NAVER

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

Boseop Kim*, HyoungSeok Kim*, Sang-Woo Lee*, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, Nako Sung

NAVER CLOVA, AI Lab, Search

1.1 Motivation

GPT-3 makes a new era of language model

- OpenAI playground and API
- AI21 Lab's Jurassic API
- Prompt-based learning
- Codex, DALL-E, ···

1.1 Motivation

GPT-3 makes a new era of language model, though…

- Discovery on non-English large-scale in-context learner is limited.

- Discovery on mid-size GPT-3 (i.e., between 13B ~ 175B) is missing.

- Applying prompt optimization method on in-context learner is not much explored.

- Further discussion is valuable on what large-scale in-context learners can do.

1.2 Our Findings

- Discovery on non-English large-scale in-context learner is limited.
  - We introduce Korean GPT-3, HyperCLOVA, which uses 560B Korean-centric corpus.
  - We discover Korean-specific tokenization scheme.
- Discovery on mid-size GPT-3 (i.e., between 13B ~ 175B) is missing.
  - We also compare 39B and 82B HyperCLOVA.
  - We find that mid-size LM like 39B is quite competitive in our experiments.
- Applying prompt optimization method on in-context learner is not much explored.
  - We first apply a prompt optimization method to a large-scale in-context learner LM.
  - p-tuning is competitive where a few thousand instances exist.
  - p-tuning can sometimes be boosted by few-shot examples in the discrete prompt.
- Further discussion is valuable on what large-scale in-context learners can do.
  - We discuss the possibility of No/Low Code AI using Language Model GUI and CUI.
  - We use the case of our in-house applications to support the argument.

2.1 HyperCLOVA: Model Structure

Transformer decoder with 82B parameter size

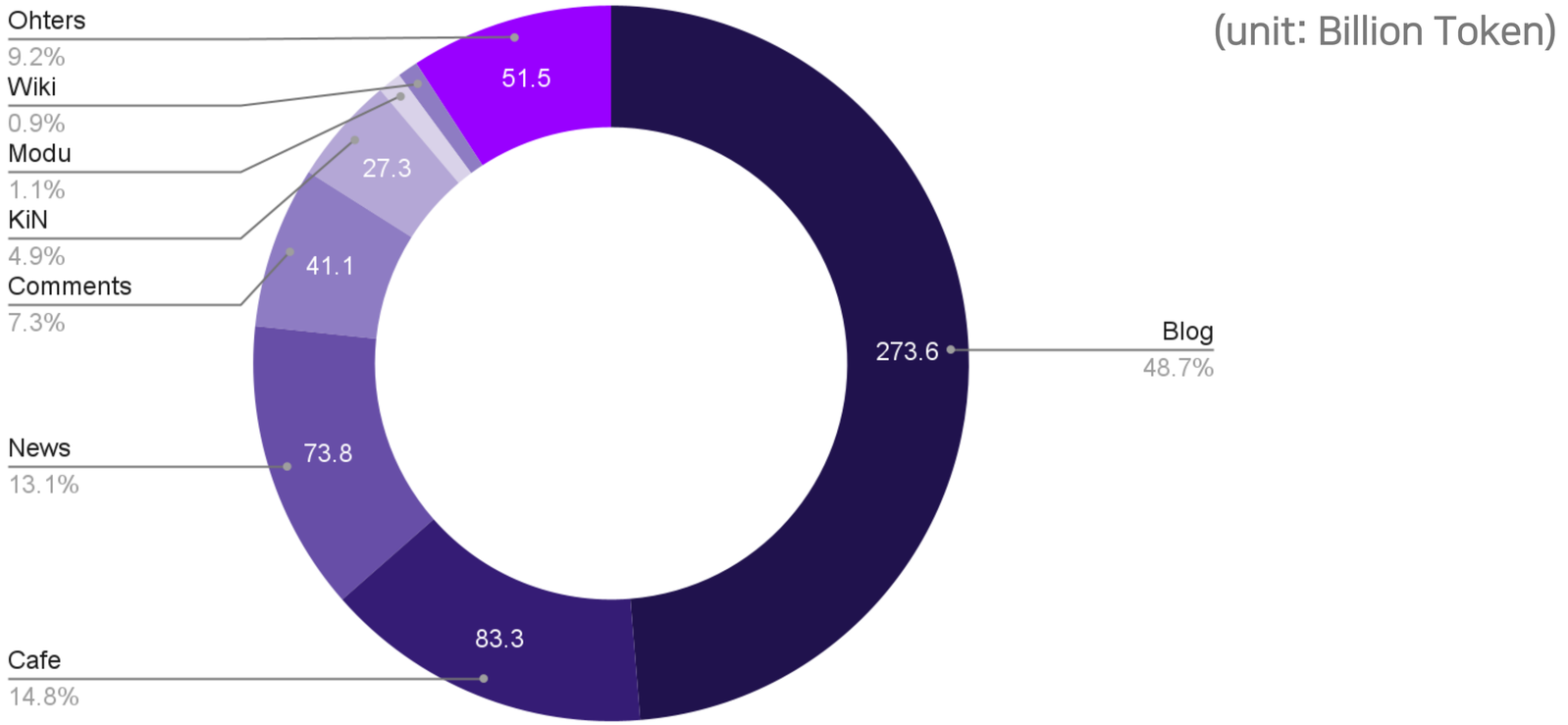
# Param	n_{layers}	d_{model}	n_{heads}	d_{head}	lr
137M	12	768	16	48	6.0e-4
350M	24	1024	16	64	3.0e-4
760M	24	1536	16	96	2.5e-4
1.3B	24	2048	16	128	2.0e-4
6.9B	32	4096	32	128	1.2e-4
13B	40	5120	40	128	1.0e-4
39B	48	8192	64	128	0.8e-4
82B	64	10240	80	128	0.6e-4

It takes 13.4 days 150B token training of 82B model with around 1,000 A100 GPUs

2.2 HyperCLOVA: Korean Corpus

562B tokens of Korean pre-training corpus

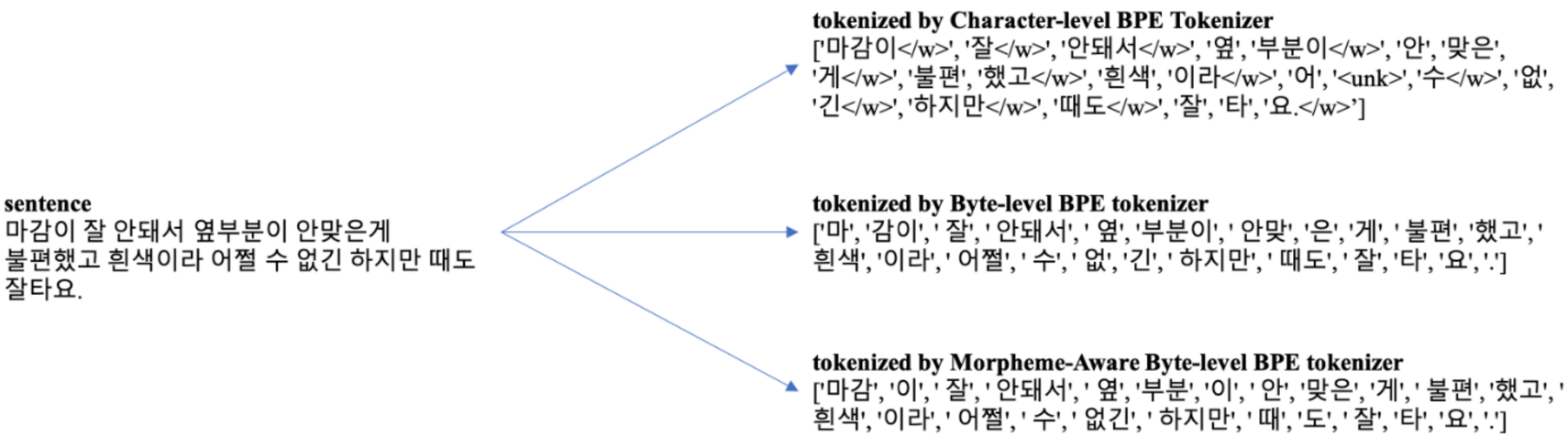
- We gathered all text data with no violation of legal issues, from both diverse services of NAVER and external sources.



2.3 Tokenization

Tokenization: morph-aware byte-level BPE

- Properly tokenizing sentences of Korean, an agglutinative language, is important.
- Our tokenization: byte-level BPE + morpheme analyzer



2.4 Scaling Law

HyperCLOVA also follows the scaling law

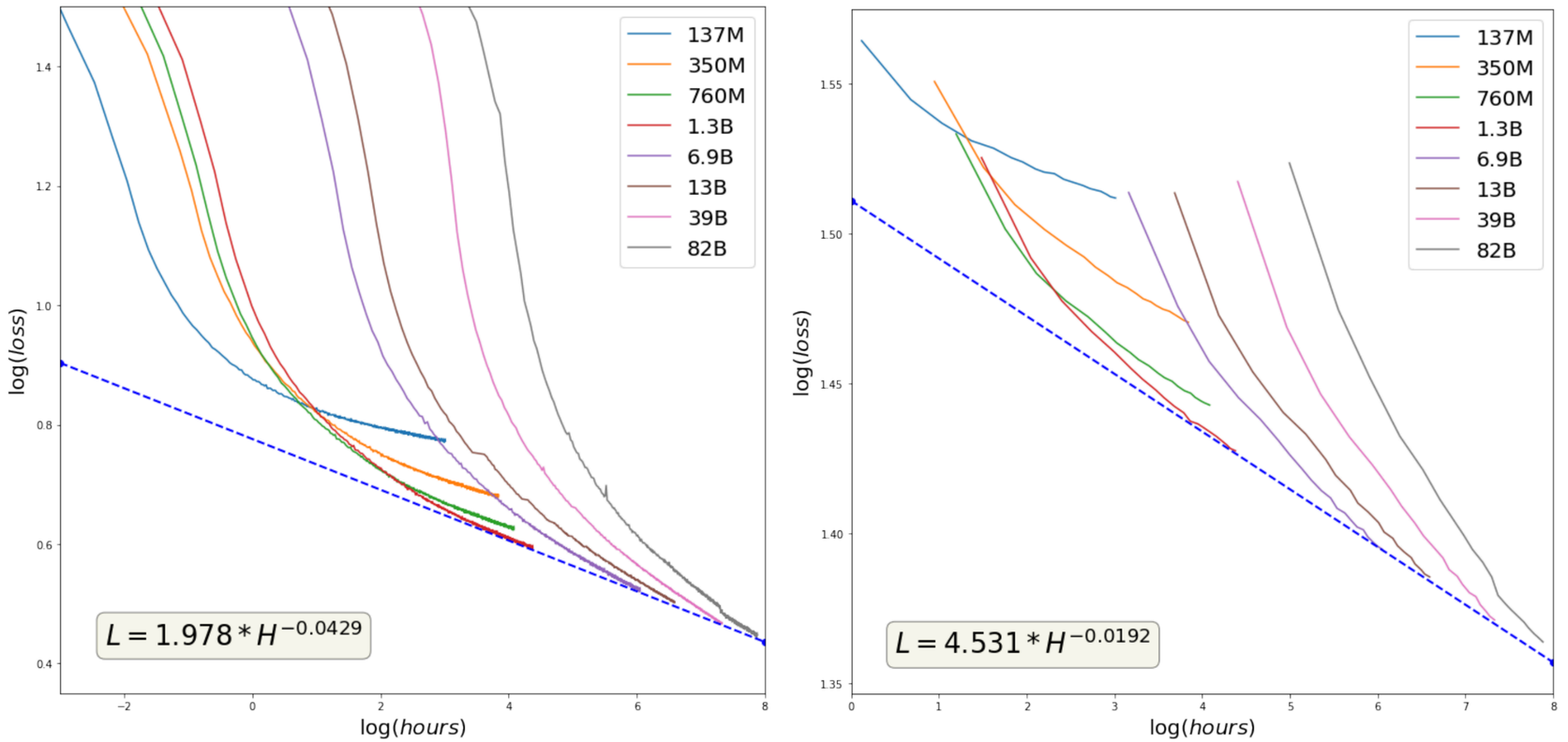


Figure 2: Scaling law (Kaplan et al., 2020; Brown et al., 2020) in training HyperCLOVA models with various parameters. The left figure presents the training and the right graph shows the loss on the testset of a Korean encyclopedia not contained in the training corpus.

2.5 In-context Learning of Korean

In-context learning: zero/few-shot learning w/o fine-tuning

(discrete) prompt

아래의 영어를 한국어로 번역합니다.
(Translate the following English sentence to Korean)

task description

영어: I want you to learn Korean so that we can talk to each other in Korean.
한국어: 나는 네가 한국 언어를 배워서 한국어로 대화했으면 좋겠어.

example (one-shot)

영어: I was surprised to hear that you are using Korean exclamation.
한국어: *너가 한국어 감탄문을 사용했다는 것에 놀랐어.*

3.1 Downstream Task Dataset

Downstream tasks for in-context learning evaluation

	Task Description	#shot
NSMC	Sentiment Classification	70
KorQuAD	Machine Reading Comprehension (MRC)	4
AI Hub	Translation (Ko->En, En->Ko)	4
YNAT	Topic Classification of News Headline	70
KLUE-STS	Semantic Textual Similarity (STS)	40

3.2 In-context Learning Performance

SOTA performance in few-shot learning

	Sentimental Classification	MRC		Translation		Topic Classification	STS
	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

3.3 Ablation Study on Tokenization

Tokenization strategy is important for some tasks

	Morpheme Analyzer	byte vs. char	BPE	OOV
Ours (Morpheme-aware byte-level BPE)	O	byte-level	O	X
byte-level BPE	X	byte-level	O	X
char-level BPE	X	char-level	O	O

	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

3.4 P-tuning on HyperCLOVA

Liu et al., GPT Understands, Too, arXiv, 2021.

P-tuning works well with even 2K instance training, in NSMC

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**

3.4 P-tuning on HyperCLOVA

In-house task: query modification task

Example 1:
User: Play IU's track AI Speaker: I am playing the track. User: How old? **Modified query: How old is IU?**
Example 2:
User: Who invented airplane? AI Speaker: Wright brothers did. User: What is the younger's name? **Modified query: What is the younger one's name of Wright brothers?**

(a) Example data of query modification task. 2,300 training data and 1,300 test data exists.

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

Example 1
User: What are the names of some albums of IU?
AI Speaker: IU's signiture albums include Love poem, Palette, and CHAT-SHIRE.
User: Which one is the most exciting album?
–
[P][P][P] User's [P][P] intent: Among Love poem, Palette,and CHAT-SHIRE, which one is the most exciting album?

Example 2
User: When did the PyeongChang Olympics take place?
AI Speaker: It is 2018.
User: Who was the president of the United States at that time?
–
[P][P][P] User's [P][P] intent:

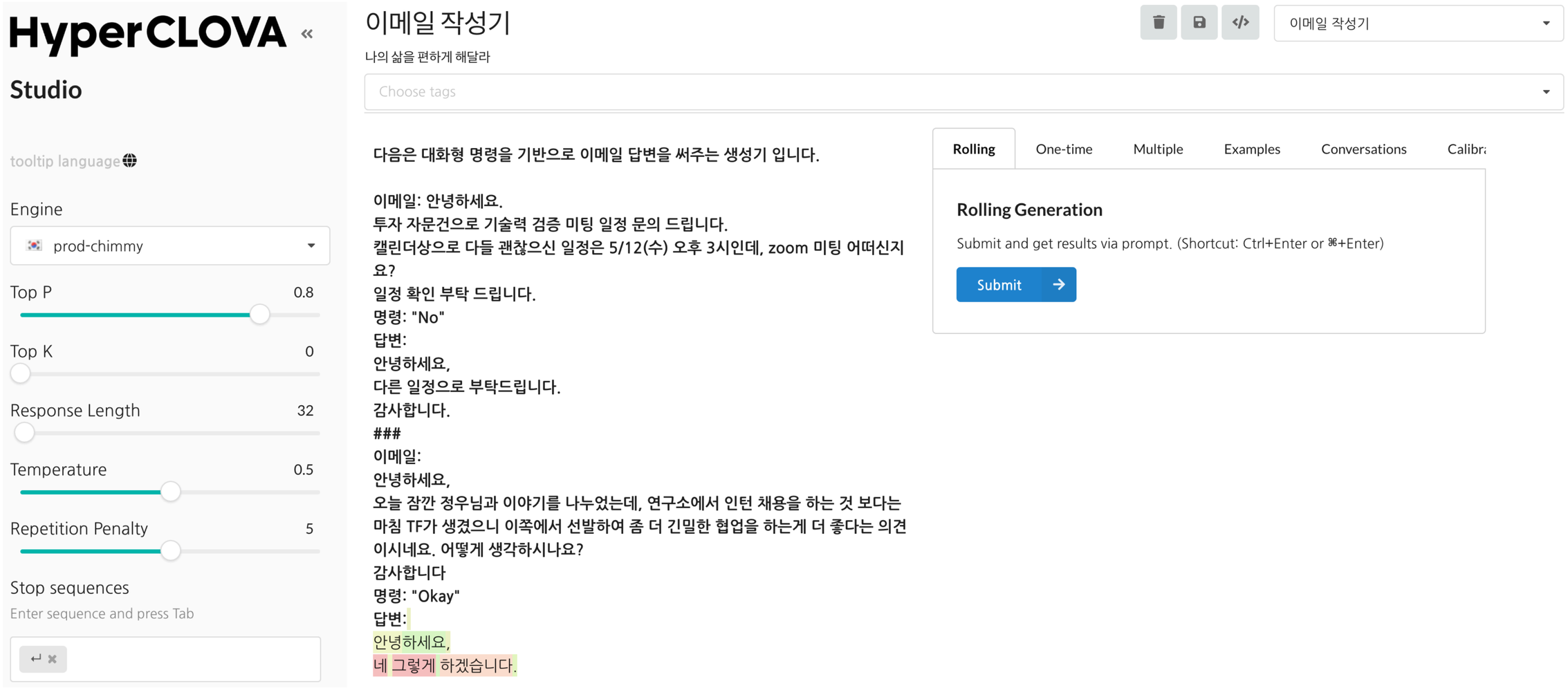
(b) Prompt used for query modification task

3.4 P-tuning on HyperCLOVA

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.

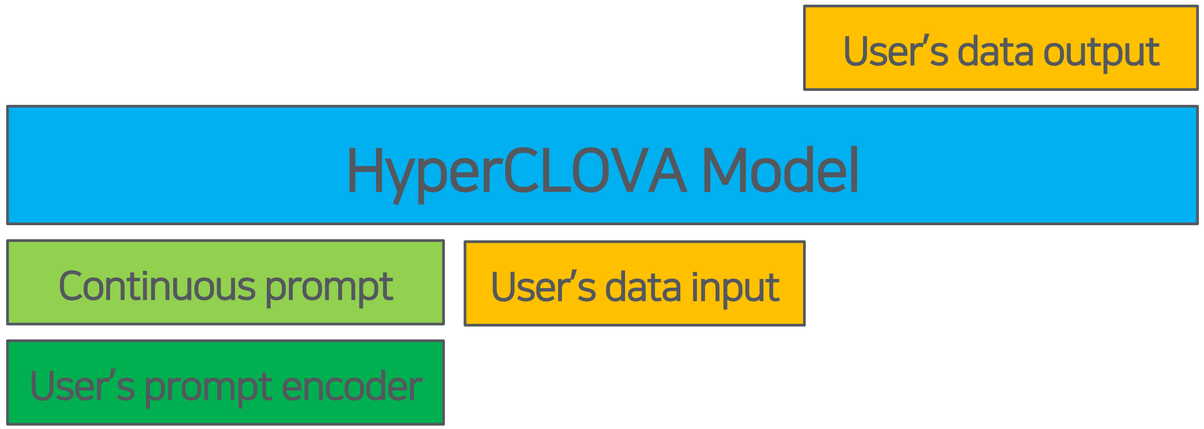
4.1 HyperCLOVA Studio & API



4.2 Input Gradient Scheme

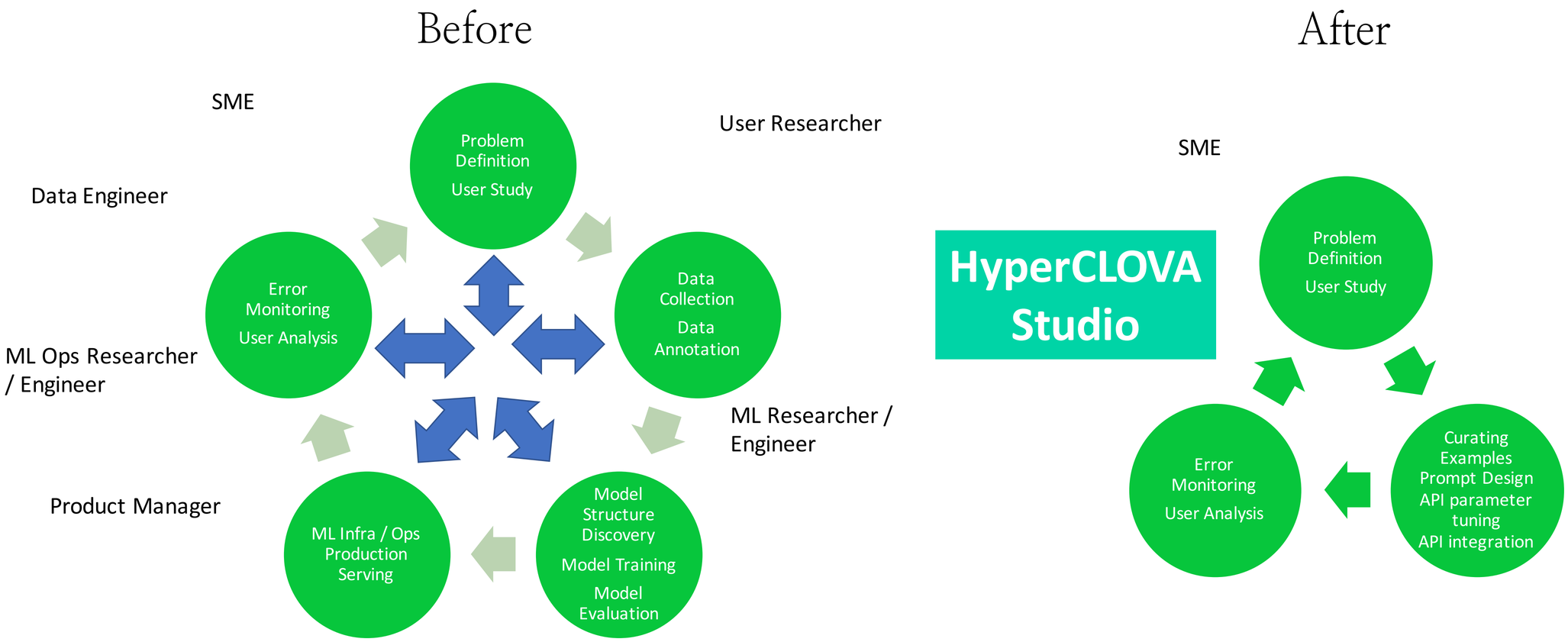
HyperCLOVA API gives input gradient

- Some prompt optimization methods (e.g., p-tuning) can be applied
- User of HyperCLOVA API can train for their own task by updating their small prompt encoder



4.3 Toward No/Low Code AI Paradigm

Decrease communication costs in AI Production



5.1 Application: Query Generation

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?

Zero-shot (Acc)					
	# of augmented samples (k)				
n	5(1)	10(2)	15(3)	25(5)	125(30)
0(0)	$60.8_{9.3}$	$68.9_{4.0}$	$71.9_{2.7}$	$74.8_{2.5}$	$78.0_{2.3}$

Few-shot (Acc)					
	# of original samples (n)				
k	1(1)	2(1)	3(1)	4(1)	5(1)
0(0)	$26.8_{6.0}$	$52.0_{4.9}$	$64.7_{5.2}$	$76.5_{4.4}$	$83.0_{3.0}$
25(5)	$79.2_{2.5}$	$81.2_{2.5}$	$82.6_{2.6}$	$83.4_{1.9}$	$84.3_{2.0}$
125(30)	$80.7_{2.2}$	$82.7_{1.9}$	$83.7_{2.1}$	$86.3_{1.5}$	$87.2_{1.7}$

Table 7: Zero-shot and few-shot performances in zero-shot transfer data augmentation. n denotes the number of original training (validation) instances per class, and k denotes the number of generated instances for training (validation) per class. Subscripted values are standard deviation.

5.2 Application: Event Title Generation

Making a high-quality advertisement copy quickly

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

	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.

6. Conclusion: Our Contribution

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

swlee's Appendix

1.1 Our Motivation

GPT-3 makes a new era of language model, though…

- Discovery on non-English large-scale in-context learner is limited.
- Discovery on mid-size GPT-3 (i.e., between 13B ~ 175B) is missing.
- Applying prompt optimization method on in-context learner is not well discovered.
- Further discussion is valuable on what large-scale in-context learner can do.