What Changes Can Large-scale Language Models Bring?

Intensive Study on HyperCLOVA : Billions-scale Korean Generative Pretrained Transformers

Sunyoung Song

Abstract

- GPT-3 shows remarkable in-context learning ability of **large-scale language models** (LMs) **trained on hundreds of billion scale data**.

- address some remaining **issues less reported** by the GPT-3 paper.
  - a non-English LM
  - the performances of different sized models
  - the effect of recently introduced prompt optimization on in-context learning

- **HyperCLOVA** : a Korean variant of 82B GPT-3 trained on a Korean-centric corpus of 560B tokens.

1. Introduction

- Due to its remarkable zero-shot and few-shot performances, GPT-3's in-context learning has gained significant attention in the AI community.

- **Issues of using GPT-3**
  - the language composition of the training corpus is **heavily skewed towards English** with 92.7%.
  - only **can have access to** a thorough analysis of models of **13B and 175B** (Brown et al., 2020) but none in between.
  - advanced prompt-based learning methods **have not yet been experimented** for an in-context large-scale LM learner

- Zero-shot learning (ZSL) : Learning to classify data that you've never seen before
- Few-shot learning (FSL) : Learning with a small amount of data.

1. Introduction

- **HyperCLOVA**
  - a Korean in-context large-scale LM with 82B parameters.
  - collect 561B tokens of Korean corpus.
  - use byte-level BPE with a morpheme analyzer.

- **P-tuning** (Prompt Tuning)

- the versatility of operating a single large-scale LM in the AI industry

- **HyperCLOVA Studio**
  - an interactive prompt engineering interface

1. Introduction - contributions

1. introduce HyperCLOVA, a **large-scale Korean in-context learning-based LM** with nearly 100B parameters, by constructing a large Korean-centric corpus of 560B tokens.

2. discover the effect of **language-specific tokenization** on large-scale in-context LMs for training corpus of non-English languages.

3. explore the **zero-shot and few-shot capabilities** of mid-size HyperCLOVA with 39B and 82B parameters and find that **prompt-based tuning** can enhance the performances, outperforming state-of-the-art models on downstream tasks when backward gradients of inputs are available.

4. argue the possibility of realizing **No Code AI** by designing and applying HyperCLOVA Studio to our in-house applications. We will release HyperCLOVA Studio with input gradients, output filters, and knowledge injection.

2. Previous Work – (1) Prompt Optimization

- As the scale of language models grows, learning via prompts is efficient regarding time and space complexity.
  -> the potential of **replacing the full finetuning paradigm with the prompt-based approach** has been reported.

- Prompt optimization : **discrete** and **continuous** approaches.
  - **discrete approach**
    - optimizes directly on the token space and has the advantage of transferability.
    - however, it has poor interpretability.
      -> so, we aims to **optimize prompts in the continuous space.**

2. Previous Work – (2) Language Model

- Language-specific language models are still in demand.
- However, due to **high cost**, language-specific language models other than English **are limited in availability.**

- **Problems**:
  - multilingual in-context learners are not even explored yet.
  - being focused on few major languages such as Chinese.

3. Pre-training – (1) Data Description

- **gathered** all available **text data**
  - user-generated content (UGC) and contents provided by external partners with no violation of legal issues from both diverse services of NAVER and external sources.

- collected a total of **561B tokens** as the final corpus.
- The corpus was **randomly sampled** for pre-training.

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

- Korean 97% / English 2%
  Japanese 0.5% / Others 0.5%

3. Pre-training – (2) Model and Learning

- Use the same transformer decoder architecture as GPT-3 of OpenAI.

- **Modeling**
  - based on megatron-LM
  - trained on the NVIDIA Superpod
  - use AdamW with cosine learning rate scheduling and weight decay as an optimizer

- **Learning**
  - It takes 13.4 days to train a model with 82B parameters with 150B tokens.

3. Pre-training – (3) Korean Tokenization

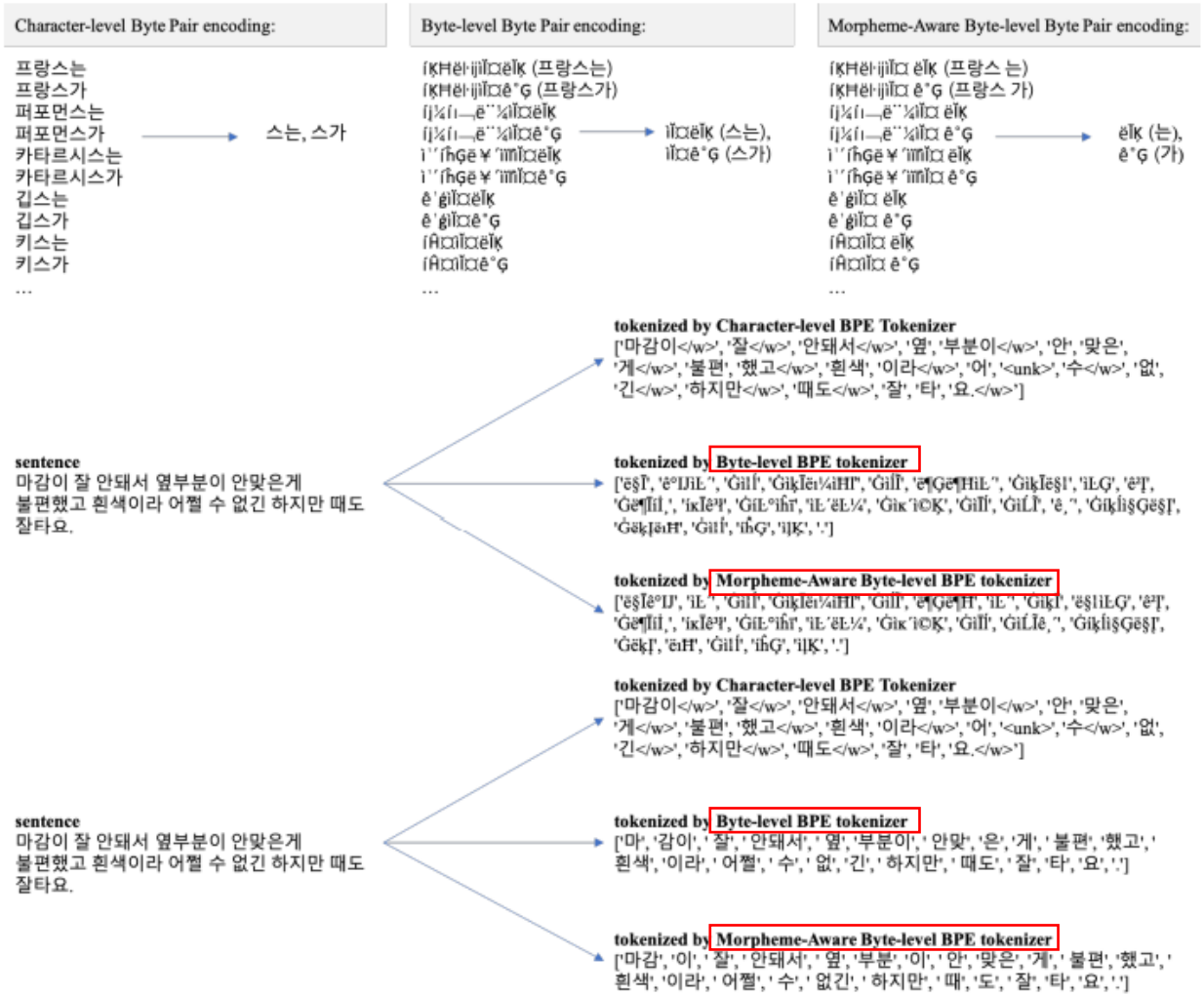
- In Korean, Properly tokenizing noun and particle, and stems and endings clarifies the semantics of each token.

- Tokenization method :
  - using **morpheme-aware byte-level BPE**.

- We pre-split sentences.
  - excludes most of non-Korean characters.

  - Using parts of the sentence presplit by our morpheme analyzer, our morpheme aware byte-level BPE learns the sentence in which most non-Korean characters are expressed as single byte characters

3. Pre-training – (3) Korean Tokenization



4. Experimental Results - (1) Setting

- Use five datasets(3 + KLUE 2) and one additional in-house dataset

- **NSMC**
  - a movie review dataset from NAVER Movies.
  - traning data: 150K / test data : 50K

  - generate 12 sets, and each set consists of 70 examples randomly sampled from the training set.
  - average the test accuracies of 12 in-context 70-shot learning models

- **KorQuAD 1.0**
  - a Korean version of machine reading comprehension dataset.
  - 10,645 training passages.
    - 66,181 training questions / 5,774 validation questions

  - Evaluation : scheme of SQuAD v2.0 which uses test paragraph, corresponding sample four question-answer pairs, and test question as the input to GPT-3.

4. Experimental Results - (1) Setting

- **AI Hub Korean-English** corpus
  - constist of 800K sentence pairs :
  - we randomly sample 1K pairs for evaluating on Ko → En and En → Ko translation tasks.
    - performed three random trials for each translation task.

  - Evalution : BLEU score

- **YNAT (one of the KLUE Benchmark tasks)**
  - Yonhap News Agency Topic Classification
  - a topic classification problem with seven classes
  - consists of 45K, 9K, and 9K annotated headlines for training, valid, and test sets
  - average the test accuracies of 3 in-context 70-shot learners.

4. Experimental Results - (1) Setting

- **KLUE-STS (one of the KLUE Benchmark tasks)**
  - predict a sentence similarity between each pair of sentences, where the similarity score has a value between 0 and 5.

  - use F1 score after binarizing the real-valued similarity as suggested in the KLUE paper
  - average the test accuracies of 3 in-context 40-shot learners

- **Query modification task**
  - for AI speaker users.
  - Target : the case where a single-turn FAQ system is already operating in AI Speakers
  - Goal : to convert the multi-turn query to a single-turn query, which can then be understood by a single-turn AI speaker.

  - 1,326 test instances

4. Experimental Results - (1) Setting

Example 1:
사용자: 아이유 노래 틀어줘 (User: Play IU's track)
스피커: 노래를 재생합니다. (AI Speaker: I am playing the track.)
사용자: 몇 살이야 (User: How old?)
사용자의 최종 의도: 아이유 몇 살이야 **(Modified query: How old is IU?)**
Example 2:
사용자: 비행기는 누가 만들었어 (User: Who invented airplane?)
스피커: 라이트형제요. (AI Speaker: Wright brothers did.)
사용자: 동생 이름 뭐야 (User: What is the younger's name?.)
사용자의 최종 의도: 라이트 형제 동생 이름 뭐야? **(Modified query: What is the younger one's name of Wright brothers?)**

Table 11: Examples of user query modified by Hyper-CLOVA. English sentences are translated by a human expert.

[P][P][P][P][P][P][P][P][P][P]

예제1
(# Example 1)
사용자: 아이유 앨범 뭐있어
(User: What are the names of some albums of IU?)
스피커: 아이유의 대표 앨범으로는 Love poem, Palette, CHAT-SHIRE가 있어요.
(AI Speaker: IU's signiture albums include Love poem, Palette, and CHAT-SHIRE.)
사용자: 가장 신나는 앨범이 뭐야
(User: Which one is the most exciting album?)
–
[P][P][P] 사용자의 [P][P] 의도: Love poem, Palette, CHAT-SHIRE 중 가장 신나는 앨범이 뭐야
([P][P][P] User's [P][P] intent: Among Love poem, Palette,and CHAT-SHIRE, which one is the most exciting album?)

예제2
(# Example 2)
사용자: 평창 동계올림픽은 몇년에 했어?
(User: When did the PyeongChang Olympics take place?)
스피커: 2018년입니다.
(AI Speaker: It is 2018.)
사용자: 그때 미국 대통령이 누구야
(User: Who was the president of the United States at that time?)
–
[P][P][P] 사용자의 [P][P] 의도: 2018년 미국 대통령이 누구야
([P][P][P] User's [P][P] intent: Who was the president of US in 2018?)

예제3
(Example 3)
사용자: 삼성전자 주가 얼마야
(User: What is Samsung Electronics' share price?)
스피커: 8만2천원입니다.
(AI Speaker: It is 82,000 Won.)
사용자: LG전자는
(User: How about LG Electronics?)
–
[P][P][P] 사용자의 [P][P] 의도: LG전자 주가 얼마야
([P][P][P] User's [P][P] intent: What is LG Electronics' share price?)

예제4
(Example 4)

Table 12: Used prompts of query modification task. [P] denotes a token for continuous prompt.

4. Experimental Results - (2) In-context Few-shot Learning.

	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

Table 3: Results of in-context few-shot tasks on question answering, machine translation, topic classification, and semantic similarity per model size. As baselines, we report the results of BERT-base for NSMC and KorQuAD, and Transformer for AI Hub from Park et al. (2020). mBERT is used for KLUE-YNAT and KLUE-STS from Park et al. (2021).

4. Experimental Results - (3) Prompt-based Tuning

Methods	Acc
Fine-tuning	
mBERT (Devlin et al., 2019)	87.1
w/ 70 data only	57.2
w/ 2K data only	69.9
w/ 4K data only	78.0
BERT (Park et al., 2020)	89.7
RoBERTa (Kang et al., 2020)	91.1
Few-shot	
13B 70-shot	87.9
39B 70-shot	88.0
82B 70-shot	88.2
p-tuning	
137M w/ p-tuning	87.2
w/ 70 data only	60.9
w/ 2K data only	77.9
w/ 4K data only	81.2
13B w/ p-tuning	91.7
w/ 2K data only	89.5
w/ 4K data only	90.7
w/ MLP-encoder	90.3
39B w/ p-tuning	**93.0**

Table 4: Comparison results of p-tuning with fine-tuned LMs and in-context few-shot learning on NSMC. MLP-encoder means the result of replacing LSTM with MLP as the p-tuning encoder on 150K NSMC training data.

Model sizes	Few-shots	p-tuning	BLEU
13B	zero-shot	×	36.15
		O	**58.04**
	3-shot	×	45.64
		O	**68.65**
39B	zero-shot	×	47.72
		O	**73.80**
	3-shot	×	65.76
		O	**71.19**

Table 5: Results of p-tuning on in-house query modification task.

4. Experimental Results - (4) Effect of Tokenization

- analyze the **effects of morpheme-aware byte-level BPE**
- As baselines, we use **byte-level BPE** and **char-level BPE**.
- char-level BPE refers to the original BPE.
  - Problems :
    - out-of-vocabulary (OOV)
    - some Korean character is not included in char-level BPE tokens (ex. "젝")

- use models of 1.3B parameters

morpheme-aware byte-level BPE

	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

Table 6: Effects of tokenization approaches on three tasks. HyperCLOVA-1.3B is used for evaluation.

5. Discussion on Industrial Impacts

- What change can large-scale LMs bring?
  -> accelerating the life-cycle of NLP ML operation

1. **HyperCLOVA Studio**
   - the place for building and **communicating the shared artifact** generated by HyperCLOVA
   - functions :
     - provide a **GUI interface**, like the OpenAI Playground.
     - support API end point in which the output can be easily acquired by an API call with diverse functions, including ones not yet provided by OpenAI Playground.

   - can be **rapid prototyping** of AI-based services while **minimizing the involvement of ML engineers.**

5. Discussion on Industrial Impacts

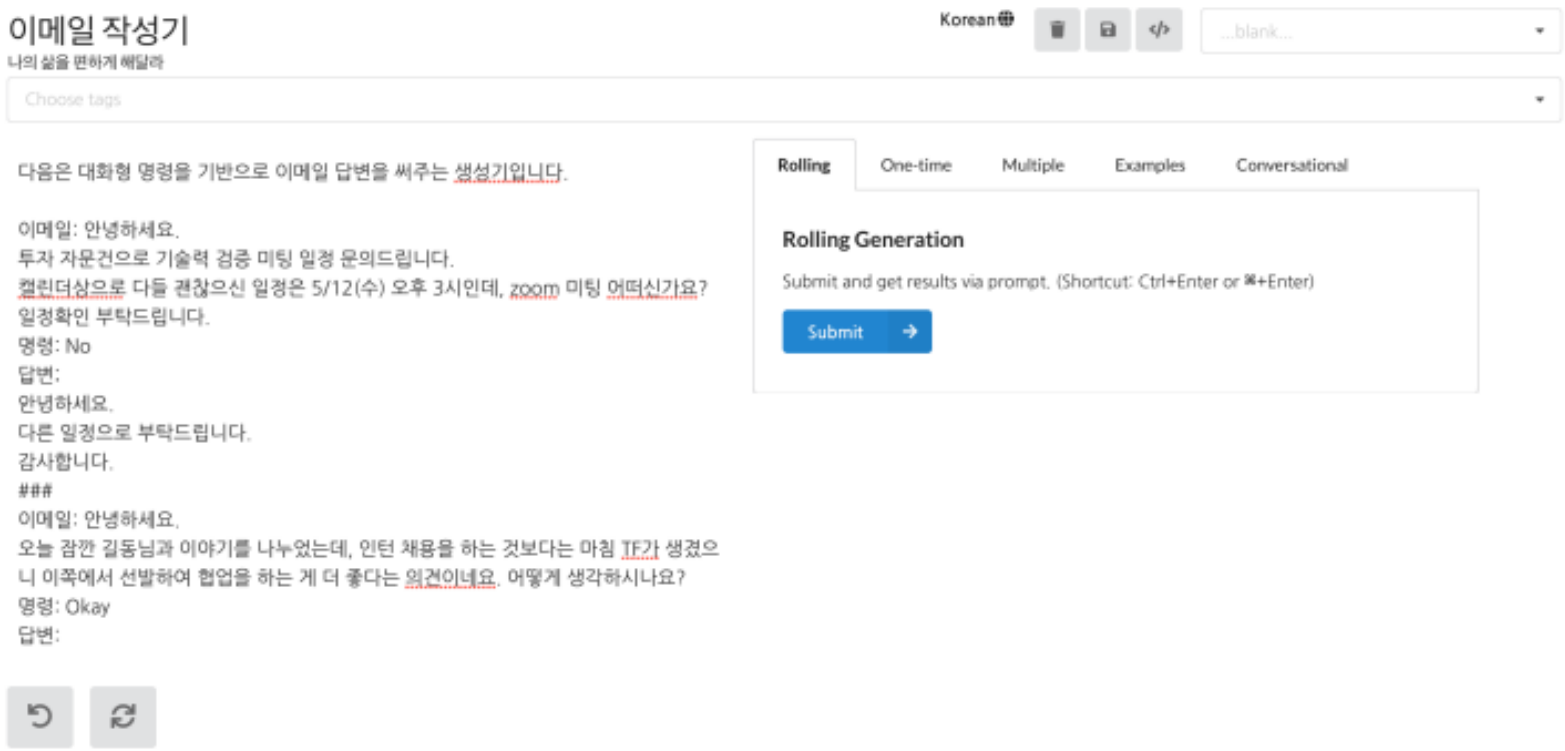
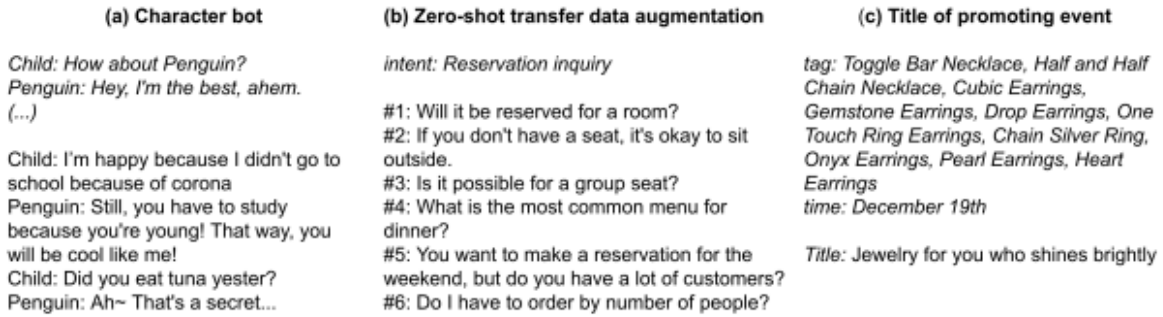


Figure 3: An example interface of HyperCLOVA Studio.

5. Discussion on Industrial Impacts

2. Case Studies on HyperCLOAV Studio



(a) Character bot	(b) Zero-shot transfer data augmentation	(c) Title of promoting event
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...	*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?	*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

Figure 1: Examples generated by HyperCLOVA with the prompts under three different tasks. Italic implies given prompts and non-italic corresponds to generated outputs. The examples are translated into English.

- **Sharing Properties** :

1) it is non-trivial to define the objective function or to evaluate the models automatically.
2) the style of the inputs and outputs is easily controlled.
3) a product designer, without programming skill nor knowledge of AI, can easily make PoC systems within few hours.

5. Discussion on Industrial Impacts

2-1. Rapidly Prototyping Chatbots with Personalities

- can build a **chatbot with the persona of a specific character** using one or two lines of description on the character property and few dialog examples.
- can be used for producing many bots in metaverse applications.

(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...

5. Discussion on Industrial Impacts

2-2. Data Augmentation

- build utterances tailored to user intent.

(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?

2-3. Event Title Generation

- for enhancing product advertisement.

(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

5. Discussion on Industrial Impacts

	BLEU	Win	Lose	Tie
mT5 vs. GT	13.28	0.311	**0.433**	0.256
HyperCLOVA vs. mT5	-	**0.456**	0.350	0.194
GT vs. HyperCLOVA	5.66	0.311	0.333	**0.356**

Table 8: Results of event title generation. GT denotes the ground truth title written by human experts. *Win* means *X* wins against *Y* under *X* vs. *Y*. BLEU is the BLEU score of each model with its corresponding GT.

상품명: 디퓨저꽃 디퓨져스틱 방향제 리드스틱 머스타드 7종
날짜: 2021년 3월 29일
카테고리: 기타아로마/캔들용품
브랜드: 캔들날다 메이릴리
태그: 72993^방향제만들기|64225^디퓨저diy|189638^디퓨저리드|139746^디퓨저만들기|198335^디퓨저만들기재료|379365^인테리어디퓨저
속성: |
광고문구: 봄을 부르는 향기
######
상품명: LYNN 린 차이나 프릴 블라우스
날짜: 2021년 3월 29일
카테고리: 블라우스/셔츠
브랜드: 린
태그: |
속성: 핏^기본핏|패턴^무지|디테일^프릴/러플|총기장^기본/하프|주요소재^폴리에스테르|소매기장^반팔
광고문구: 여성스러운 프릴 블라우스
######
상품명: 맥 아이 섀도우 1.5g
날짜: 2021년 3월 29일
카테고리: 아이섀도
브랜드: 맥
태그: 75984^선물로좋은|76503^포인트주기좋은|281615^자연스러운발색|240838^지속력좋은|235326^포인트연출|665375^파우더리|1228492^부드러운사용감|8360416^자연스러운스모키|5279^청순메이크업|78091^선물포장
속성: 형태^압축/팩트형|세부제품특징^고운입자|세부제품특징^은은함|세부제품특징^웜톤용|색상^골드|주요제품특징^고발색|색상^핑크|세부제품특징^눈매연출|세부제품특징^펄있음|주요제품특징^부드러운발림|색상^브라운|타입^싱글|주요제품특징^지속력
광고문구: 매트한 질감과 선명한 발색
######
상품명: 케이스 아쿠아텍스 이지클린 패브릭 원단 저상형 패밀리 침대 SS.Q
날짜: 2021년 05월 17일
카테고리: 패밀리침대
브랜드: ss퍼니처
태그: 사이즈^슈퍼싱글+퀸|부가기능^안전가드포함|프레임^저상형|자재등급^E0(친환경)|부가기능^유해물질차단|프레임소재^패브릭
속성: 5554855641
광고문구: *안전한 소재로 제작된 저상형 패밀리 침대*

Table 15: Prompt for advertisement headline design

Models	Product event titles
mT5	봄맞이 인테리어 발매트 모음전 (Interior foot-mat event for spring season.)
HyperCLOVA	욕실 분위기를 바꿔줄 아이템 (Items that can change bathroom mood.)
mT5	타이니러브 바운서 (Tiny love bouncer.)
HyperCLOVA	엄마와 아기를 위한 편안함 (Comfort for mommy and baby.)
mT5	한끼 요리 탕요리 반조리 (A meal, stew, semi-cooked.)
HyperCLOVA	저녁 걱정 뚝! 간편한 탕요리 (No worry on dinner! Simple semi-cooked stew.)
mT5	가을맞이 면접룩 기획전 (Interview fashion event for fall season.)
HyperCLOVA	면접 때 입을 옷 고민하지 마세요 (No worry on your fashion for the interview.)

Table 17: Examples of product event titles generated by mT5 and HyperCLOVA. English phrases in parenthesis are translated by human experts for preserving their nuances and semantics.

5. Discussion on Industrial Impacts

3. Opportunity of HyperCLOVA Studio

- **AI functions to improve functions :**
  1) input gradient API
     - can be applied to enhance the performance of local downstream tasks.

  2) prompt injection module

  3) filters for input and output
     - prevent misuse of HyperCLOVA.

5. Discussion on Industrial Impacts

4. No/ Low Code AI Paradigm

- **typical machine learning development pipeline**
  1) problem definition and user research
  2) data gathering and annotation
  3) training and validating models
  4) operating machine learning systems
  5) error analysis and user monitoring

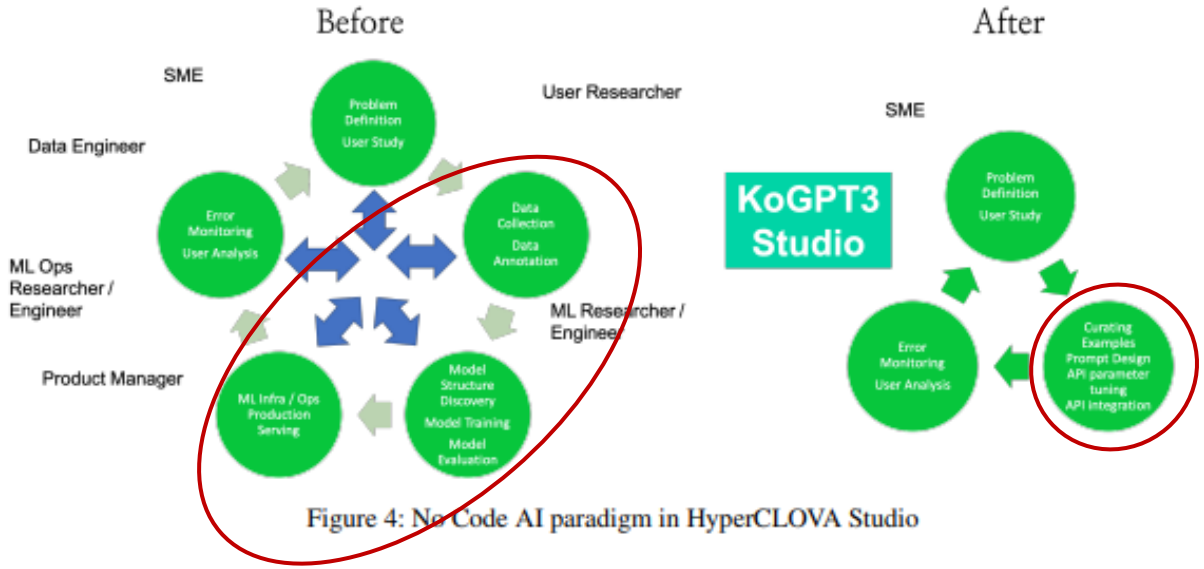


Figure 4: No Code AI paradigm in HyperCLOVA Studio

6. Conclusion

- They present HyperCLOVA, various billions-scale Korean-centric LMs.
- HyperCLOVA with 82B parameters shows state-of-the-art in-context zero-shot and few-shot performance and can further be boosted by prompt-based learning method.
- They will share our model by HyperCLOVA Studio where non-developers can easily build their own AI-backed products.
- They argue that a framework like HyperCLOVA Studio can potentially achieve No Code AI paradigm.
- Goal is to create an ecosystem using HyperCLOVA studio in Korea and help people not familiar with machine learning make their own AI models.

Thank You

감사합니다.