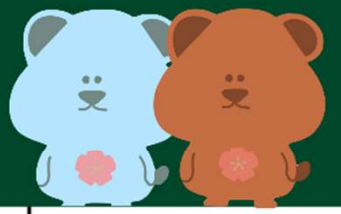


What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers

유런 5기 고급심화팀 황채원

EWHA EURON

#00 Background

EWHA EURON

#00 Background

In-context Learning

: 언어 모델이 지시문, 예제 등 입력 문서(context, prompt)의 의미를 파악해 요약, 번역, 대화 등 구체적인 과제(downstream task)를 해결하는 패러다임

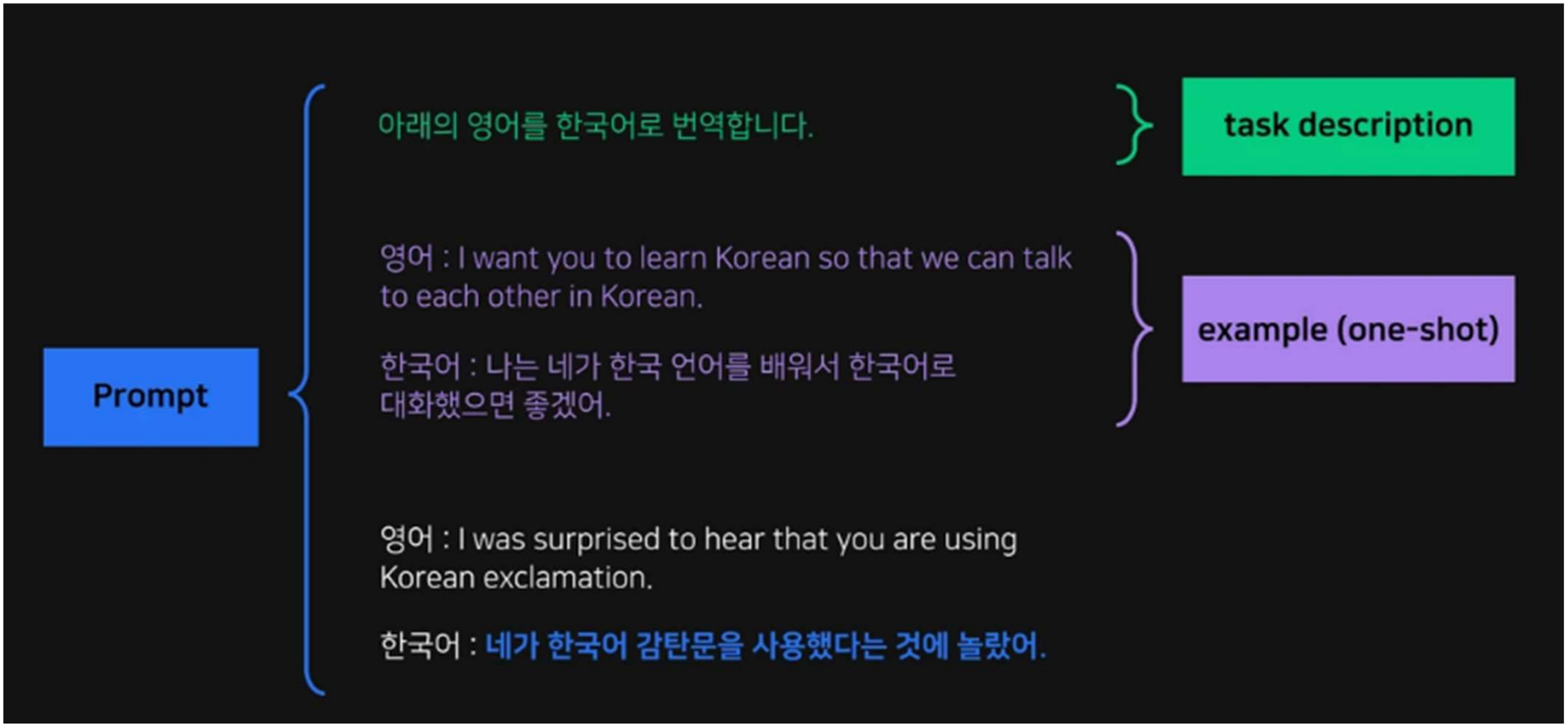
- 파인 튜닝(fine-tuning) 등과는 달리 모델을 업데이트하지 않는다.

https://engineering.clova.ai/posts/2022/05/hyperclova-corpus

EWHA EURON

#00 Background

Zero-shot, one-shot, few-shot



EWHA EURON

#00 Background

BLEU, ROUGE, PPL

BLEU(Bilingual Evaluation Understudy)
: Generated Sentence의 단어가 Reference Sentence에 포함되는 정도

ROUGE(Recall-Oriented Understudy for Gisting Evaluation)
: Reference Setence의 단어가 Generated Sentence에 포함되는 정도

PPL(Perplexity)
: 언어 모델의 성능을 평가하는 정량 지표. 언어 모델이 테스트 문장이 나타날 확률을 높게 예측할수록 PPL이 작아진다. 대개 PPL이 작으면 언어 모델의 성능이 좋은 것으로 알려져 있다.

EWHA EURON

#01 Introduction

EWHA EURON

#01 Introduction

GPT-3 의 이슈

1. **학습 corpus의 92.7퍼센트가 영어에 치우쳤다. 따라서 다른 언어들의 태스크를 수행하기 어렵다.**

2. **13B, 175B 사이즈의 모델에 대한 정보만 있고 다른 사이즈의 모델에 대해선 모른다. 다양한 사이즈의 모델과 그 계산 비용에 대해 아는 것이 LLM의 사용에 유용할 것이라 여겨진다.**

3. **In-context LLM에서 프롬프트 기반의 학습과 튜닝이 시도된 적 없다.**

EWHA EURON

#01 Introduction

Hyper CLOVA를 통해 제안하고자 하는 것
- 한국어 in-context LLM, 820억 개의 파라미터

1. 특정 언어에 적합한 tokenization 기법이 non-English LLM의 학습에 미치는 영향을 파악함. 한국어에 적합한 tokenization strategy 사용.

2. 중간 사이즈의 모델에 zero-shot, few-shot을 이용했고 프롬프트 기반의 튜닝을 향상시킬 수 있음을 제시함.

3. No Code AI – Hyper CLOVA 스튜디오를 통해 ML 전문가가 아니더라도 모델을 쉽게 이용할 수 있도록 하여, 소통으로 인한 비용을 대폭 절감할 수 있는 가능성을 보임.

EWHA EURON

#02 Previous Work

EWHA EURON

#02 Previous Work

Prompt Optimization

1. Discrete : 직관적인 해석에는 용이하지만 최적의 결과를 내지 못할 수 있다.
2. Continuous : 가상의 임베딩 벡터를 설정하여 더 좋은 성능을 낼 수 있다.

Discrete Prompt

Continuous Prompt

https://codingsmu.tistory.com/162

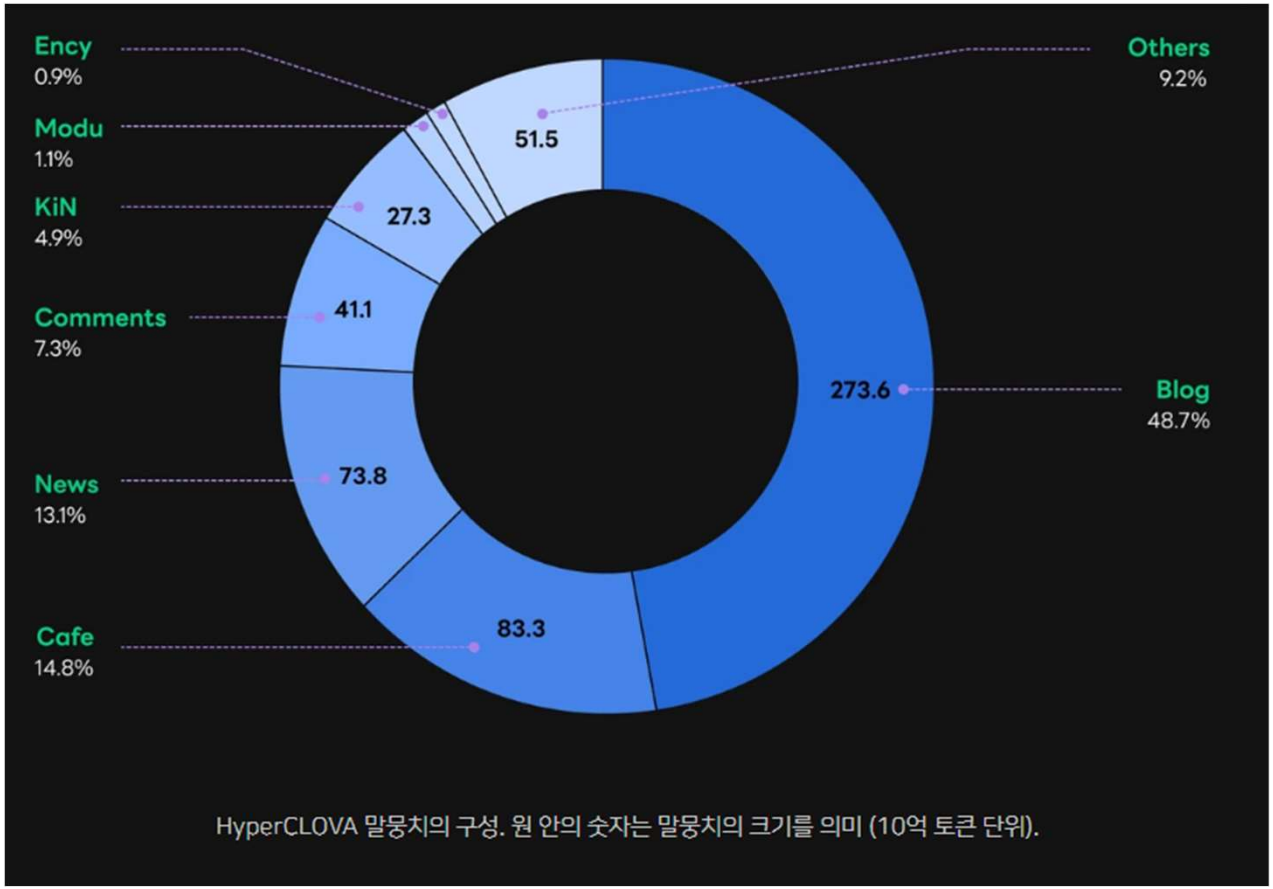
EWHA EURON

#03 Pre-training

EWHA EURON

#03 Pre-training

Data Description



HyperCLOVA 말뭉치의 구성. 원 안의 숫자는 말뭉치의 크기를 의미 (10억 토큰 단위).

Name	Description	Tokens
Blog	Blog corpus	273.6B
Cafe	Online community corpus	83.3B
News	News corpus	73.8B
Comments	Crawled comments	41.1B
KiN	Korean QnA website	27.3B
Modu	Collection of five datasets	6.0B
WikiEn, WikiJp	Foreign wikipedia	5.2B
Others	Other corpus	51.5B
Total		561.8B

Table 1: Descriptions of corpus for HyperCLOVA

EWHA EURON

#03 Pre-training

데이터 전처리에서 흥미로웠던 부분

- 각 문서의 퀄리티를 측정하기 위해 LR 모델을 학습시킴.
- high qualit의 encyclopedia documents -> positive
- crawled web documents -> negative

- 중복된 문서 삭제를 위해 해시함수로 유사도를 측정함.
- 리뷰타입의 문서는 반복된 표현이 과도하게 많았음.

- 뉴스 문서는 구독자의 감정적 동요를 불러일으키는, 불필요한 표현들을 삭제

- Data Anonymization : 주민번호, 이메일 주소, 전화번호 등의 개인정보는 마스킹.
- 그러나, 개인을 특정할 수 없는 [주민번호의 지역과 나이, 성별을 나타내는 부분], [이메일 주소의 도메인], [전화번호의 다이얼 코드] 등의 정보는 남겨놓았음.

EWHA EURON

#03 Pre-training

Korean tokenization

토큰화(tokenization)란?
- 주어진 코퍼스(corpus)에서 토큰(token)이라 불리는 단위로 나누는 작업

엘리스는
엘리스가
엘리스에게

<- **교착어인 한국어에서 단어는**
의미적 기능을 하는 부분과 문법적 기능을 하는 부분이 나뉜다.

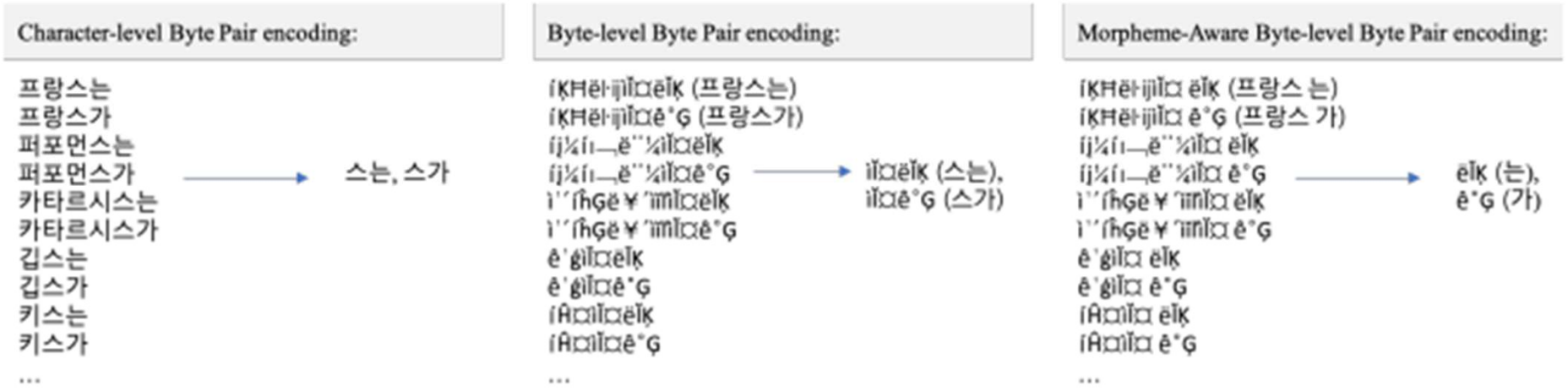
먹다
먹었다
먹는다

EWHA EURON

#03 Pre-training

Korean tokenization

- 세 종류의 토큰화 작업 중 형태소(morpheme) 단위의 토큰화를 선택



#04 Experimental Results

#04 Experimental Results

평가 대상 과제

1, NSMC 영화리뷰감성분석

2. KorQuAD 문서독해

3. AiHub Ko->En 한영번역

4. AiHub En->Ko 영한번역

5. KLUE-YNAT 뉴스제목분류

EWHA EURON

#04 Experimental Results

	NSMC (Acc)	KorQuAD (EA / F1)		AI Hub (BLEU) Ko→En	AI Hub (BLEU) En→Ko	YNAT (F1)	KLUE-STS (F1)
Baseline	89.66	74.04	86.66	40.34	40.41	82.64	75.93
137M	73.11	8.87	23.92	0.80	2.78	29.01	59.54
350M	77.55	27.66	46.86	1.44	8.89	33.18	59.45
760M	77.64	45.80	63.99	2.63	16.89	47.45	52.16
1.3B	83.90	55.28	72.98	3.83	20.03	58.67	60.89
6.9B	83.78	61.21	78.78	7.09	27.93	67.48	59.27
13B	87.86	66.04	82.12	7.91	27.82	67.85	60.00
39B	87.95	67.29	83.80	9.19	31.04	71.41	61.59
82B	88.16	69.27	84.85	10.37	31.83	72.66	65.14

- 모델 사이즈가 커짐에 따라 성능도 좋아진다. 그러나 large scale이 아니더라도 프롬프트 엔지니어링을 이용한다면 성능을 충분히 개선할 수 있다.

- 단, 한영 번역과 KLUE의 경우 베이스라인 모델에 비해 성능이 확연히 떨어지는 결과가 나왔다.

- 이에 대해 논문에서는 영어 코퍼스 학습량의 부족을 그 원인으로 예측했고, 프롬프트 엔지니어링으로 성능을 향상시킬 수 있는 부분이라고 이야기했다.

EWHA EURON

#04 Experimental Results

Effect of Tokenization

OOV란? Out-Of-Vocabulary

- 모델이 사전훈련 중에 접하지 못한 단어나 구

1. OOV가 있다면 모델은 해당 문장의 의미를 완전히 이해하지 못할 수 있다.

2. OOV는 모델의 어휘사전에 없기 때문에 해당 단어를 대체하거나 무시하면서 정보의 손실이 발생할 수 있다.

3. 언어는 계속 변화하는데, 새로운 표현이 OOV로 처리되어 언어의 변화가 모델에 반영되지 못한다.

EWHA EURON

#04 Experimental Results

Effect of Tokenization
: Char-level, byte-level, morpheme-aware byte-level 비교

- 영어에서는 char-level, byte-level이 주로 사용된다. 그러나 한국어에서 char-level BPE의 사용은 OOV를 발생시킨다.

- YNAT에서 보이는 char-level과 byte-level의 격차. 이는 뉴스 기사 헤드라인에 쓰인 단어들을 char-level로 토큰화 할 때 문제가 발생하기 때문.

- 해당 언어에 맞는 토큰화 방식을 사용하는 것이 모델의 성능에 영향을 미친다는 결과가 도출되었다.

	KorQuAD (EA / F1)		AI Hub (BLEU) Ko→En	AI Hub (BLEU) En→Ko	YNAT (F1)	KLUE-STS (F1)
Ours	**55.28**	**72.98**	3.83	**20.03**	**58.67**	**60.89**
byte-level BPE	51.26	70.34	**4.61**	19.95	48.32	60.45
char-level BPE	45.41	66.10	3.62	16.73	23.94	59.83

EWHA EURON

#04 Experimental Results

Effect of Tokenization

sentence
마감이 잘 안돼서 옆부분이 안맞은게
불편했고 흰색이라 어쩔 수 없긴 하지만 때도
잘타요.

tokenized by Character-level BPE Tokenizer
['마감이</w>', '잘</w>', '안돼서</w>', '옆', '부분이</w>', '안', '맞은',
'게</w>', '불편', '했고</w>', '흰색', '이라</w>', '어', '<unk>', '수</w>', '없',
'긴</w>', '하지만</w>', '때도</w>', '잘', '타', '요.</w>']

tokenized by Byte-level BPE tokenizer
['ëŠ', 'ê°IJiL', 'ĠìĹ', 'ĠìķĪëı¼ìĦł', 'ĠìĪ', 'ë¶Ġë¶ĦiL', 'ĠìķĪëŠ', 'ìŁĠ', 'ëª',
'Ġë¶Ĺìĺ,', 'íĸĪêª', 'ĠiL°íĥ', 'ìL´ëŁ¼', 'ĠìĸˇìŒĶ', 'ĠìĪ', 'ĠìĹĪ', 'ê,', 'ĠìķĺìşĠëŠ',
'ĠëķĮëıĦ', 'ĠìĹĺ', 'íĥĠ', 'ìĮĶ', '.']

tokenized by Morpheme-Aware Byte-level BPE tokenizer
['ëŠĪê°IJ', 'ìL', 'ĠìĹĺ', 'ĠìķĪëı¼ìĦł', 'ĠìĪ', 'ë¶Ġë¶Ħ', 'ìL', 'ĠìķĪ', 'ëŠìĺĿĠ', 'ëª',
'Ġë¶Ĺìĺ,', 'íĸĪêª', 'ĠiL°íĥ', 'ìL´ëŁ¼', 'ĠìĸˇìŒĶ', 'ĠìĪ', 'ĠìĹĪê,', '', 'ĠìķĺìşĠëŠ',
'ĠëķĮ', 'ëıĦ', 'ĠìĹĺ', 'íĥĠ', 'ìĮĶ', '.']

sentence
마감이 잘 안돼서 옆부분이 안맞은게
불편했고 흰색이라 어쩔 수 없긴 하지만 때도
잘타요.

tokenized by Character-level BPE Tokenizer
['마감이</w>', '잘</w>', '안돼서</w>', '옆', '부분이</w>', '안', '맞은',
'게</w>', '불편', '했고</w>', '흰색', '이라</w>', '어', '<unk>', '수</w>', '없',
'긴</w>', '하지만</w>', '때도</w>', '잘', '타', '요.</w>']

tokenized by Byte-level BPE tokenizer
['마', '감이', ' 잘', ' 안돼서', ' 옆', '부분이', ' 안맞', '은', '게', ' 불편', '했고', '
흰색', '이라', ' 어쩔', ' 수', ' 없', '긴', ' 하지만', ' 때도', ' 잘', '타', '요', '.']

tokenized by Morpheme-Aware Byte-level BPE tokenizer
['마감', '이', ' 잘', ' 안돼서', ' 옆', '부분', '이', ' 안', '맞은', '게', ' 불편', '했고', '
흰색', '이라', ' 어쩔', ' 수', ' 없긴', ' 하지만', ' 때', '도', ' 잘', '타', '요', '.']

EWHA EURON

#05 Discussion on Industrial Impacts

EWHA EURON

#05 Discussion on Industrial Impacts

기존 파이프라인의 문제
- LLM을 이용한 플랫폼을 개발하기 위해서는,
프로그래밍 지식이 없는 사람들과의 소통이 필수적이며 이 과정에서 큰 비용이 발생한다.

Hyper CLOVA Studio는 GUI와 API를 제공하여 ML 엔지니어의 개입을 최소화한다.
No Coding으로AI 기반 서비스의 프로토타입을 빠르게 완성할 수 있도록 돕는다.

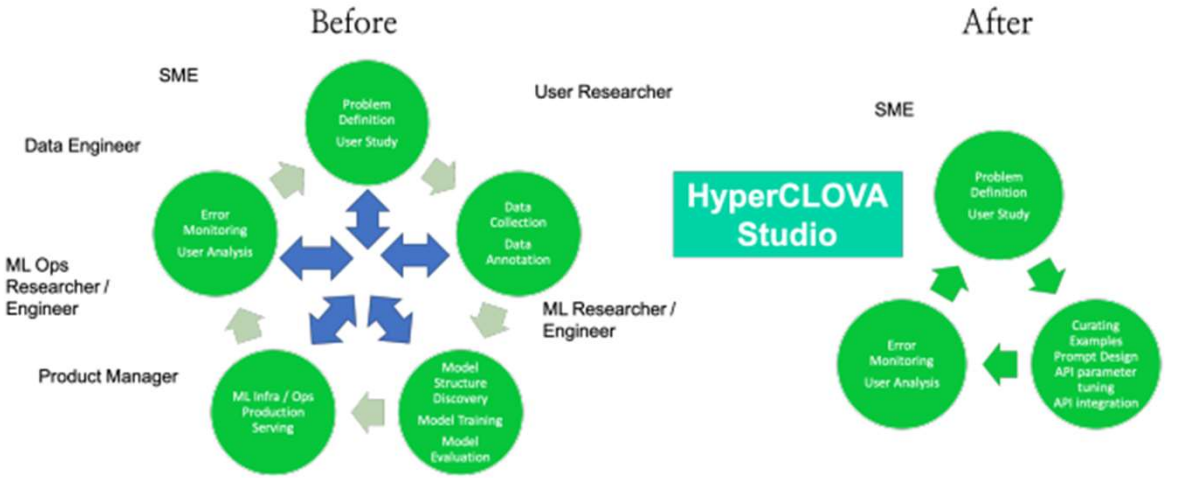


Figure 4: No Code AI paradigm in HyperCLOVA Studio.

① 문제정의
② 데이터 수집과 요약
③ 모델 학습 및 평가
④ MLOps
⑤ 모니터링 및 유지관리

EWHA EURON

#05 Discussion on Industrial Impacts

1. **Rapidly Prototyping Chatbots with Personalities - 캐릭터성을 가진 챗봇 프로토타입 생성**

2. **Zero-shot Transfer Data Augmentation – 제로샷 전이 데이터 증강. 사용자의 의도에 맞는 발화 생성.**

3. **Event Title Generation – 이벤트 문구 생성**

(a) Character bot

Child: How about Penguin?
Penguin: Hey, I'm the best, ahem.
(...)

Child: I'm happy because I didn't go to school because of corona
Penguin: Still, you have to study because you're young! That way, you will be cool like me!
Child: Did you eat tuna yester?
Penguin: Ah~ That's a secret...

(b) Zero-shot transfer data augmentation

intent: Reservation inquiry

#1: Will it be reserved for a room?
#2: If you don't have a seat, it's okay to sit outside.
#3: Is it possible for a group seat?
#4: What is the most common menu for dinner?
#5: You want to make a reservation for the weekend, but do you have a lot of customers?
#6: Do I have to order by number of people?

(c) Title of promoting event

tag: Toggle Bar Necklace, Half and Half Chain Necklace, Cubic Earrings, Gemstone Earrings, Drop Earrings, One Touch Ring Earrings, Chain Silver Ring, Onyx Earrings, Pearl Earrings, Heart Earrings
time: December 19th

Title: Jewelry for you who shines brightly

EWHA EURON

#06 Conclusion

EWHA EURON

#06 Conclusion

LMs concerning issues

1. LLM의 오용 – 이루다 사례와 같은 사생활 침해, 혐오 발언

2. 공정성, 편향, 대표성 – 훈련 데이터의 편향이 LLM에도 영향을 미친다. 편향을 줄이기 위한 데이터 전처리, 혹은 문장 생성 시의 필터링이 필요하다.

3. 과도한 에너지 소비 – LLM의 훈련에는 많은 에너지가 소모된다. 효율적인 하드웨어의 사용이 요구된다.

긍정적 방향으로 나아가기 위한 노력

No/Low Code AI의 가능성 – AI에 대한 접근성을 높여 많은 이들에게 AI의 혜택을 가져다 줄 수 있다.

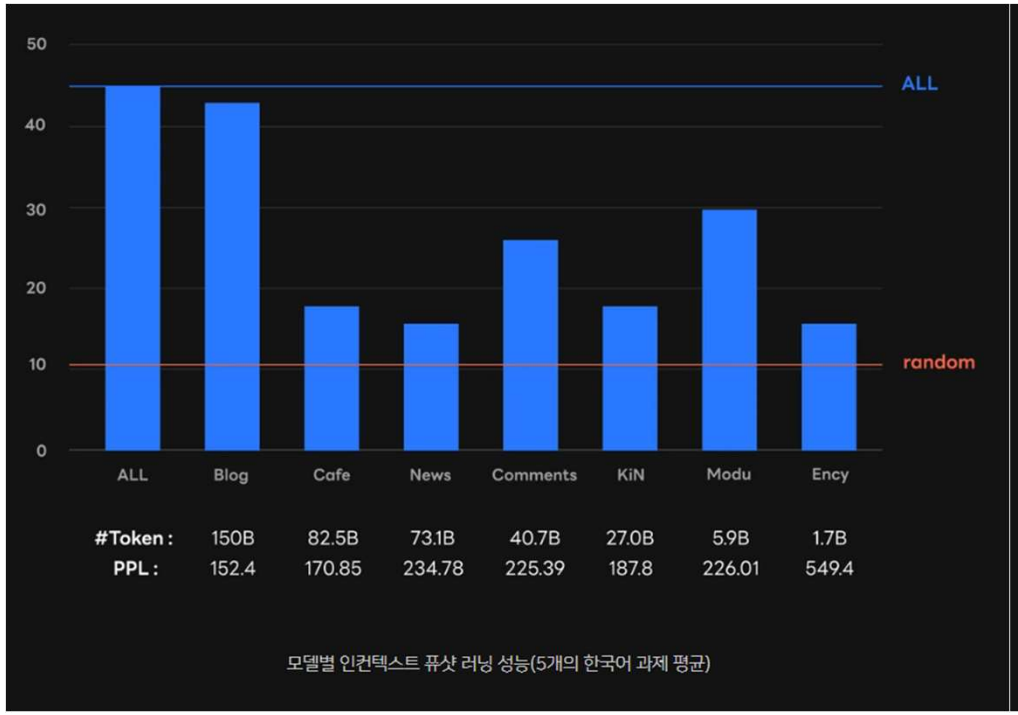
EWHA EURON

#07 ETC

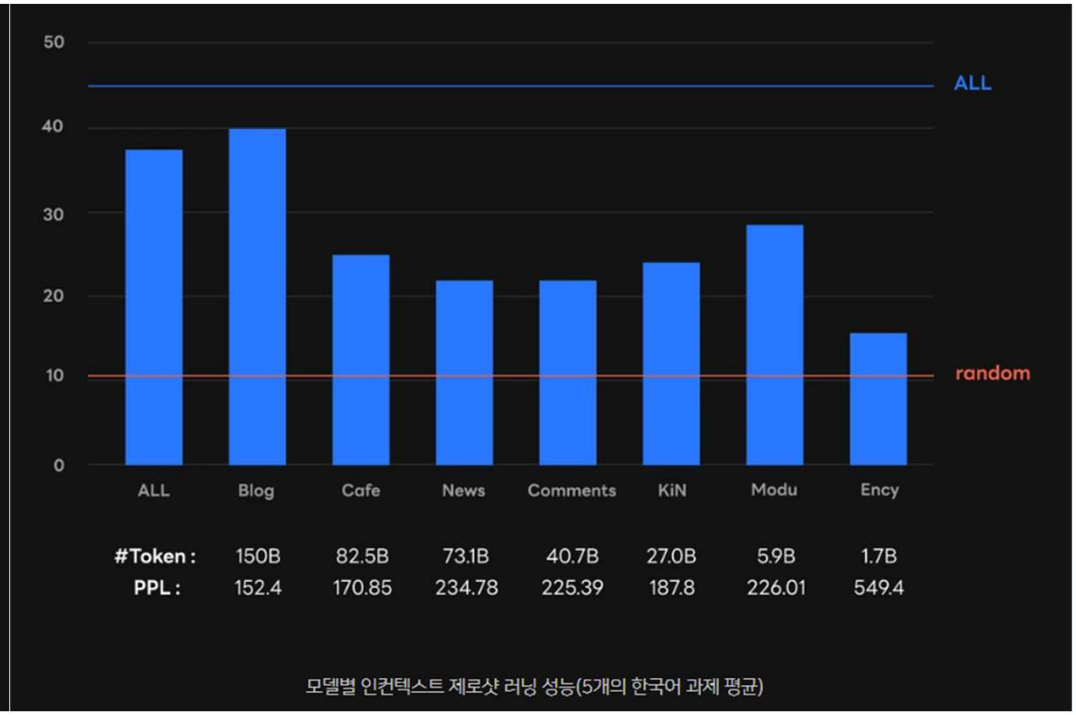
EWHA EURON

#07 ETC

발견 1: 말뭉치에 따라 성능이 크게 달라지며, PPL이 낮다고 성능이 꼭 좋은 것은 아니다



모델별 인컨텍스트 퓨샷 러닝 성능(5개의 한국어 과제 평균)

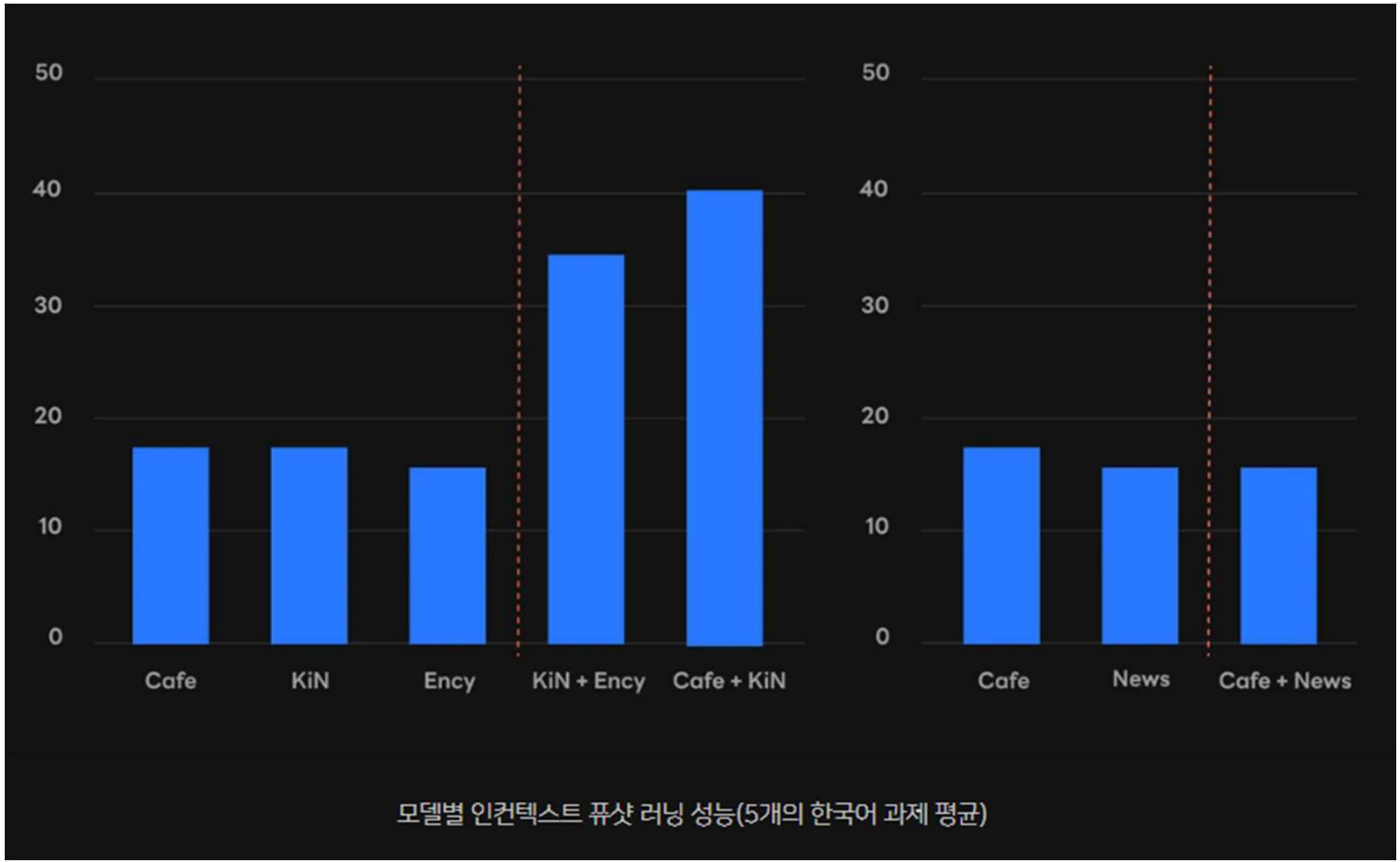


모델별 인컨텍스트 제로샷 러닝 성능(5개의 한국어 과제 평균)

#07 ETC

발견 2: 말뭉치를 잘 섞으면 없던 능력이 생기기도 한다

- 말뭉치를 잘 섞으면 인컨텍스트 러닝 능력이 생기는 경우를 발견

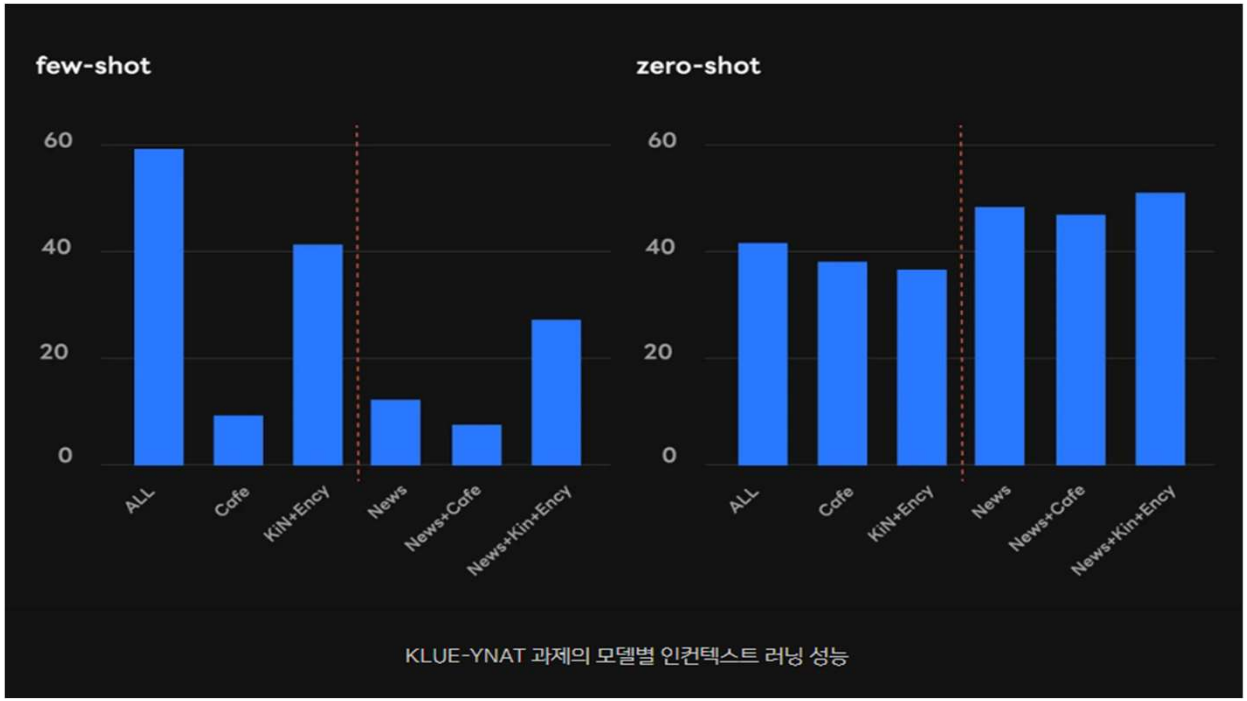


모델별 인컨텍스트 퓨샷 러닝 성능(5개의 한국어 과제 평균)

#07 ETC

발견 3: 과제와 비슷한 말뭉치가 사전 훈련에 포함된다고 해서 높은 성능을 보장하지는 않는다

- 뉴스제목을 바탕으로 토픽을 예측하는 과제에서 뉴스가 많이 배운 모델의 성능이 높지 않았다.



KLUE-YNAT 과제의 모델별 인컨텍스트 러닝 성능