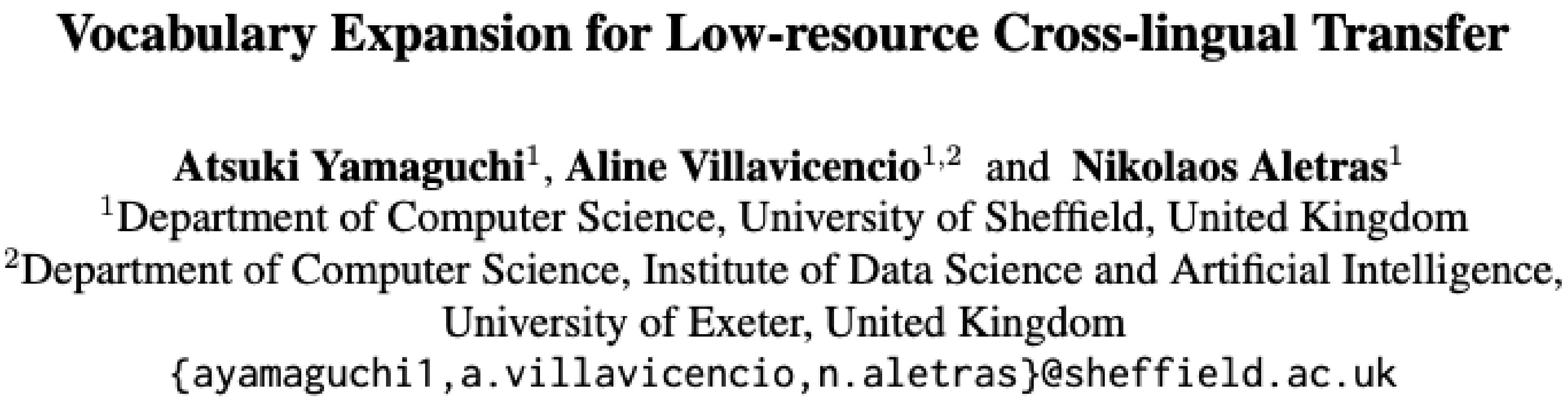
instruction-pretrain/instruction-synthesizer
Text Generation • Updated about 4 hours ago • ⤓230 • ♡ 34

instruction-pretrain/InstructLM-500M
Text Generation • Updated about 4 hours ago • ⤓431 • ♡ 30

instruction-pretrain/InstructLM-1.3B
Text Generation • Updated about 4 hours ago • ⤓443 • ♡ 32

## 무조건 GPT-4를 이용해서 만드는 것이 좋겠지만, 꼭 그게 능사는 아니다

- PT레벨로 데이터 수가 늘어나면 현실적으로 GPT-4를 전부 돌리는 것은 비용적으로 불가능
- 한국어에서 자체 Synthesizer를 구축해서 만들어보는 것도 가능



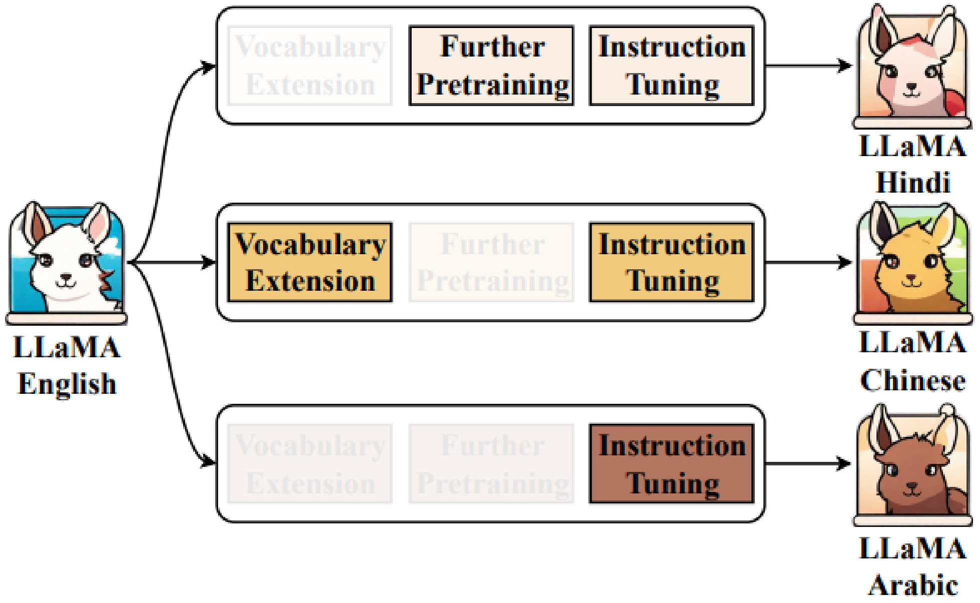
# Vocabulary Expansion for Low-resource Cross-lingual Transfer

**Atsuki Yamaguchi**[1], **Aline Villavicencio**[1,2] and **Nikolaos Aletras**[1]
[1]Department of Computer Science, University of Sheffield, United Kingdom
[2]Department of Computer Science, Institute of Data Science and Artificial Intelligence,
University of Exeter, United Kingdom
{ayamaguchi1,a.villavicencio,n.aletras}@sheffield.ac.uk

# Embedding Initialization

## 기본적으로 Multilingual Model에서 Monolingual Model로 가는 것을 가정

⇒ Vocab Embedding 초기화 어떻게?



## 크게 두가지 방법

- **Vocab Expansion**
- **Vocab Replacement**

⇒ 현재까지의 두가지 방법들은 대규모 코퍼스를 필요로 하는 Language Adaptive Pre-Training(LAPT) 이 필수적

⇒ 그러나, 극한의 Low-resource에서는 이러한 규모의 자원의 사용이 비현실적, 비효율적

이전의 방법보다 더 간단하고, 자원사용을 덜하는 효율적 방법에 대한 systematic study

# Heuristic-based embedding initialization

## Problem Statement

- 목적: 다국어 Source Model을 target 언어에 대해 잘 적합 시키는 것
- Target corpus로부터 tokenizer를 학습시켜서 auxiliary tokenizer 를 생성
- 해당 tokenizer의 vocab으로부터 Source vocab과 겹치지 않는 $V_{new}$들 만을 추출
- $V_s$에서 $V_{new}$를 추가해 target vocabulary $V_t$를 도출 => $V_s + V_{new} = V_t$

## 세가지 휴리스틱한 Rule 기반의 방법을 실험

⇒ Mean, Merge, Token Align

# Method

## Mean

- $V_{new}$ 를 Source tokenizer로 분절되는 vocab embedding의 평균으로 임베딩을 부여

## Merge

- 가정: 기존 subtoken이 결합 시에 도출되는 새 토큰과 임베딩이 유사해야 결합 시 동일한 의미를 낼 수 있을 것
- BPE의 merge rule을 사용: (‘a’, ‘t’) → ‘at’

## Token Alignment

- Target corpus의 문장들을 source, target tokenizer로 분할
- 토큰화된 문장을 비교해서 몇 번째 source token index가 target token index와 align되는지 align list를 생성 (일대 다 매칭)

  예) 32000 => [(1234, 12345), (2345, 23456, 3456)]
- 매칭된 토큰들에 대해서 target corpus에서 등장 빈도를 계산하여 가중 평균을 통해 임베딩을 부여

# Experiments

**이 논문에서는 초점을 맞추는 것은 결국 2가지**

**1. 기존 FOCUS와 같은 외부자원을 이용하는 방법들이 극한 상황의 low-resource setting에서는 제대로 작동하지 않는다**

=> 휴리스틱한 방법만으로도 충분히 잘 작동한다

**2. Low-resource setting Vocab Expansion의 최적화를 위한 조건 탐색**

: the size of $V_{new}$

: the amount of Adaptation Samples

# Setup

## Model

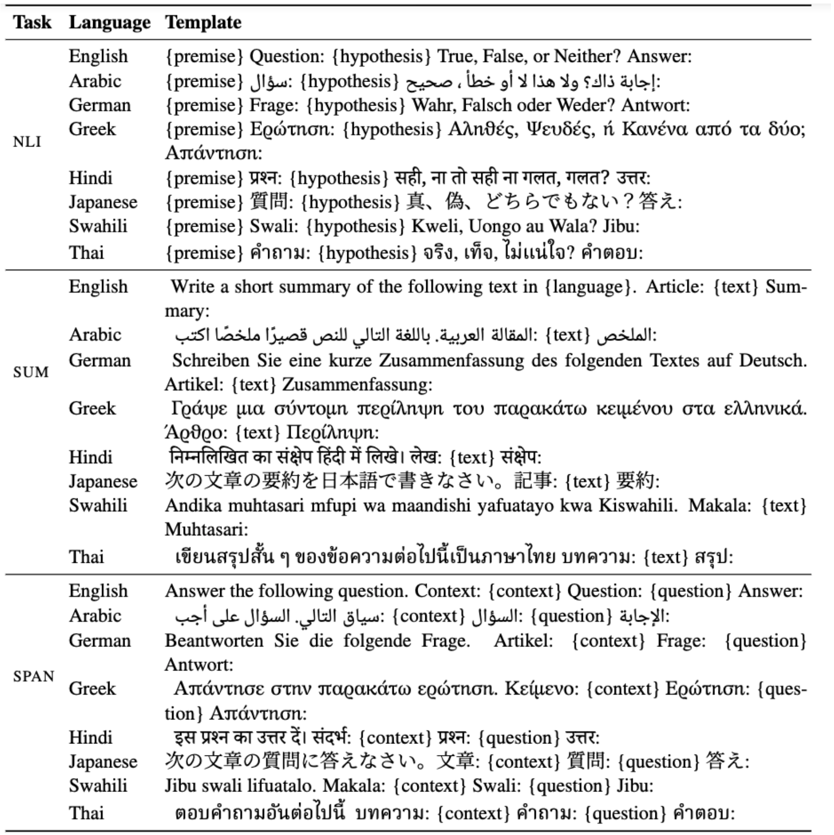
- Llama2 7B / Mistral 7B
- Baseline: Source / LAPT(LoRA) / Random / FOCUS

## Task

- NLI / Summarization / QA

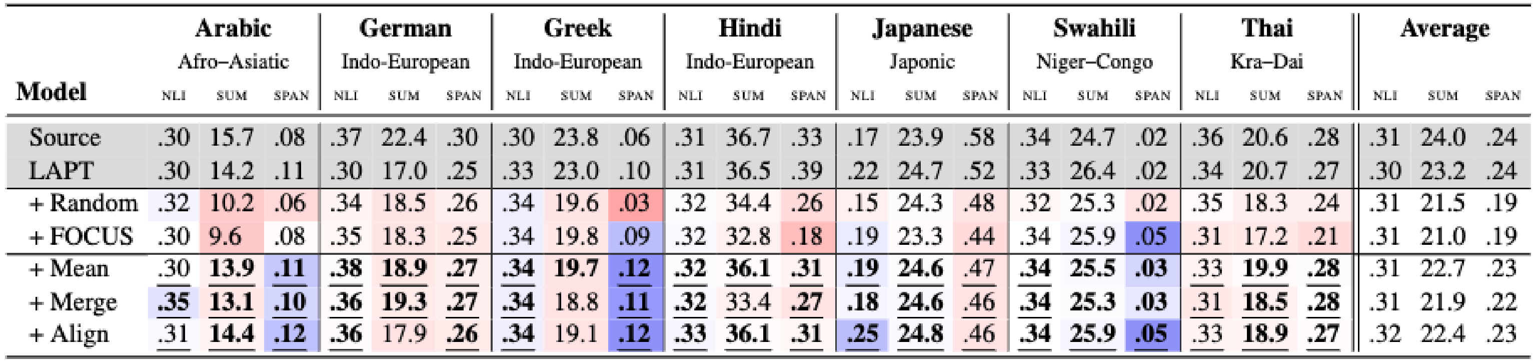
## Target Dataset

- CC-100: low-resource setting $2^{15}$~30k / high-resource setting $2^{20}$~1M sentences
- 기본적으로 추가되는 토큰 갯수 k는 빈도기반 상위 100개



| Task | Language | Template |
|---|---|---|
| NLI | English | {premise} Question: {hypothesis} True, False, or Neither? Answer: |
| | Arabic | {premise} سؤال: {hypothesis} صحيح ، خطأ أو لا هذا ولا ذاك؟ إجابة: |
| | German | {premise} Frage: {hypothesis} Wahr, Falsch oder Weder? Antwort: |
| | Greek | {premise} Ερώτηση: {hypothesis} Αληθές, Ψευδές, ή Κανένα από τα δύο; Απάντηση: |
| | Hindi | {premise} प्रश्न: {hypothesis} सही, ना तो सही ना गलत, गलत? उत्तर: |
| | Japanese | {premise} 質問: {hypothesis} 真、偽、どちらでもない？答え: |
| | Swahili | {premise} Swali: {hypothesis} Kweli, Uongo au Wala? Jibu: |
| | Thai | {premise} คำถาม: {hypothesis} จริง, เท็จ, ไม่แน่ใจ? คำตอบ: |
| SUM | English | Write a short summary of the following text in {language}. Article: {text} Summary: |
| | Arabic | اكتب ملخصًا قصيرًا للنص التالي باللغة العربية. المقالة: {text} الملخص: |
| | German | Schreiben Sie eine kurze Zusammenfassung des folgenden Textes auf Deutsch. Artikel: {text} Zusammenfassung: |
| | Greek | Γράψε μια σύντομη περίληψη του παρακάτω κειμένου στα ελληνικά. Άρθρο: {text} Περίληψη: |
| | Hindi | निम्नलिखित का संक्षेप हिंदी में लिखे। लेख: {text} संक्षेप: |
| | Japanese | 次の文章の要約を日本語で書きなさい。記事: {text} 要約: |
| | Swahili | Andika muhtasari mfupi wa maandishi yafuatayo kwa Kiswahili. Makala: {text} Muhtasari: |
| | Thai | เขียนสรุปสั้น ๆ ของข้อความต่อไปนี้เป็นภาษาไทย บทความ: {text} สรุป: |
| SPAN | English | Answer the following question. Context: {context} Question: {question} Answer: |
| | Arabic | أجب على السؤال التالي. سياق: {context} السؤال: {question} الإجابة: |
| | German | Beantworten Sie die folgende Frage. Artikel: {context} Frage: {question} Antwort: |
| | Greek | Απάντησε στην παρακάτω ερώτηση. Κείμενο: {context} Ερώτηση: {question} Απάντηση: |
| | Hindi | इस प्रश्न का उत्तर दें। संदर्भ: {context} प्रश्न: {question} उत्तर: |
| | Japanese | 次の文章の質問に答えなさい。文章: {context} 質問: {question} 答え: |
| | Swahili | Jibu swali lifuatalo. Makala: {context} Swali: {question} Jibu: |
| | Thai | ตอบคำถามอันต่อไปนี้ บทความ: {context} คำถาม: {question} คำตอบ: |

# Zero-shot performance



| Model | Arabic | | | German | | | Greek | | | Hindi | | | Japanese | | | Swahili | | | Thai | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Afro–Asiatic | | | Indo-European | | | Indo-European | | | Indo-European | | | Japonic | | | Niger–Congo | | | Kra–Dai | | | | | |
| | NLI | SUM | SPAN | NLI | SUM | SPAN | NLI | SUM | SPAN | NLI | SUM | SPAN | NLI | SUM | SPAN | NLI | SUM | SPAN | NLI | SUM | SPAN | NLI | SUM | SPAN |
| Source | .30 | 15.7 | .08 | .37 | 22.4 | .30 | .30 | 23.8 | .06 | .31 | 36.7 | .33 | .17 | 23.9 | .58 | .34 | 24.7 | .02 | .36 | 20.6 | .28 | .31 | 24.0 | .24 |
| LAPT | .30 | 14.2 | .11 | .30 | 17.0 | .25 | .33 | 23.0 | .10 | .31 | 36.5 | .39 | .22 | 24.7 | .52 | .33 | 26.4 | .02 | .34 | 20.7 | .27 | .30 | 23.2 | .24 |
| + Random | .32 | 10.2 | .06 | .34 | 18.5 | .26 | .34 | 19.6 | .03 | .32 | 34.4 | .26 | .15 | 24.3 | .48 | .32 | 25.3 | .02 | .35 | 18.3 | .24 | .31 | 21.5 | .19 |
| + FOCUS | .30 | 9.6 | .08 | .35 | 18.3 | .25 | .34 | 19.8 | .09 | .32 | 32.8 | .18 | .19 | 23.3 | .44 | .34 | 25.9 | .05 | .31 | 17.2 | .21 | .31 | 21.0 | .19 |
| + Mean | .30 | **13.9** | **.11** | **.38** | **18.9** | **.27** | **.34** | **19.7** | **.12** | **.32** | **36.1** | **.31** | **.19** | **24.6** | .47 | **.34** | **25.5** | **.03** | .33 | **19.9** | **.28** | .31 | 22.7 | .23 |
| + Merge | **.35** | **13.1** | **.10** | **.36** | **19.3** | **.27** | **.34** | 18.8 | **.11** | **.32** | 33.4 | **.27** | **.18** | **24.6** | .46 | **.34** | **25.3** | **.03** | .31 | **18.5** | **.28** | .31 | 21.9 | .22 |
| + Align | .31 | **14.4** | **.12** | **.36** | 17.9 | **.26** | **.34** | 19.1 | **.12** | **.33** | **36.1** | **.31** | **.25** | **24.8** | .46 | **.34** | **25.9** | **.05** | .33 | **18.9** | **.27** | .32 | 22.4 | .23 |

Table 1: Zero-shot performance in low-resource settings (30K sentences) on 500 randomly selected test samples for each dataset using LLaMA2-7B as source. The baselines without vocabulary expansion are in grey. **Bold** indicates comparable or better results than Random. Underlined indicates comparable or better results than FOCUS. Darker blue and red shades indicate higher positive and negative relative performance change over Source per language and task, respectively.

- Mean & Align 은 Source / LAPT 와 대등하거나 높은 성능

- 특히 low-resource language(Swahili, Thai, Greek)에서, SOTA인 FOCUS는 오히려 성능이 떨어지는 양상

⇒Low-resource Setting하에서는 제대로 작동하지 않음

- German Sum과 Japanese SPAN에서 급격한 성능 감소

⇒German / Japanese와 같은 **High resource language에서는 Expansion이 효과 미비**

⇒PT단계에서 이미 해당 언어에 대해 학습이 충분히 되었으므로, 오히려 작은 표본의 데이터를 기반의 expansion은 역효과

# Size of New Target Tokens

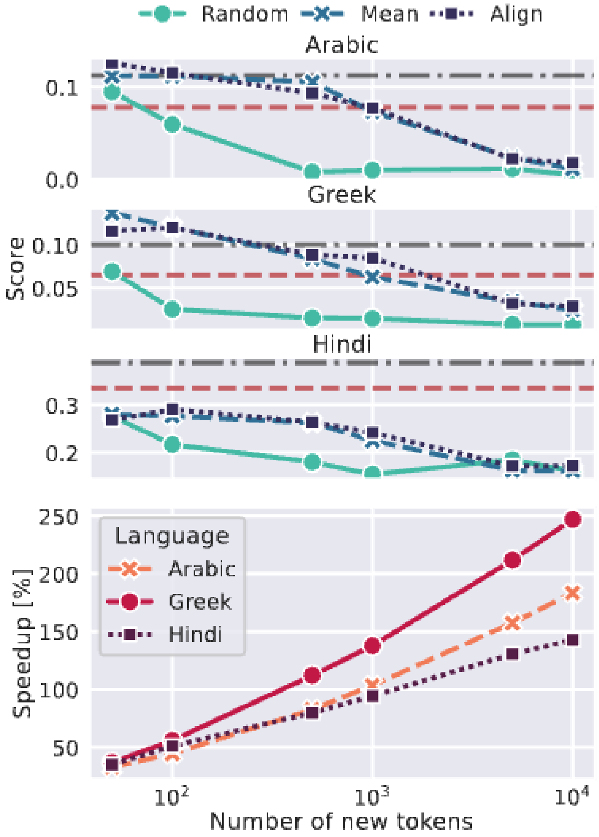


Figure 2: Performance in zero-shot SPAN across different numbers of new target tokens, including inference speedup. Red and grey dotted lines denote the performance of Source and LAPT, respectively.

### What is the proper size of new target tokens?

{50, 100, 500, 1K, 5K, 10K} 에 대한 성능변화와 Inference 속도

- **Performance & Inference Speed Tradeoff**

⇒ New vocab이 커질 수록 성능이 떨어지는 양상

⇒ 휴리스틱 방법은 Random에 비해 robustness를 비교적 잘 유지

### Recommendation

- Setting $|V_{new}|$ **to around 100 to 500** can be a suitable threshold to maintain competitive performance while benefiting from inference Speed

# Adaptation Samples

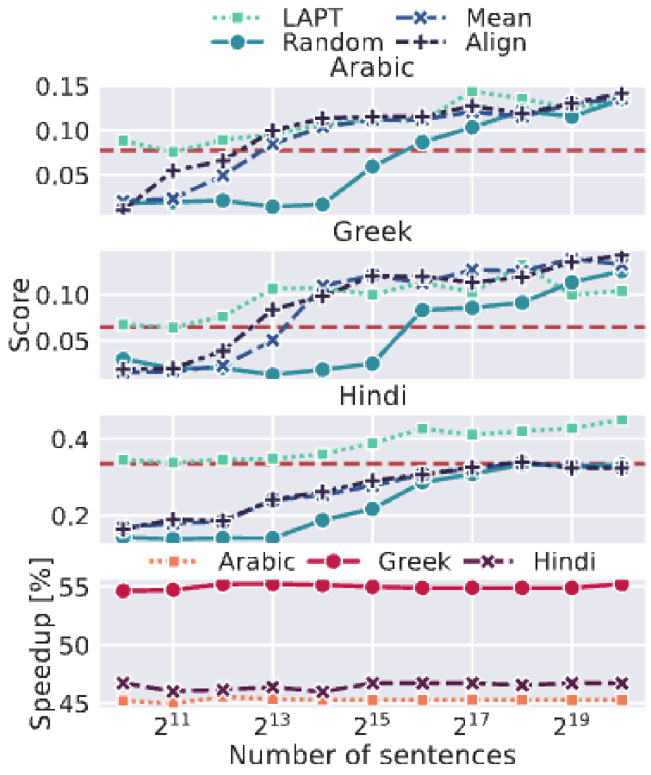


Figure 3: Performance and inference speedup in zero-shot SPAN across different numbers of training sentences. Red dotted line denotes Source performance.

## How many samples are needed for adaptation?

- LAPT는 꾸준히 높은 성능을 유지

⇒ Align 방법이 Mean의 경우보다는 더 빠르게 adaptation

- **$2^{15}$==30k 로도 충분히 source 성능을 넘음**

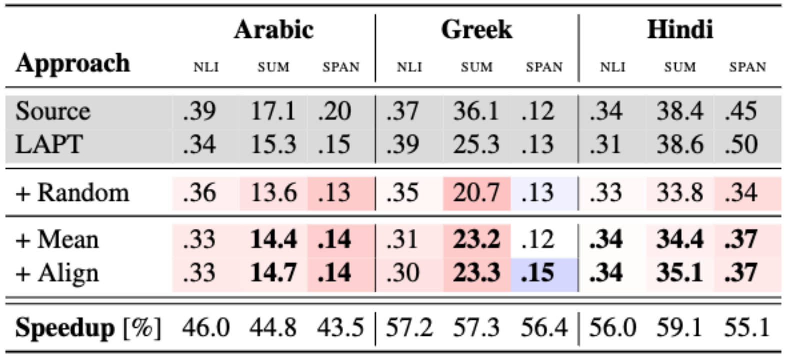
⇒ **30K의 소규모 데이터 만으로도 inference속도를 높이면서, source 모델을 뛰어넘는 성능을 달성이 가능**

## Recommendation

- **Need at least |D| = $2^{14}$ to $2^{15}$** to achieve competitive performance with Source and LAPT when $|V_{new}|$ = 100

- Using Align might provide better performance than Mean

# Other LLMs



| Approach | Arabic |  |  | Greek |  |  | Hindi |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | NLI | SUM | SPAN | NLI | SUM | SPAN | NLI | SUM | SPAN |
| Source | .39 | 17.1 | .20 | .37 | 36.1 | .12 | .34 | 38.4 | .45 |
| LAPT | .34 | 15.3 | .15 | .39 | 25.3 | .13 | .31 | 38.6 | .50 |
| + Random | .36 | 13.6 | .13 | .35 | 20.7 | .13 | .33 | 33.8 | .34 |
| + Mean | .33 | **14.4** | **.14** | .31 | **23.2** | .12 | **.34** | **34.4** | **.37** |
| + Align | .33 | **14.7** | **.14** | .30 | **23.3** | **.15** | **.34** | **35.1** | **.37** |
| **Speedup** [%] | 46.0 | 44.8 | 43.5 | 57.2 | 57.3 | 56.4 | 56.0 | 59.1 | 55.1 |

Table 2: Zero-shot performance and inference speedup using 30K sentences with Mistral-7B as Source. **Bold** indicates comparable or better results than Random. Darker blue and red indicate higher positive and negative relative performance change over Source.

## Mistral-7B

- Source에 비해서 대부분 성능이 저하

⇒ Mistral-7B 자체가 Llama2보다 그 자체적으로 성능이 높기 때문

⇒ Remaining Challenges 로 연결

This implies that different base models could have different requirements of |D| and $|V_{new}|$

⇒ 랜덤보다는 더 좋은 성능

⇒ 각 LLMs별로 학습한 언어와 그 양이 다르기때문에, 모든 LLMs에 일반화는 어렵다. 하지만 동일한 자원을 사용하는 Random의 경우 보다는 좋다.

# Conclusion

### 단일 언어에 대해 Adaptation시에 Vocabulary Expansion에 대한 고찰

- 특히 low-resource에 초점을 맞추어 자원이 극도로 부족한 상황에서의 최적의 대안 찾기
- 고려해야 할 요소: 새로운 토큰의 갯수 / 코퍼스 규모

### 실용적인 관점에서 사용한다기보다는 강력한 베이스라인으로써 고려가 적절

- 모델별 편차가 존재. BPE의 사용하는 경우만 고려
- FOCUS, LAPT의 사용이 불가능하다면 Random보다는 나은 대안

⇒ 모델규모가 지나치게 크지 않는 이상 학습이 불가능한 상황이 있을까?

- High resource setting에서 expansion 자체는 좋은 대안이 아닐 수 있다는 제언

# Thank you

# Q&A