Paper Review

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

Kim et al., EMNLP, 2021

Myeongsup Kim

Master Student
Data Science & Business Analytics Lab.
School of Industrial Management Engineering
Korea University

Myeongsup_kim@korea.ac.kr

<What Are Not Covered in This Presentation>

- **Details of Transformer**

  Vaswani et al., Attention is All You Need, NIPS, 2017

- **Details of Generative Pre-trained Transformer Series (GPT, GPT-2, GPT-3)**

  Radford et al., Improving Language Understanding by Generative Pre-Training, 2018

  Radford et al., Language Models Are Unsupervised Multitask Learners, 2019

  Brown et al., Language Models Are Few-Shot Learners, NeurIPS, 2020

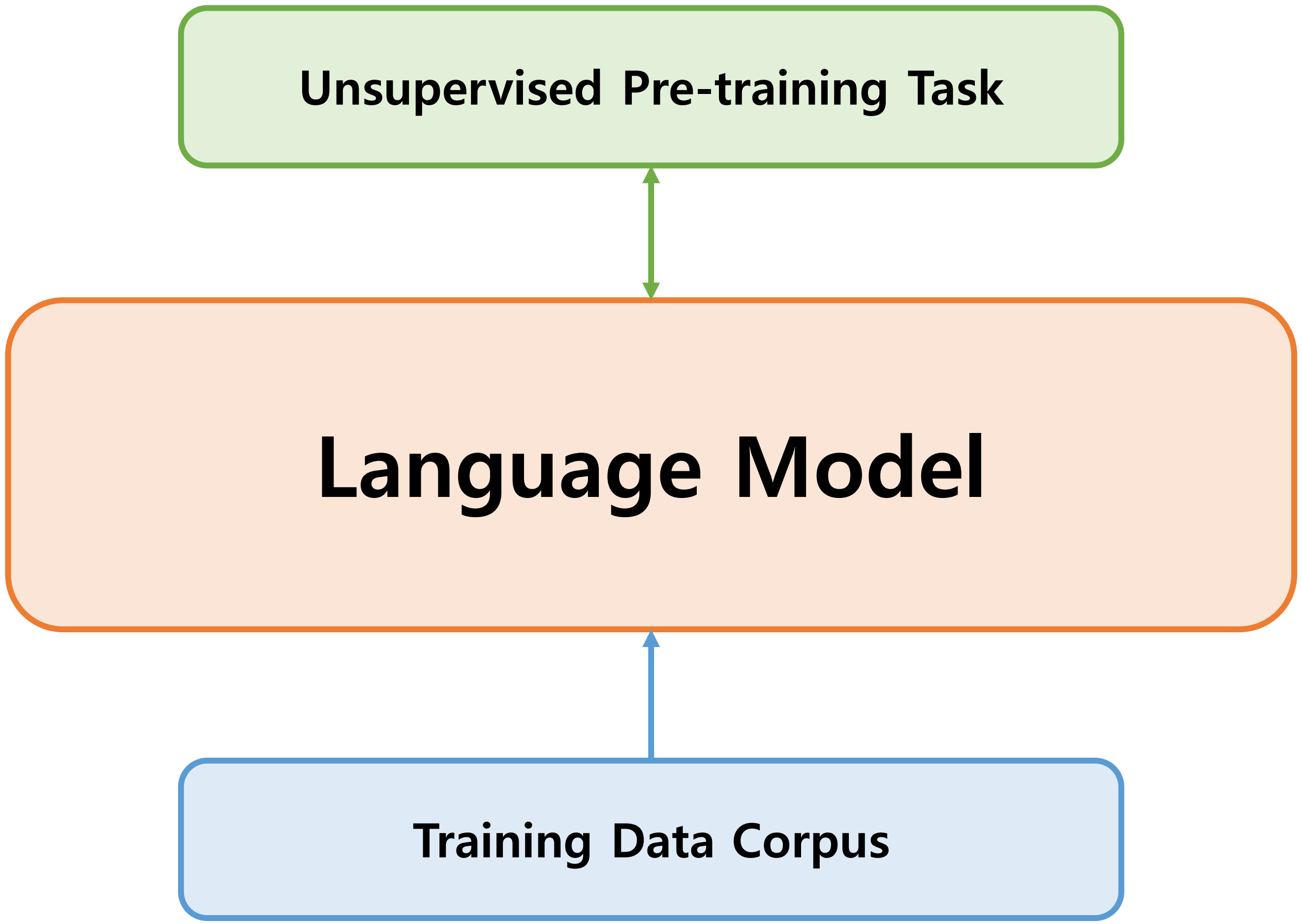
Introduction

- Hyper Scale Language Model
- Few Shot Learning for Language Model
- In-Context Learning

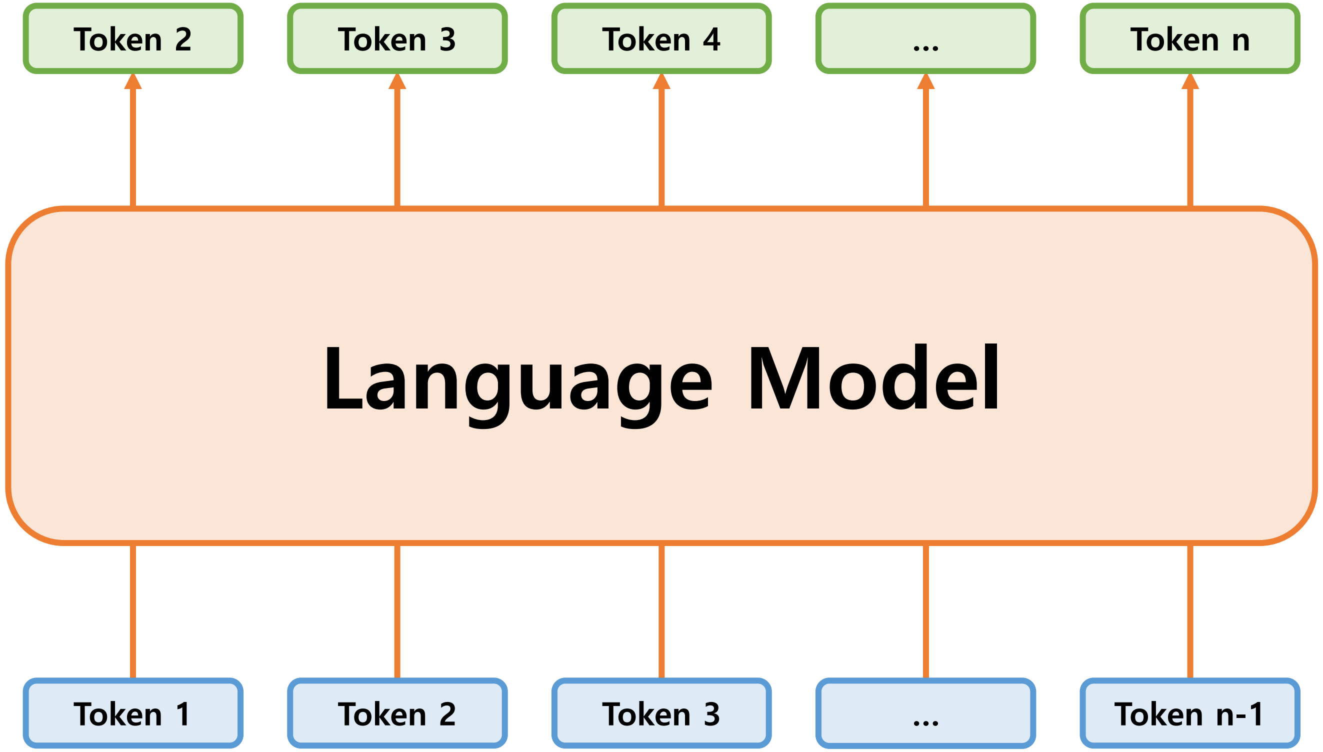
Introduction

-Hyper Scale Language Model

<Language Model>



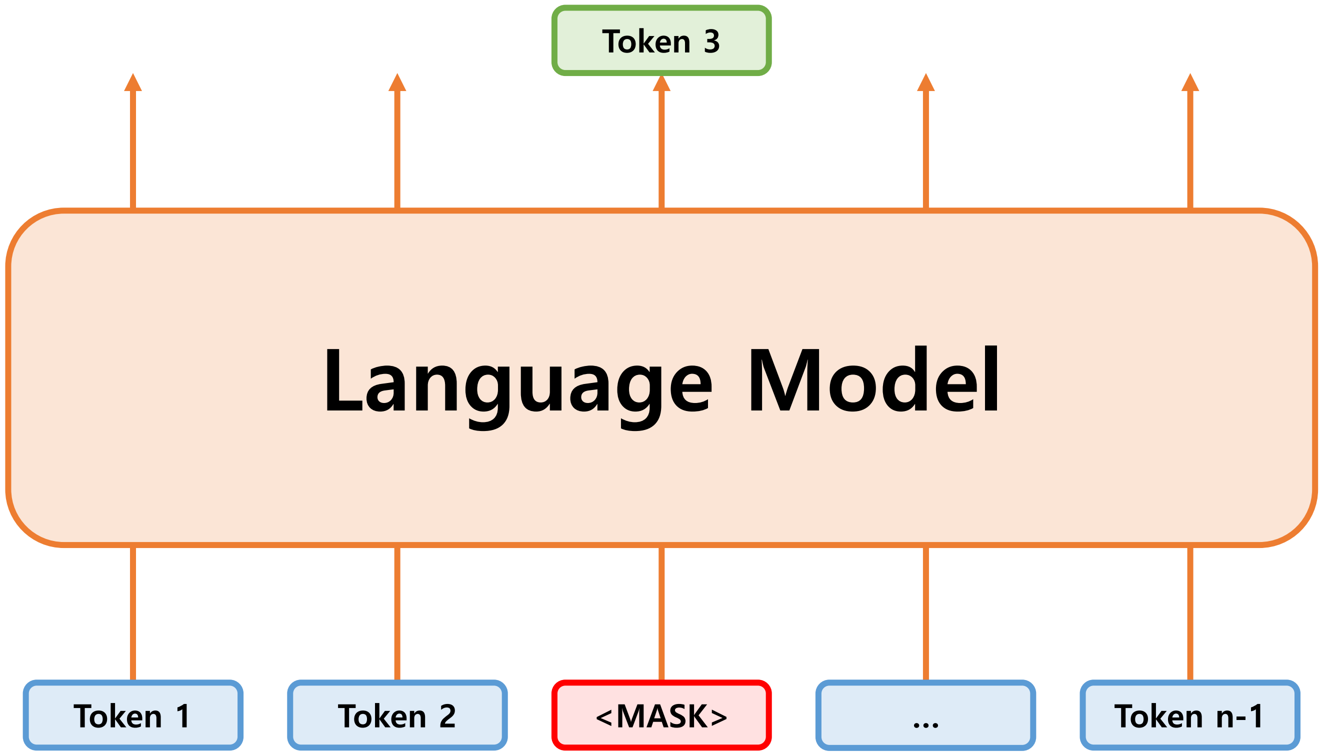
<Language Modeling>



Introduction

-Hyper Scale Language Model

<Masked Language Modeling>



<Language Model>

Document Classification

Sentiment Analysis

…

ELMO
Embeddings from Language Model

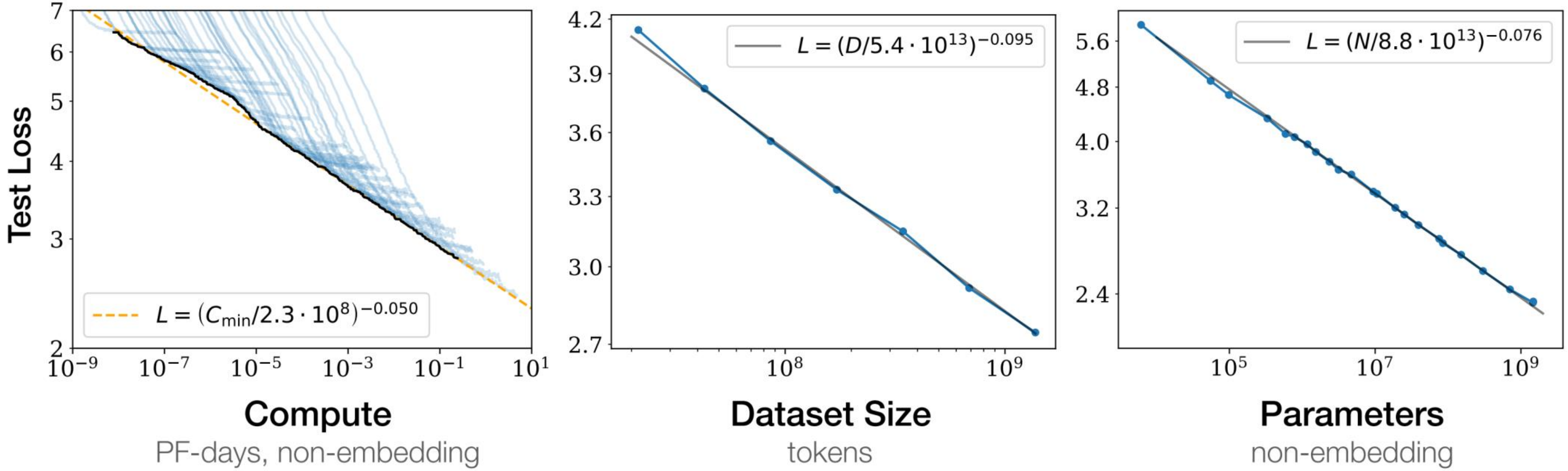
BERT
Bidirectional Encoder Representation from Transformer

GPT
Generative Pre-trained Transformer

Introduction

-Hyper Scale Language Model

<The Scaling Laws for LMs>

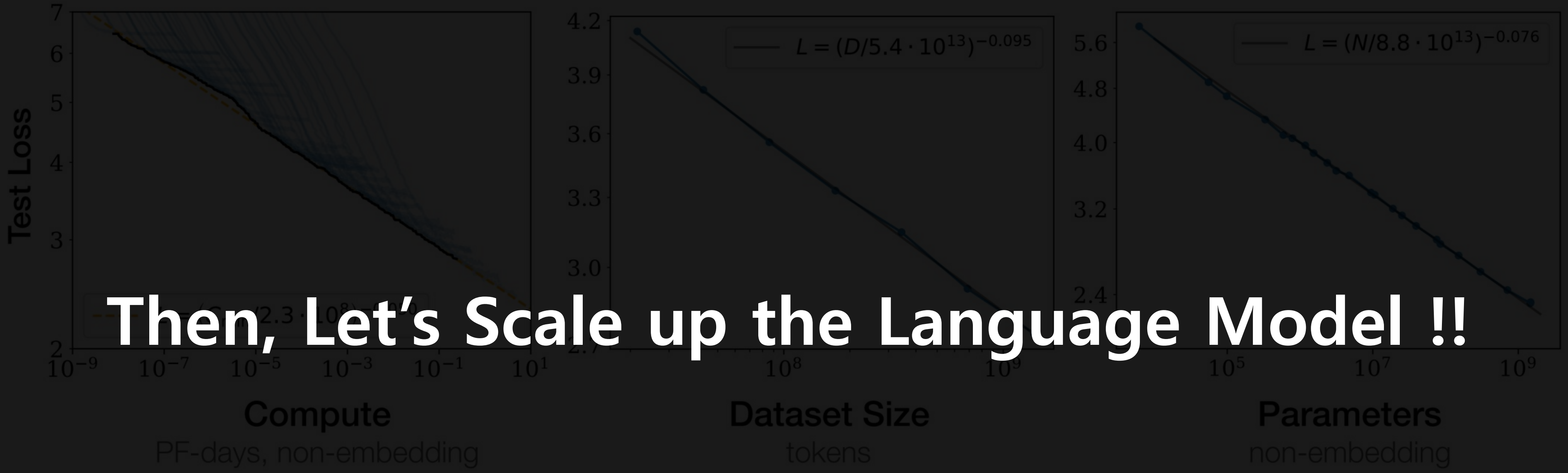


“**Language modeling performance improves** smoothly as **we increase** **<u>the model size</u>**, **<u>dataset size</u>**, and **<u>amount of compute</u> used for training**.” (Kaplan et al., 2020)

Introduction

-Hyper Scale Language Model

<The Scaling Laws for LMs>



Then, Let's Scale up the Language Model !!

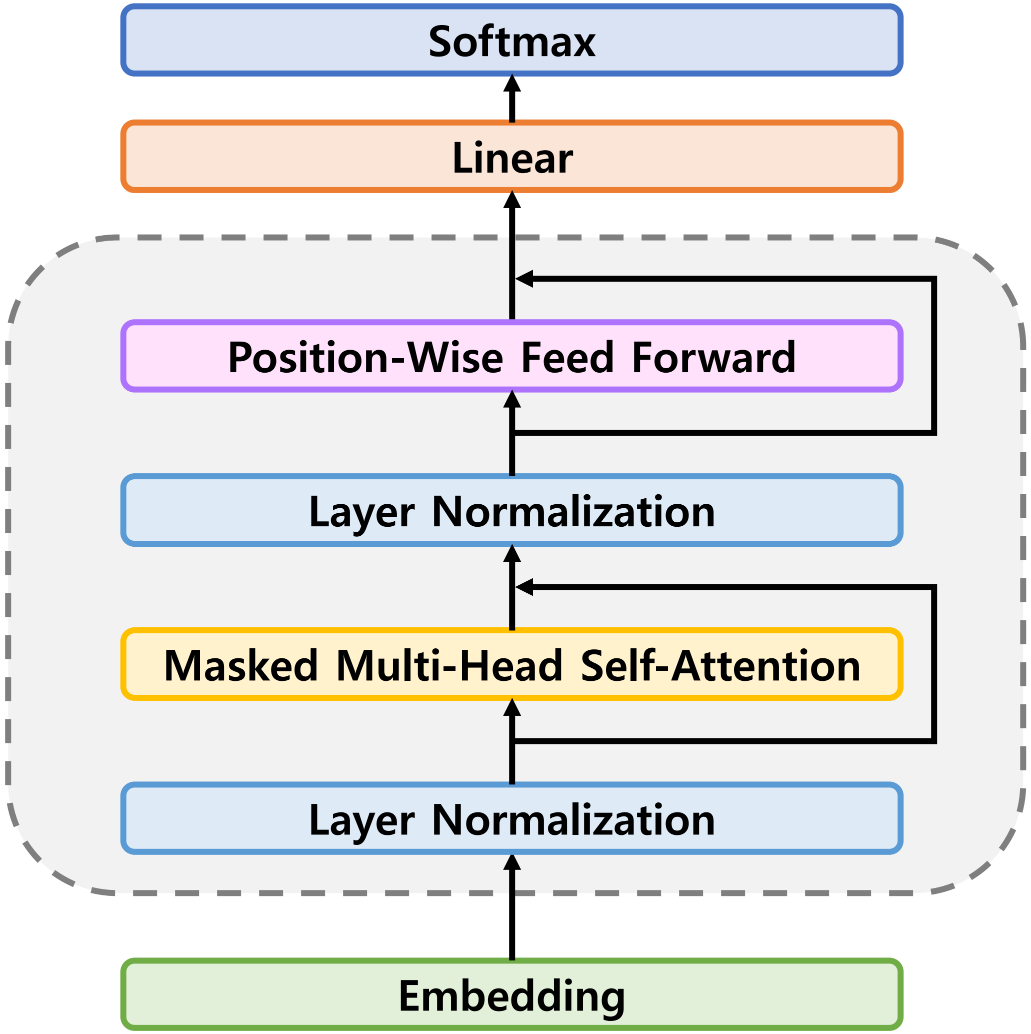
"**Language modeling performance improves** smoothly as **we increase <u>the model size</u>**, **<u>dataset size</u>**, and **<u>amount of compute</u> used for training**." (Kaplan et al., 2020)

Brown et al., Language Models Are Few-Shot Learners, NeurIPS, 2020

Introduction

-Hyper Scale Language Model

<Generative Pre-trained Transformer-3>



$$\boldsymbol{n_{params} = 175.0B}$$

$$n_{layers} = 96$$

$$d_{model} = 12288$$

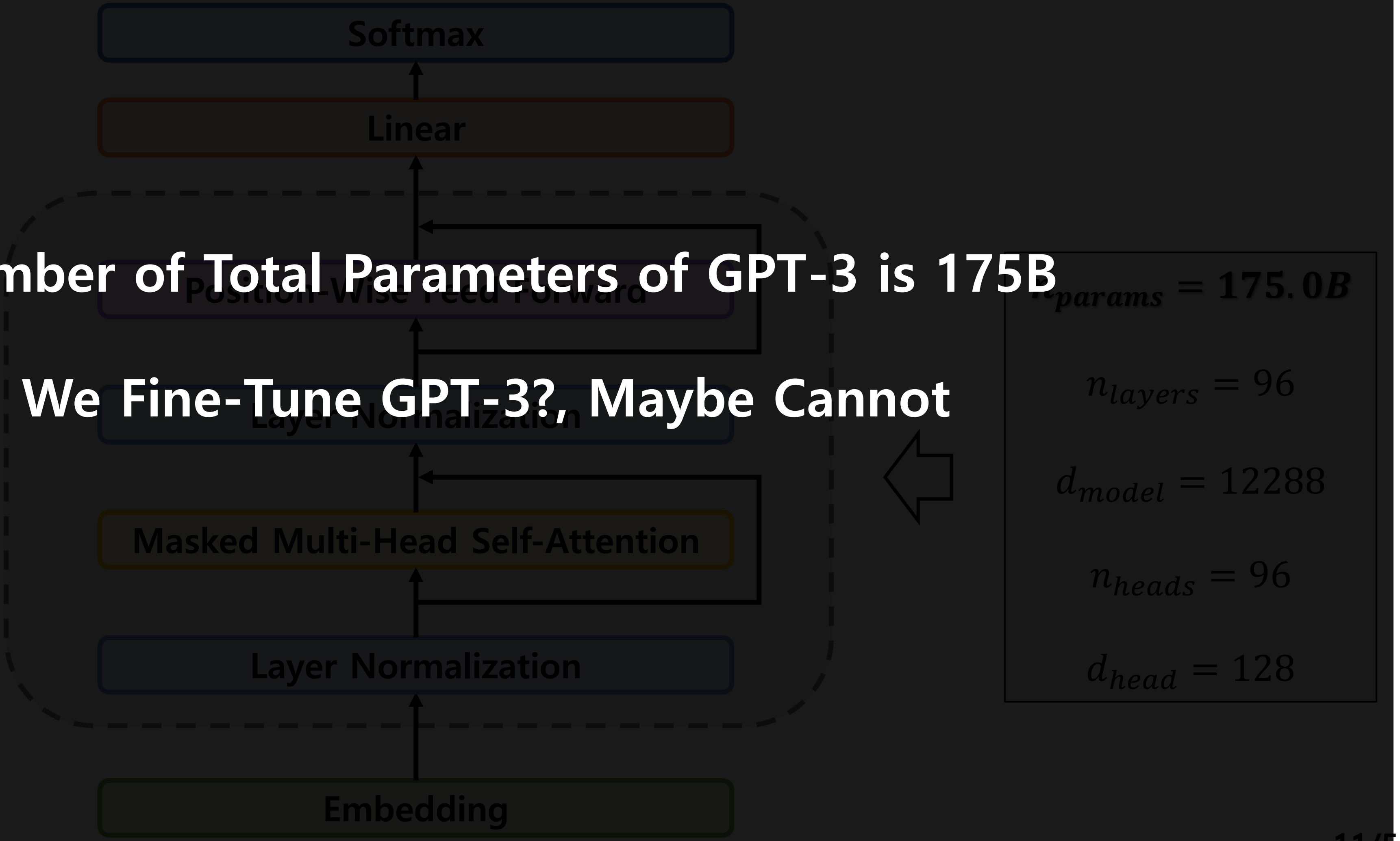
$$n_{heads} = 96$$

$$d_{head} = 128$$

Introduction

-Hyper Scale Language Model

<Generative Pre-trained Transformer-3>



The Number of Total Parameters of GPT-3 is 175B

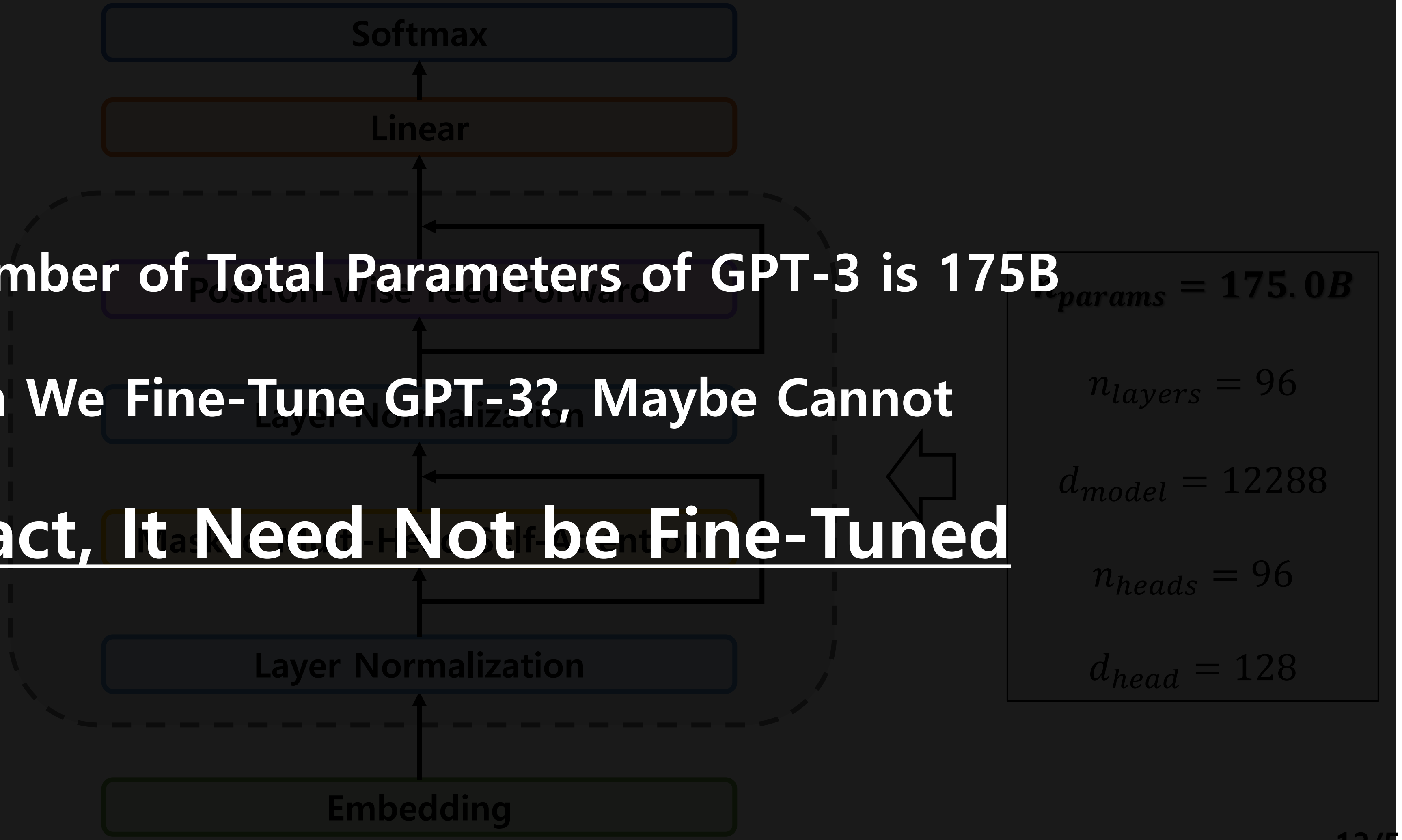
Can We Fine-Tune GPT-3?, Maybe Cannot

[Brown et al., Language Models Are Few-Shot Learners, NeurIPS, 2020]

Introduction

-Hyper Scale Language Model

<Generative Pre-trained Transformer-3>



The Number of Total Parameters of GPT-3 is 175B

Can We Fine-Tune GPT-3?, Maybe Cannot

<u>In Fact, It Need Not be Fine-Tuned</u>

Introduction

-Few Shot Learning for LM

<Few Shot Learning for LM>

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
cheese => ..........              ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
sea otter => loutre de mer        ← example
cheese => ..........              ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
sea otter => loutre de mer        ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese => ..........              ← prompt
```

Introduction

-Few Shot Learning for LM

<Few Shot Learning for LM>

Translate English to Korean:

I am a student. -> 나는 학생이다.

I like pizza. -> 나는 피자를 좋아한다.

How are you? -> ______________

잘 지내고 있니?

Introduction

-Few Shot Learning for LM

<Few Shot Learning for LM>

Answer the question:

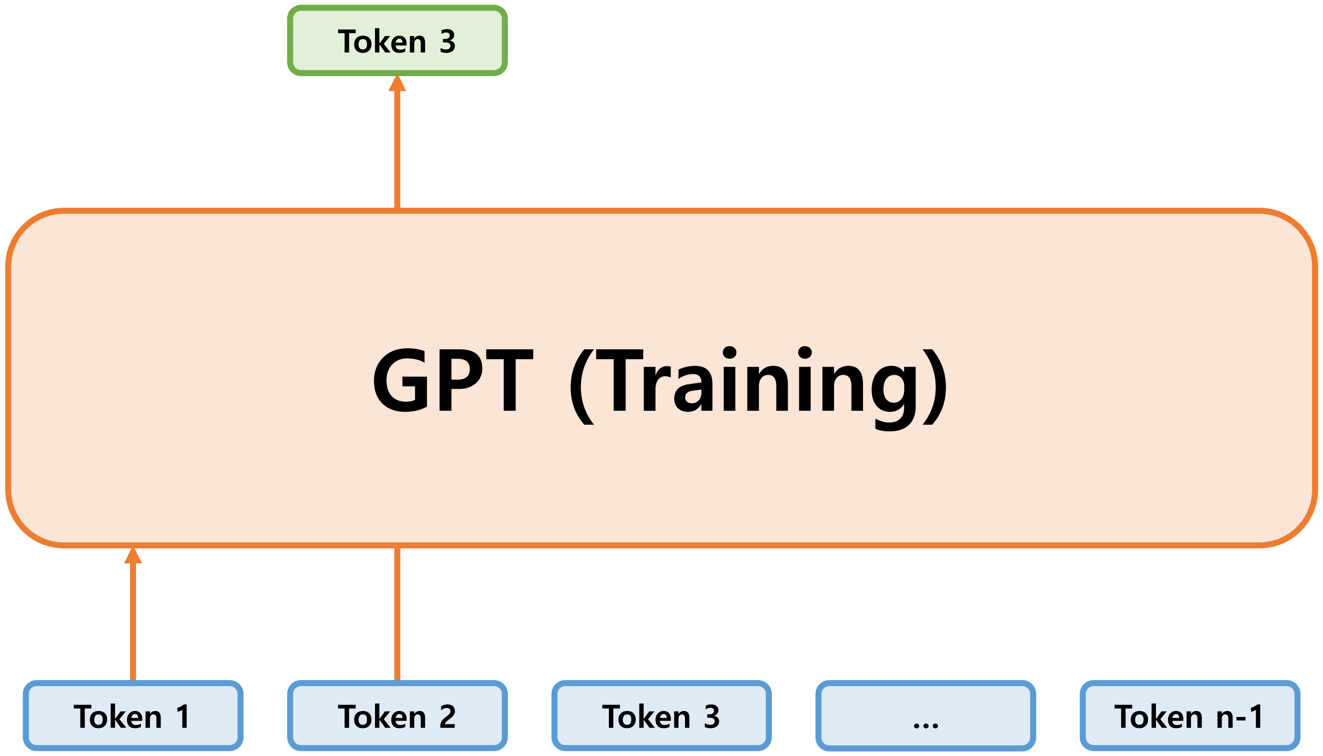
Where is the capital of UK? -> London

Who founded Apple? -> __________

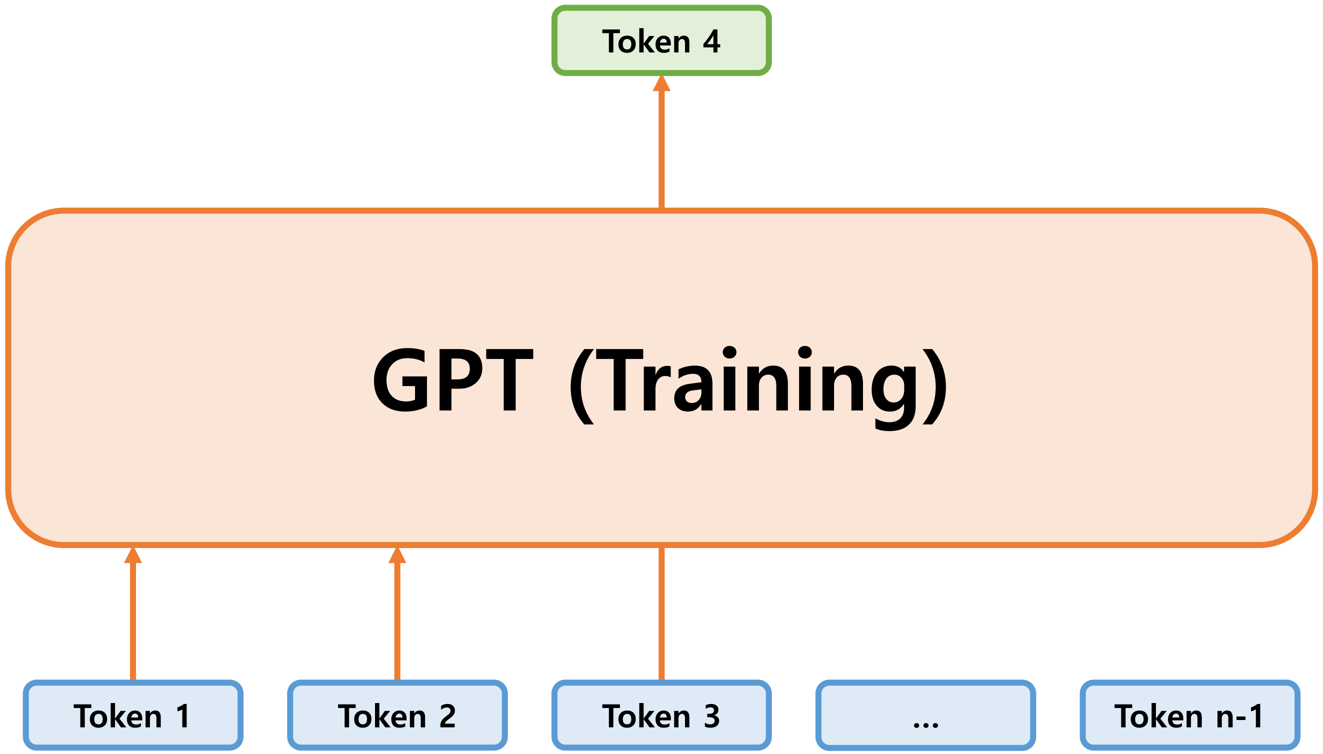
Steve Jobs

<In-Context Learning>

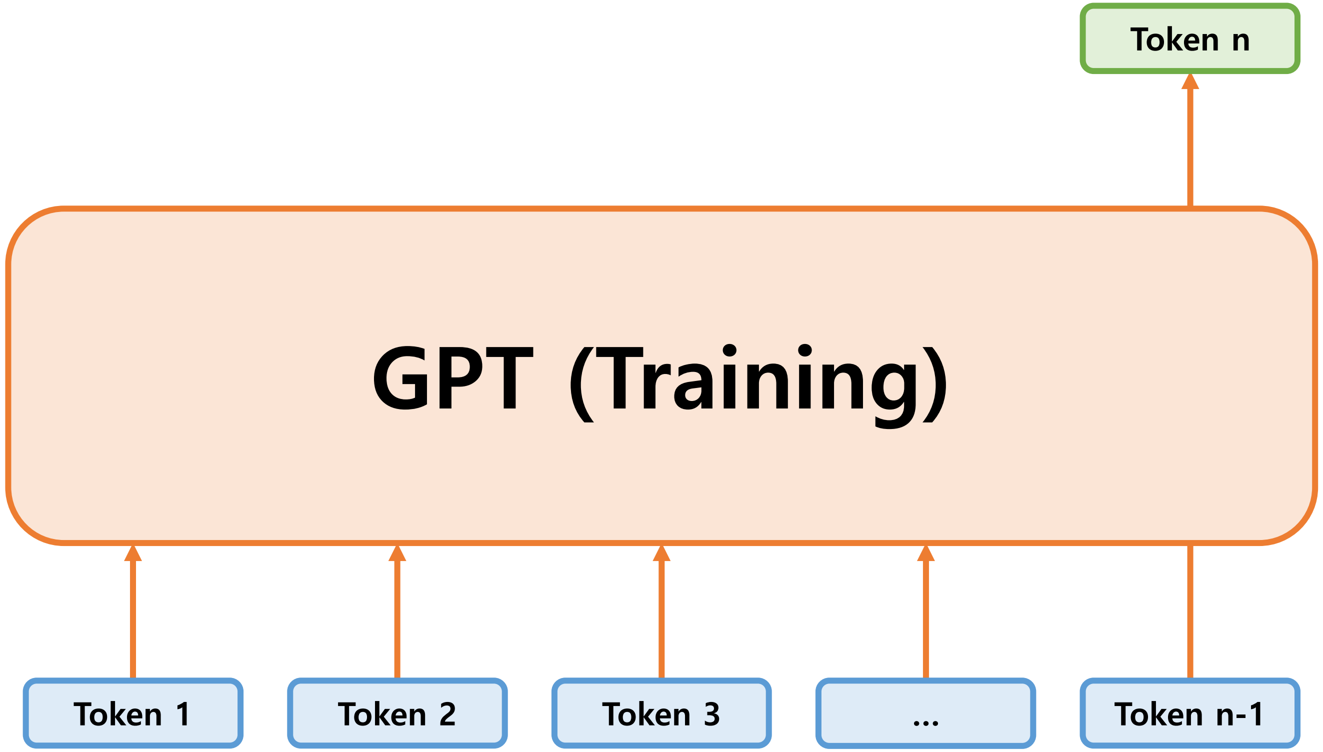
<In-Context Learning>



<In-Context Learning>



<In-Context Learning>



<In-Context Learning>

GPT (Training)

I heard someone saying “I am a student” which means “나는 학생이다” in Korean

<In-Context Learning>

I heard someone saying “I am a student” which means “나는 학생이다” in Korean

GPT (Training)

I heard someone saying “I am a student” which means “나는 학생이다” in Korean

<In-Context Learning>

I heard someone saying “I am a student” which means “나는 학생이다” in Korean

GPT (Training)

I heard someone saying “I am a student” which means “나는 학생이다” in Korean

Introduction

-In-Context Learning

<In-Context Learning>

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

“Examples of **naturally occurring demonstrations of English to French and French to English translation** found throughout the WebText training set.” (Radford et al., 2019)

Introduction

-In-Context Learning

<In-Context Few Shot Learning>

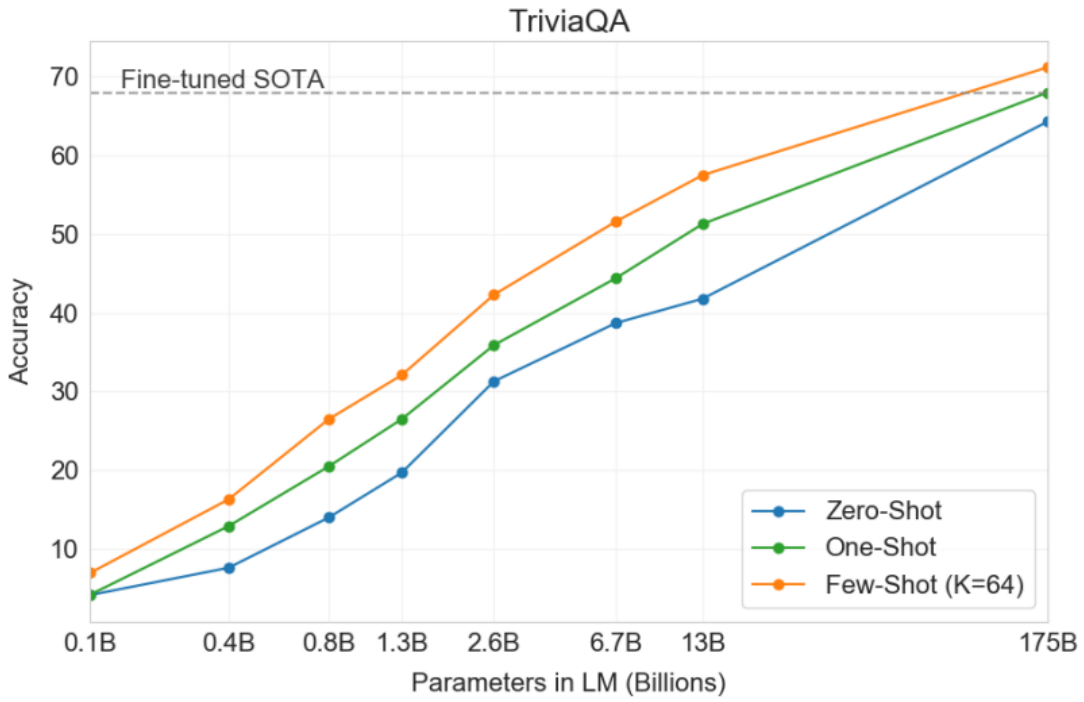


Figure 3.3: On TriviaQA GPT3's performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG [LPP+20]

Brown et al., Language Models Are Few Shot Learners, NeurIPS, 2020

<In-Context Few Shot Learning>

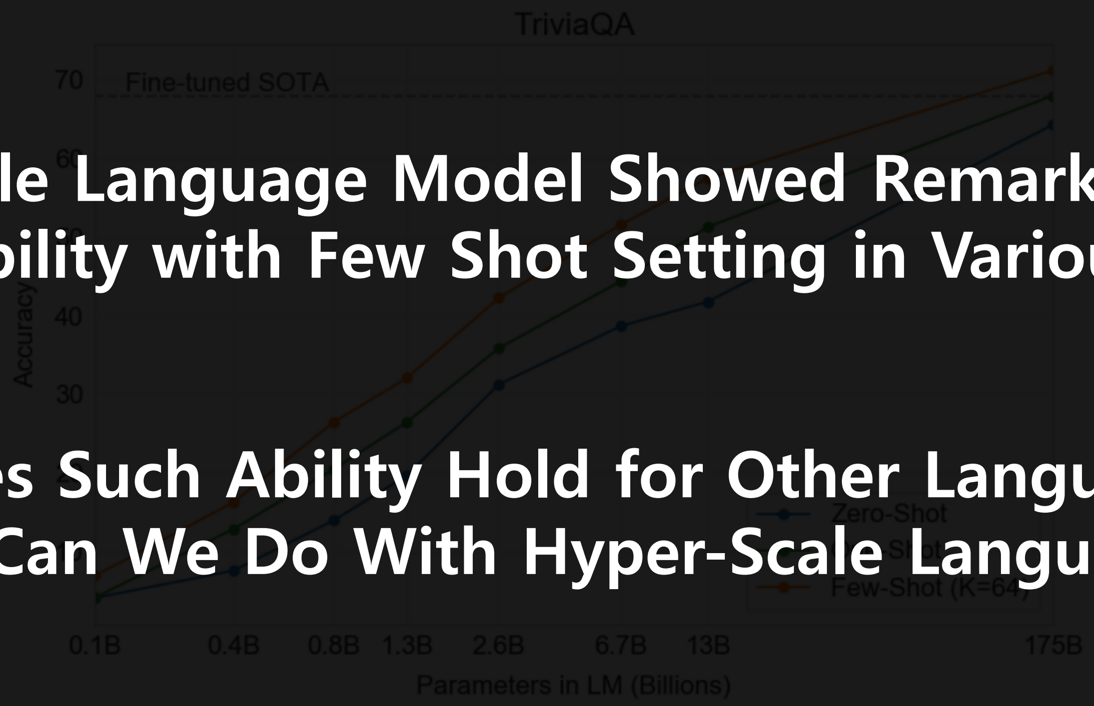


Figure 3.3: On TriviaQA GPT3's performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG [LPP+20]

The Hyper-Scale Language Model Showed Remarkable In-Context Learning Ability with Few Shot Setting in Various NLP Tasks

Does Such Ability Hold for Other Language?
And What Can We Do With Hyper-Scale Language Models?

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

Kim et al., 2021, EMNLP

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

Kim et al., 2021, EMNLP

Pre-Training

- Data Description
- Model and Learning
- Tokenization

<Data Description>

Name	Description	Tokens
Blog	Blog corpus	273.6B
Café	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**

"The ratio of Korean data for OpenAI GPT-3 is very small, with less than 0.02 by character count."

"Therefore, it is crucial to construct a large Korean-centric corpus in advance to training HyperCLOVA."

<Configuration per Size of HyperCLOVA>

# $\boldsymbol{Param}$	$\boldsymbol{n_{layers}}$	$\boldsymbol{d_{model}}$	$\boldsymbol{n_{heads}}$	$\boldsymbol{d_{head}}$	$\boldsymbol{lr}$
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

"We make our model design similar to GPT-3, and we set near exponential interpolation from 13B to 175B OpenAI GPT-3."

"We aim to explore the capability and representation power of the models with mid-size parameters."

<Training>

"Our model is based on **megatron-LM**"

"and Trained on **NVIDIA Superpod**, which includes **128 strongly clustered DGX servers** with **1,024 A100 GPUs**"

"We use **AdamW** with **cosine learning rate scheduling** and **weight decay** as an optimizer"

"**mini-batch size of 1,024**"

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

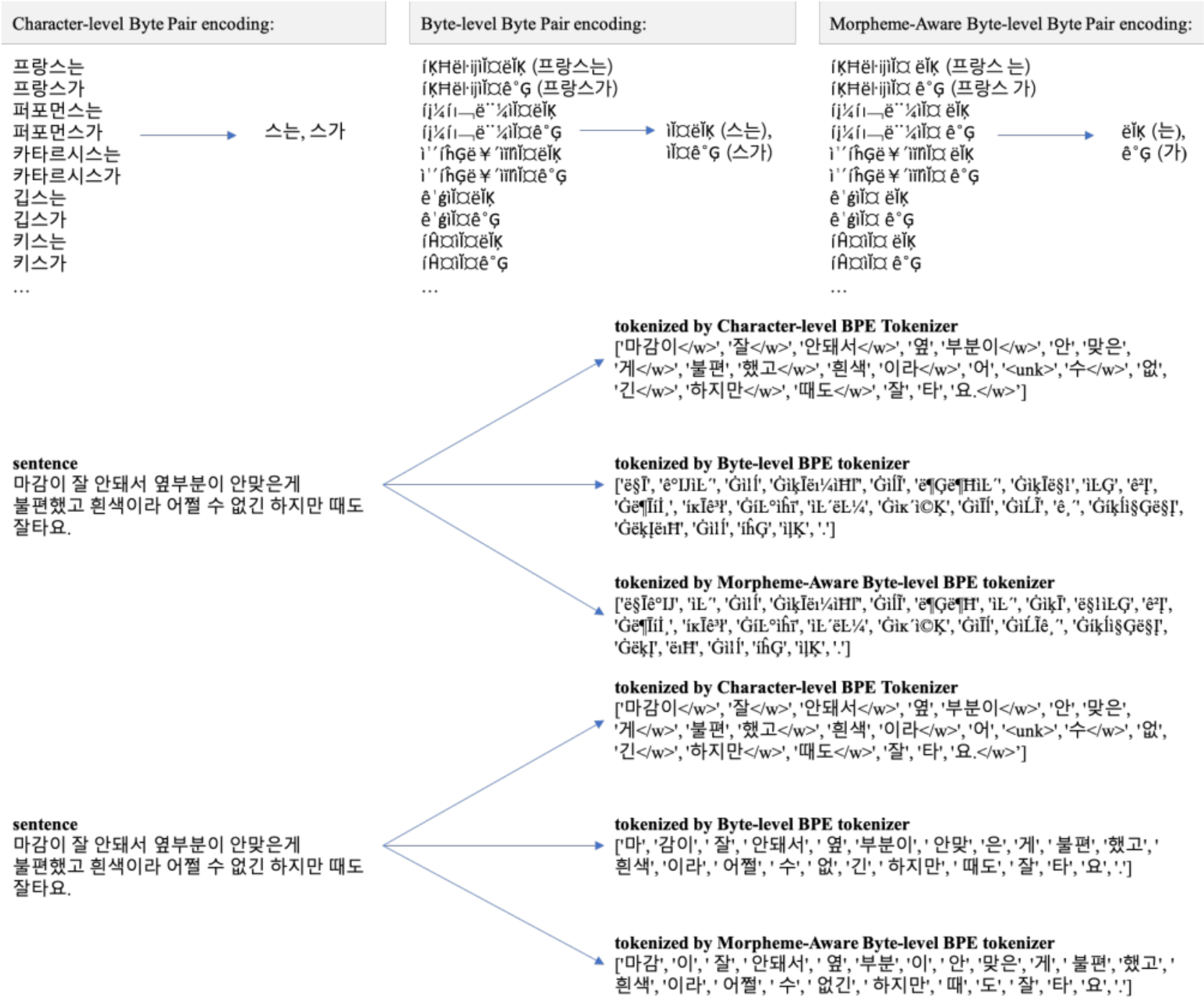
Method
-Model Training

<Training>

Method

-Tokenization

<Tokenization>



Experiments

- Experimental Setting
- In-Context Few-Shot Learning
- Prompt-Based Tuning
- Effect of Tokenization

Experiments

- Experimental Setting

<Dataset>

Dataset	Description	Task	Setting
NSMC	A movie review dataset from NAVER Movies 150K training data and 50K test data	**Sentiment Classification**	12 sets, 70 examples average accuracy
KorQuAD 1.0	Korean version opouf machinie reading comprehension dataset 10,645 training passage with 66,181 training questions, 5,774 valiation questions	**Question Answering**	1 paragraph 4 question set zero-shot learning for paragraph four-shot learning for question
AI Hub Korean-English	Korean-English parallel sentences from news, government websites, legal documents, etc 800K sentence pairs	**Machine Translation**	randomly sample 1K pairs for evaluating three random trials for each task four shot learning
YNAT	Yonhap news agency topic classification, seven classes 45K, 9K, and 9K annotated headlines for training, valid, and test	**Topic Classification**	Average accuracy of 3 in-context 70-shot learners
KLUE-STS	Predict a sentence similarity between each pair of sentences similarity score between 0 and 5	**Semantic Textual Similarity**	Average accuracy of 3 in-context 70-shot learners
Query modification task	Query modification task for AI speaker users converting multi-tern query into single-tern query	**Query Modification**	…

Experiments
- Experimental Setting

<Dataset>

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?)**

<Example of user query modification task>

<In-Context Few-Shot Learning>

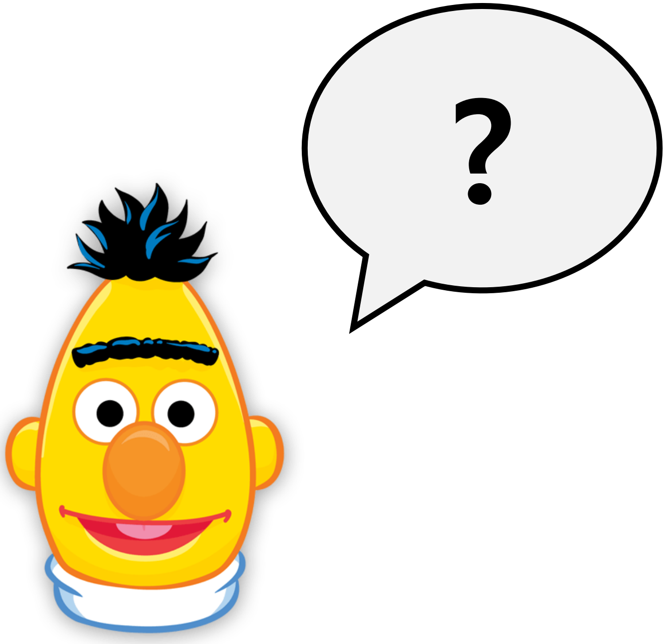
	NSMC (Acc)	KorQuAD (EA / F1)		AI Hub (BLEU) Ko -> En	AI Hub (BLEU) En -> Ko	YNAT (F1)	KLUE-STS (F1)
Baseline	89.66	74.04	86.66	40.34	40.41	82.64	75.93
137M	73.11	8.87	23.92	0.80	2.78	29.01	59.54
350M	77.55	27.66	46.86	1.44	8.89	33.18	59.45
760M	77.64	45.80	63.99	2.63	16.89	47.45	52.16
1.3B	83.90	55.28	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

<Result of in-context few-shot learning>

<Prompt-Based Tuning>

Question Answering

Where is the capital of UK?



<Prompt-Based Tuning>

The capital of UK is <MASK>

<Prompt-Based Tuning>

Question Answering

Where is the capital of UK?	→	The capital of UK is <MASK>

Sentiment Classification

My phone doesn’t work	→	My phone doesn’t work. So, I feel <MASK>

<Prompt-Based Tuning>

Method	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 (Part 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**

<Comparison results of p-tuning with fine-tuning and in-context few-shot learning on NSMC>

<Prompt-Based Tuning>

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

<Result of p-tuning on in-house query modification task>

"p-tuning enables HyperCLOVA to outperform comparatives with no parameter update"

"p-tuning with only 4K examples provides comparable results to RoBERTa fine-tuned on 150K data"

"this is the first report of applying input-side p-tuning to generation tasks with an in-context LM"

<Effect of Tokenization>

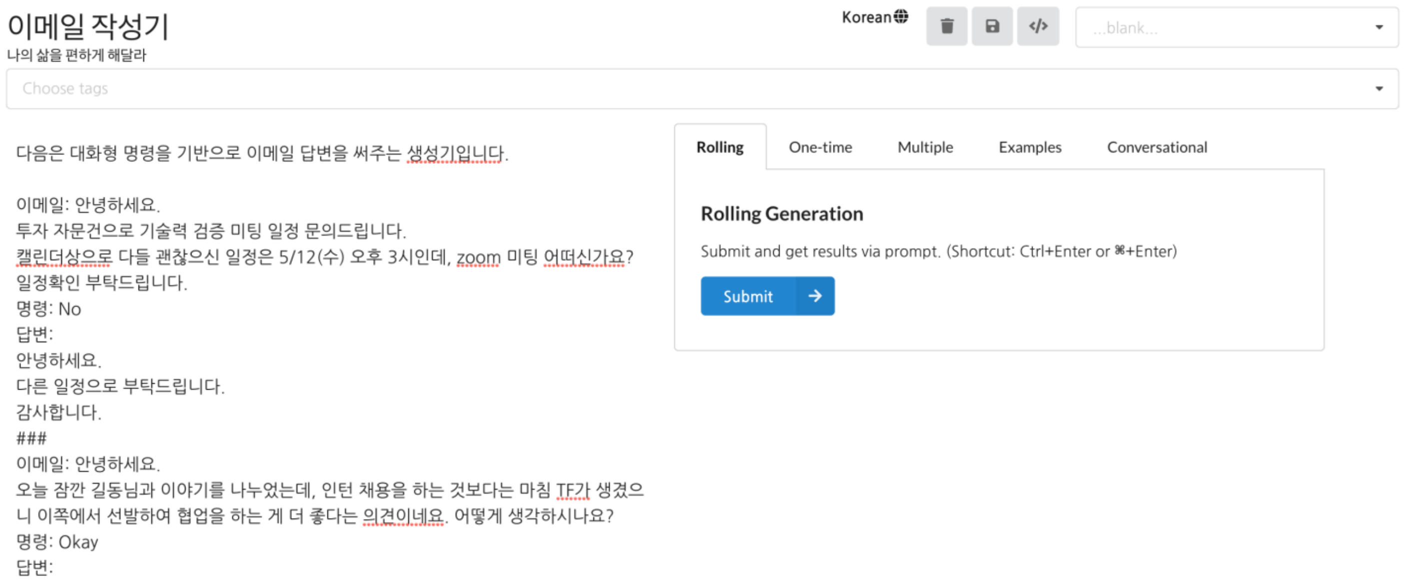
	KorQuAD (EA / F1)		AI Hub (BLEU) Ko -> En	En -> Ko	YNAT (F1)	KLUE-STS (F1)
Ours	**55.28**	**72.98**	3.83	**20.03**	**58.67**	**60.89**
byte-level BPE	51.26	70.34	**4.61**	19.95	48.32	60.45
char-level BPE	45.41	66.10	3.62	16.73	23.94	59.83

<Effect of tokenization approaches on three tasks, HyperCLOVA-1.3B is used for evaluation>

Discussion on Industrial Impacts

- HyperCLOVA Studio
- Case Studies on HyperCLOVA Studio
- No/Low Code AI Paradigm

<HyperCLOVA Studio>



<An Example Interface of HyperCLOVA Studio>

“HyperCLOVA Studio is the place for building and communicating the shared artifact generated by HyperCLOVA”

“HyperCLOVA Studio serves to functions, 1) it can provide a GUI interface, and 2) support API point”

Discussion on Industrial Impacts

- Case Studies on HyperCLOVA Studio

<Case Studies>

(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

<Examples Generated by HyperCLOVA with the Three Different Tasks>

Discussion on Industrial Impacts

- Case Studies on HyperCLOVA Studio

<Zero-shot Transfer Data Augmentation>

사용자 인텐트에 맞는 문장 5개를 만드시오.
(Create five sentences which match the user intent.)

@ 사용자인텐트 : 포장 가능 문의
(@ User intent: Inquiry on takeout)
예시 발화
(Example utterances)
1. 칼국수나 돈까스 같은 음식도 포장되요?
(1. Can I get food like Kalguksu or pork cutlet to go?)
2. 죄송한데 테이크아웃 되죠?
(2. Excuse me, can I takeout this?)
3. 메뉴 포장 되나요?
(3. Can I get this menu to go?)
4. 아이스크림 포장해주세요
(4. I'd like to get this ice cream to go.)
5. 집에서도 먹을 수 있게 포장해주시나요?
(5. Can I get this menu to go so I can eat this at home?)

@ 사용자인텐트 : 배달음식 환불
(@ User intent: refund on delivery food)
예시 발화
(Example utterances)
1. 보쌈에서 시큼한 냄새가 나는데 환불부탁드립니다
(1. Bossam smells sour, please give me a refund.)
2. 메뉴가 잘못 배달 되었습니다. 환불부탁드립니다
(2. The menu was delivered incorrectly. Please give me a) refund.)
3. 간장게장 맛이 이상해요. 환불 가능 한가요?
(3. Soy Sauce Marinated Crab tastes weird. Can I get a refund?)
4. 치킨이 너무 식어서 왔어요. 환불 부탁드려요
(4. The chicken is too cold. I'd like a refund, please.)
5. 음식에서 벌레가 나왔네요. 환불 해주세요
(5. There's a bug in the food. Please give me a refund.)

<A Prompt for Zero-Shot Transfer Data Augmentation>

Discussion on Industrial Impacts

- Case Studies on HyperCLOVA Studio

<Event Title Generation>

키워드: 캔디주얼리, 프로포즈목걸이, 커플링, 은반지, 다이아가드링, 로즈골드목걸이, 하트귀걸이, 하트목걸이
(Keywords: candy jewelry, proposal necklace, coupling, silver ring, diamond guard ring, rose gold necklace, heart earring, heart necklace
날짜: 2021년3월7일
(Date: March 7, 2021)
제목: 화이트데이 커플주얼리 세일
(Title: White Day Couple Jewelry Sale)

키워드: 수입그릇, 빈티지그릇, 법랑냄비, 수저세트, 튼튼한컵, 레트로냄비
(Keywords: imported bowl, vintage bowl, enamel pot, spoon and chopsticks set, strong cup, retro pot
날짜: 2020년4월21일
(Date: April 21, 2020)
제목: 주방용품 해외직구 할인전
(Title: Kitchenware overseas direct purchase discount exhibition)

키워드: 미세먼지, 차량용핸드폰거치대, 세차용품, 자동차용품, 차량용품, 차량무선충전거치대, 차량악세사리, 논슬립패드, 자동차악세사리
(Keywords: fine dust, mobile phone holder for vehicles, car washing products, automobile supplies, vehicle supplies, vehicle wireless charging cradle vehicle accessories, non-slip pads, car accessories)
날짜: 2021년4월1일
(Date: April 1, 2021)
제목: 각종 차량용품 할인 모음전
(Title: Collection of discounts on various vehicle supplies)

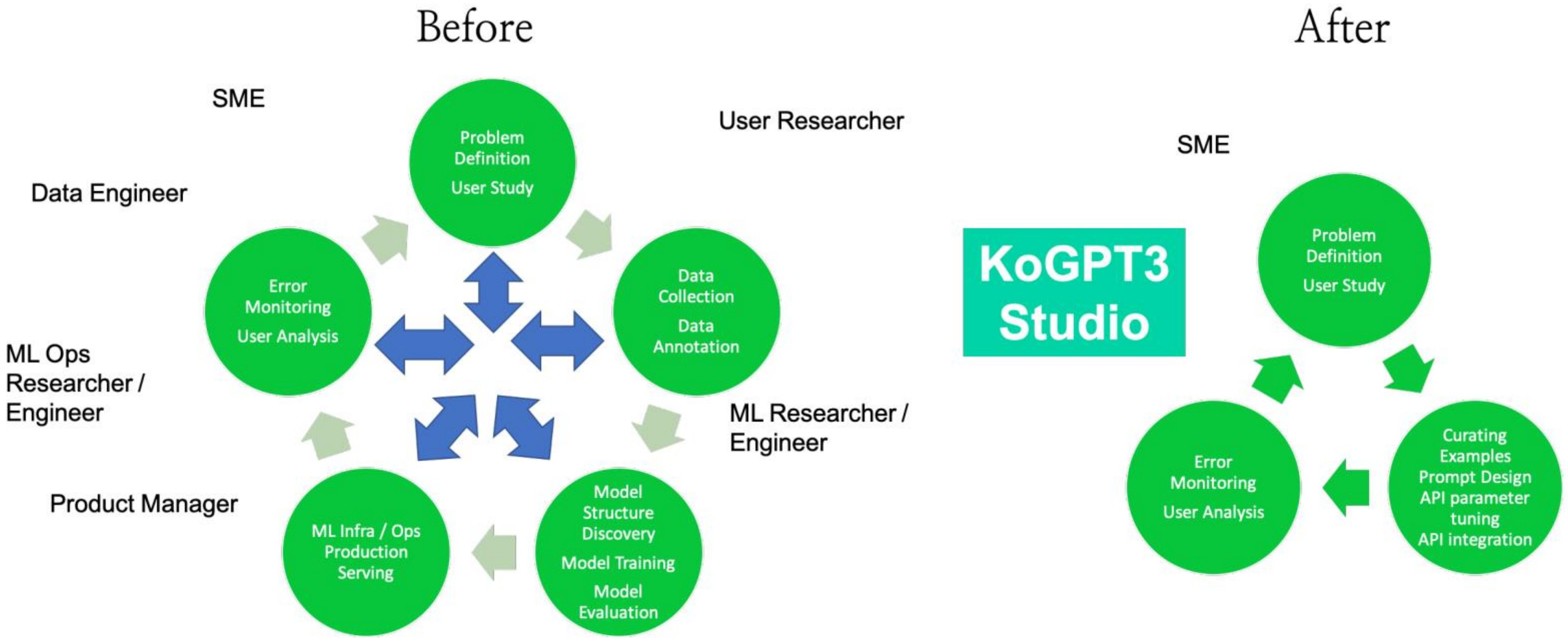
키워드: 슬리퍼, 실내용슬리퍼, 사무용슬리퍼, 하이힐, 봄신상신발, 봄신발, 여자슬리퍼, 여성슬리퍼, 여성하이힐, 여자하이힐
(Keywords: slippers, indoor slippers, office slippers, high heels, spring new arrival shoes, spring shoes, women's slippers, female slippers, women's high heels, female high heels)
날짜: 2021년3월1일
(Date: March 1, 2021)
제목: 봄 여성 사무용 슬리퍼 하이힐 SALE
(Title: Spring women's office slippers high heels SALE)

<Controlling Style by Change In-Context Examples for Title of Online Special Events>

Discussion on Industrial Impacts

- No/Low Code AI Paradigm

<No/Low Code AI Paradigm>



<No Code AI Paradigm in HyperCLOVA Studio>

Conclusion

<Conclusion>

- Introduced HyperCLOVA, a large-scale Korean in-context learning basd LM with nearly 100B parameters, by constructing a large Korean-centric corpus of 360B tokens

- Discovered the effect of language-specific tokenization on large-scale in-context LMs for training corpus of non-English languages

- Explored zero-shot and few-shot capabilities of mid-size HyperCLOVA with 39B and 82B parameters and find that prompt-based tuning can enhance the performance outperforming state-of-the-art on downstream tasks

- Argued the possibility of realizing No Code AI by designing and applying HyperCLOVA Studio to three in-house applications

Any Questions?

Thank You