이해, 생성, 효율: 세 마리 토끼 다 잡는 Seq2Seq HyperCLOVA

유강민
NAVER Cloud

박동주
NAVER Search

가순원
NAVER Search

NAVER DEVIEW 2023

CONTENTS

유강민

1. Sequence-to-Sequence (Seq2Seq) 언어모델 소개

박동주

2. CT5: 현존 가장 강력한 한국어 Seq2Seq 언어모델

가순원

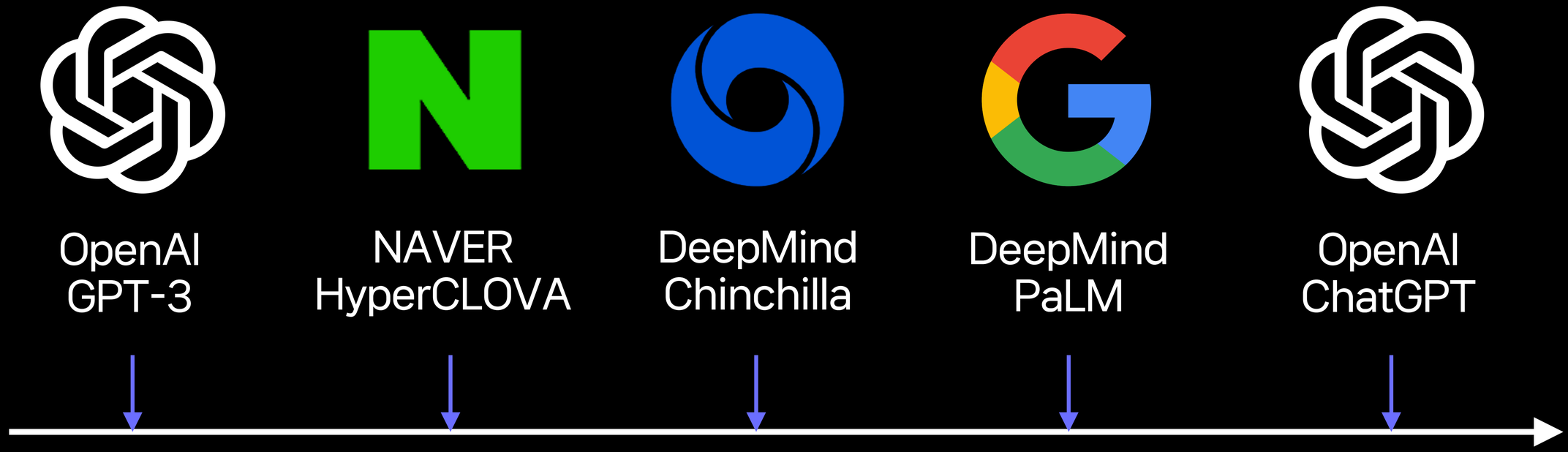
3. DialogCT5: Domain Adaptation으로 데이터 효율 UP!

1. Seq2Seq 언어 모델 소개

1.1. GPT 언어모델의 발전

NAVER
DEVIEW
2023

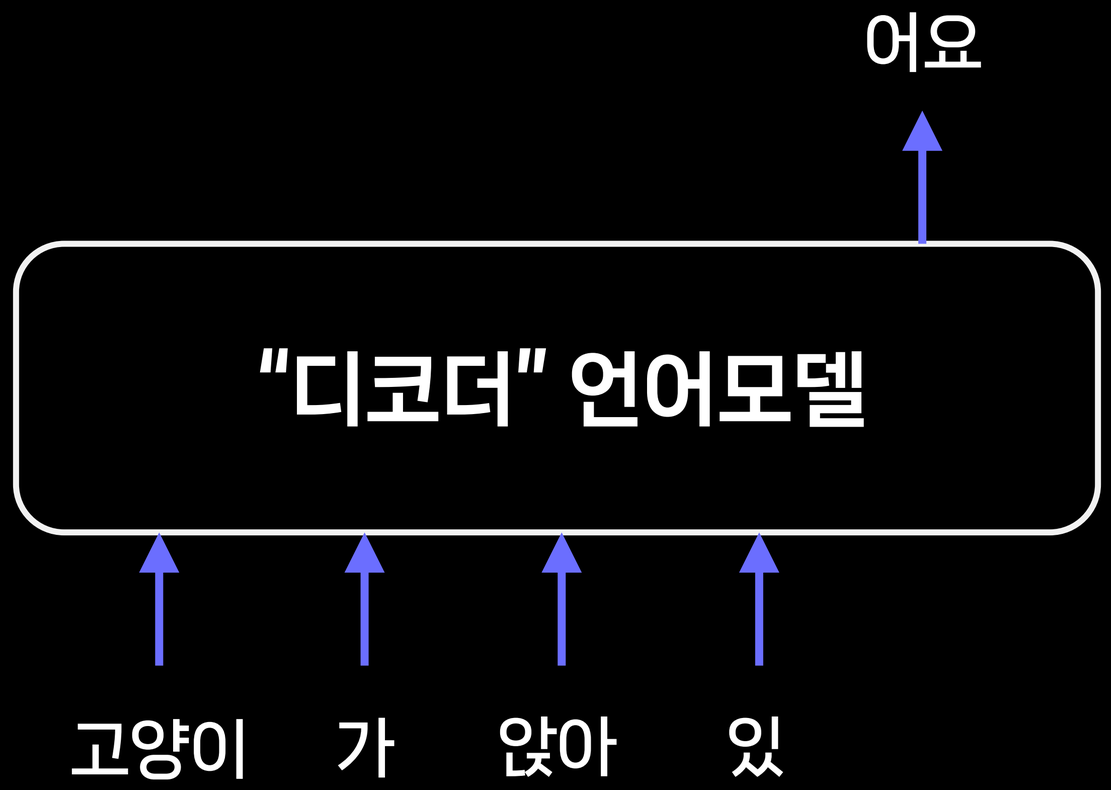
- 최근 GPT 계열의 초거대 언어모델들이 발전하여 일상생활까지 침투



1.1. GPT 언어모델의 발전

NAVER
DEVIEW
2023

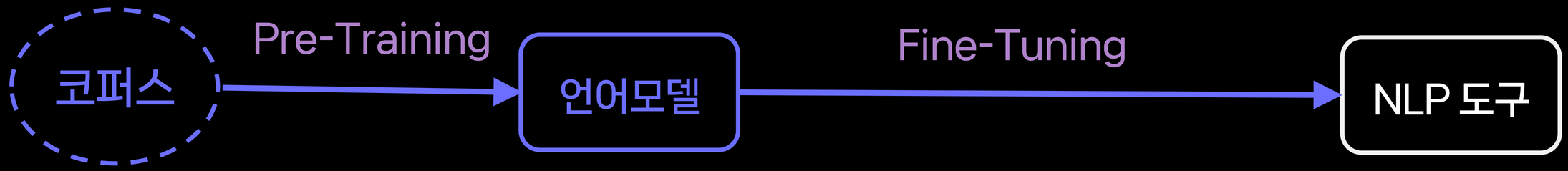
- 언어모델은 다음 자언어 덩어리(단어, 어절 등)를 예측하도록 학습된 기계모델
- Transformer(Vaswani et al., 2017)기반 인공신경망 구조



1.1. GPT 언어모델의 발전

NAVER
DEVIEW
2023

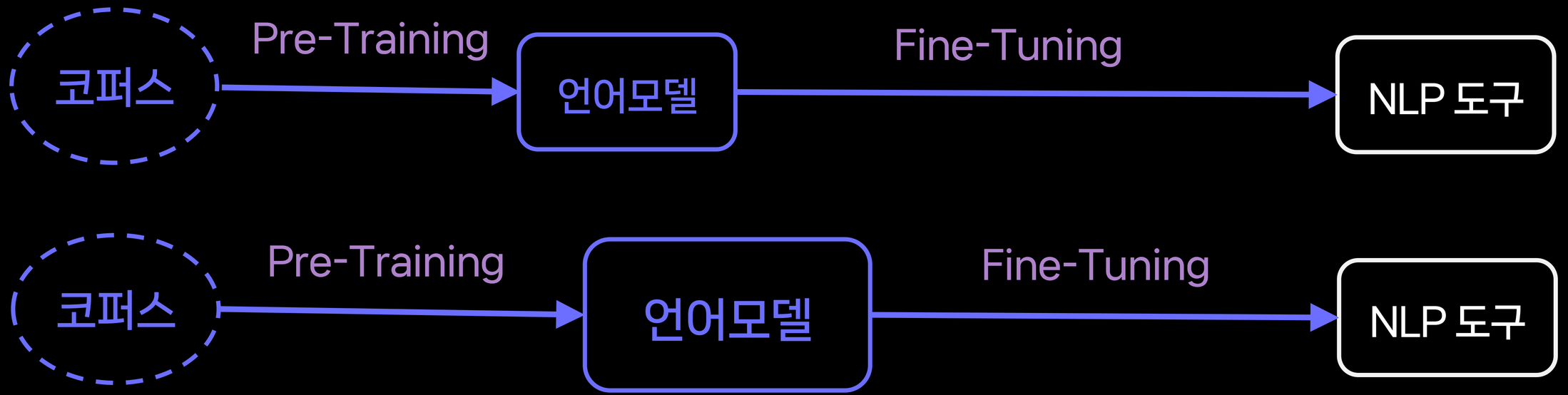
- 언어모델은 Pre-Training과 Fine-Tuning 단계로 나뉨



1.1. GPT 언어모델의 발전

NAVER
DEVIEW
2023

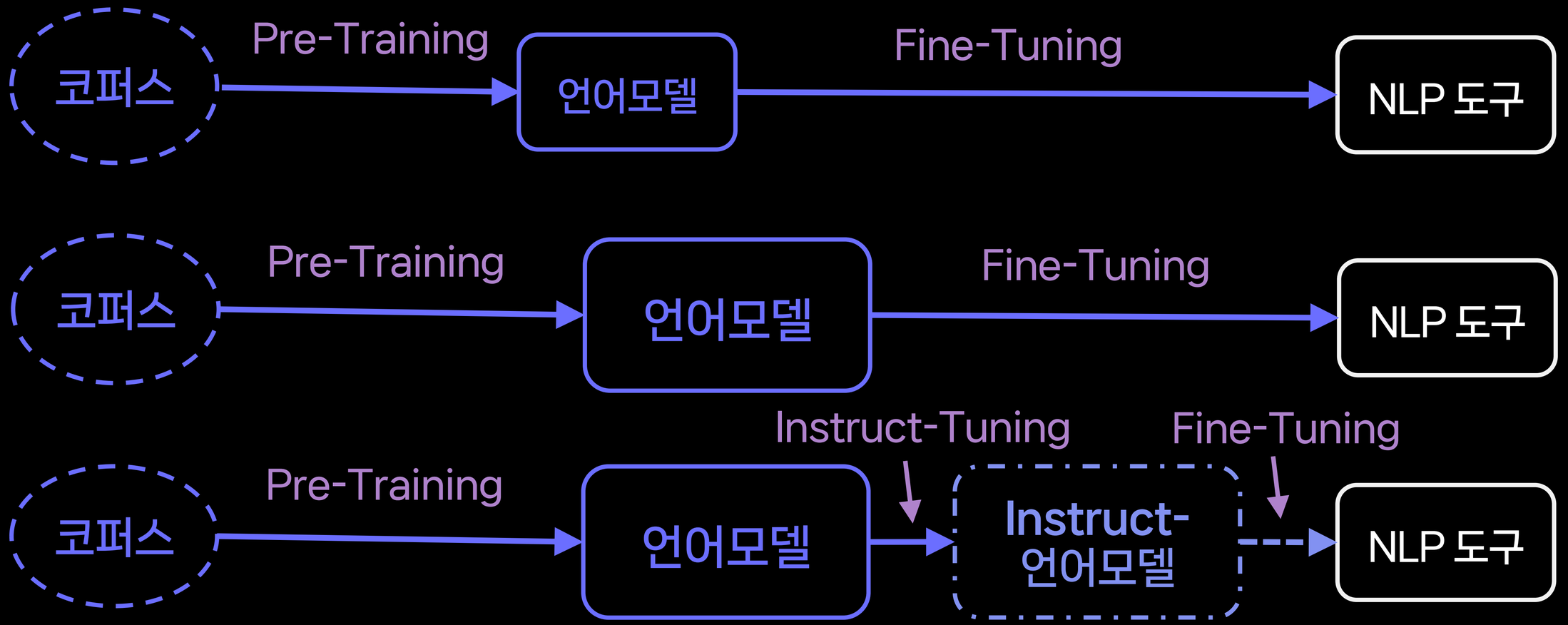
- 언어모델은 Pre-Training과 Fine-Tuning 단계로 나뉨



1.1. GPT 언어모델의 발전

NAVER
DEVIEW
2023

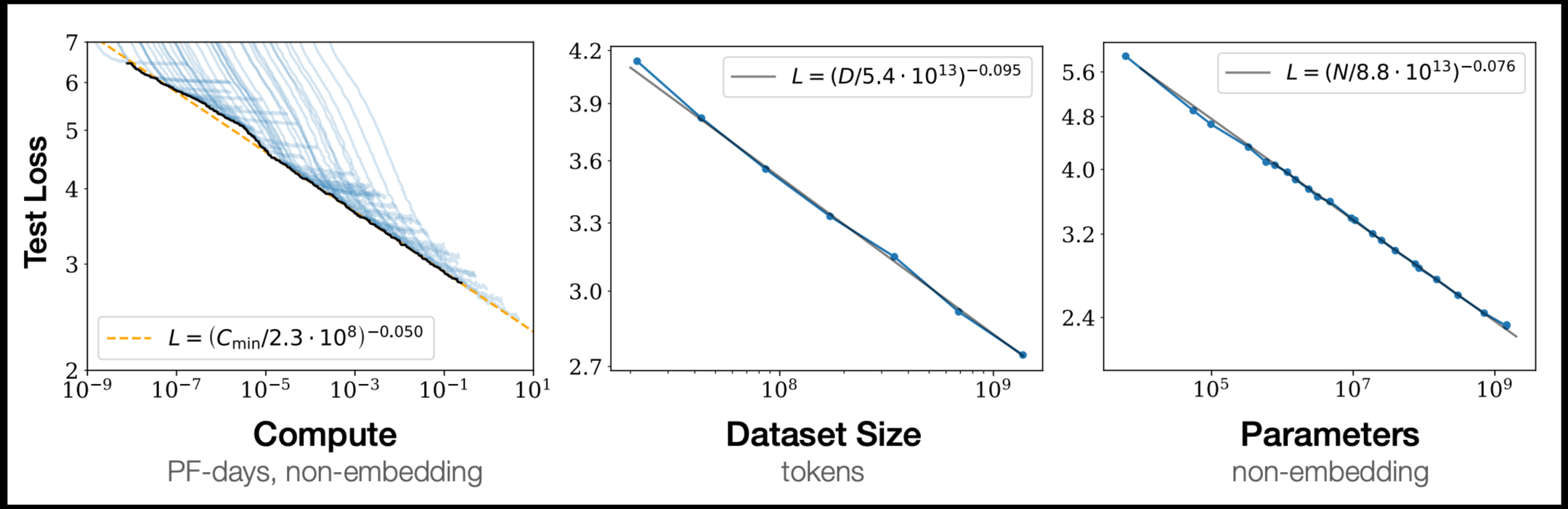
- Instruction-Tuning을 통해 언어모델의 사용자 활용 가치 증폭



1.1. GPT 언어모델의 발전

NAVER
DEVIEW
2023

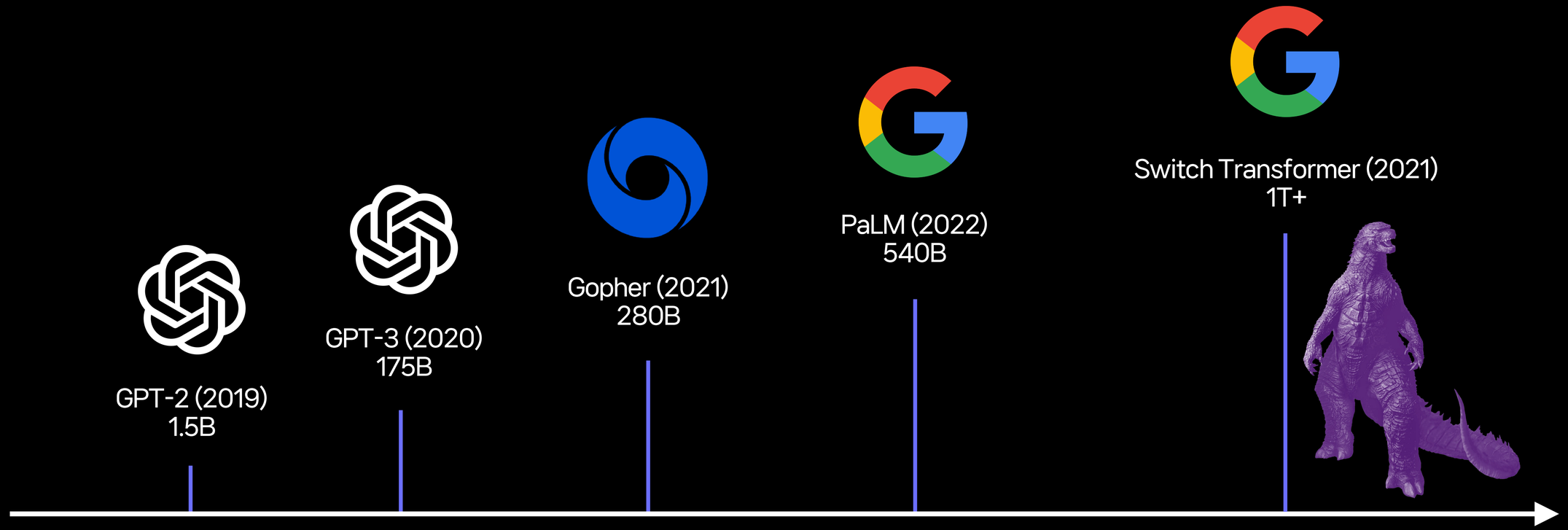
- "Scaling Laws for Neural Language Models" (Kaplan et al., 2020)
- 연산 자원 (Compute), 데이터양, 모델 크기를 투입할 수록 성능 개선 확인



1.1. GPT 언어모델의 발전

NAVER
DEVIEW
2023

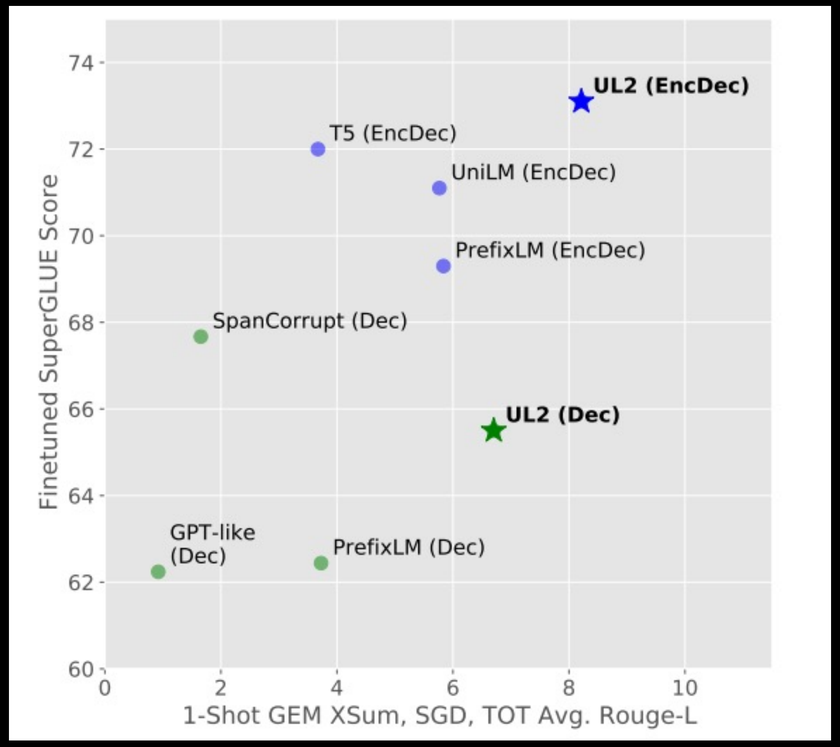
- Scaling Laws에 따라 언어모델의 크기가 비대해짐
- 개인이나 일반적인 환경에서 범접할 수 없는 규모



1.2. Seq2Seq vs GPT

NAVER
DEVIEW
2023

- Seq2Seq은 작은 모델로 뛰어난 성능을 달성할 수 있는 것이 장점 중 하나
- Few-shot learning 문제에서 GPT보다 더 우월한 성능 보임 (UL2)

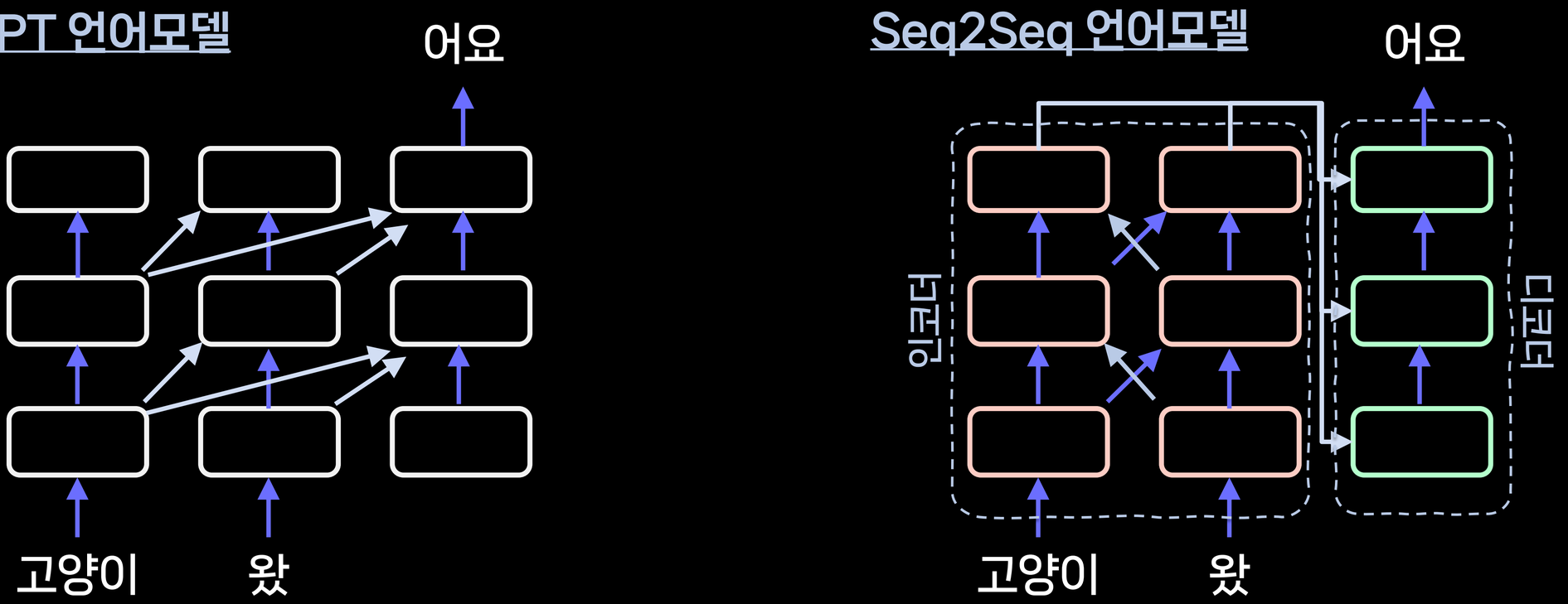


UL2 (Tay et al., 2022)

1.2. Seq2Seq vs GPT

NAVER
DEVIEW
2023

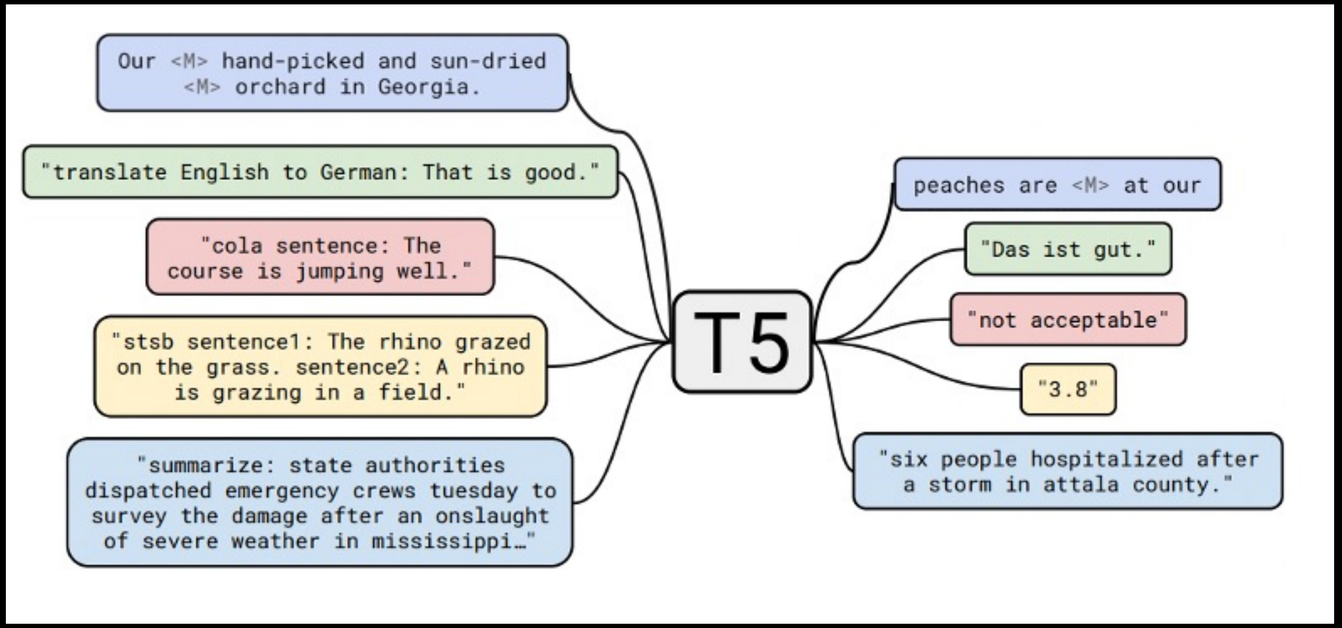
- 두 언어모델 구조 모두 attention mechanism기반 Transformer 사용
- Encoder-Decoder 사이 한정된 정보 통로로 Information Bottleneck 발생



1.3. Seq2Seq을 선택해야하는 이유

NAVER
DEVIEW
2023

- 첫 번째 이유: 손쉬운 태스크 확장
- Unified Text-to-Text Transformer 계열 언어 모델의 강력한 장점
- 숫자, 레이블 등 모든 정보를 텍스트로 변환하여 추론, 예측

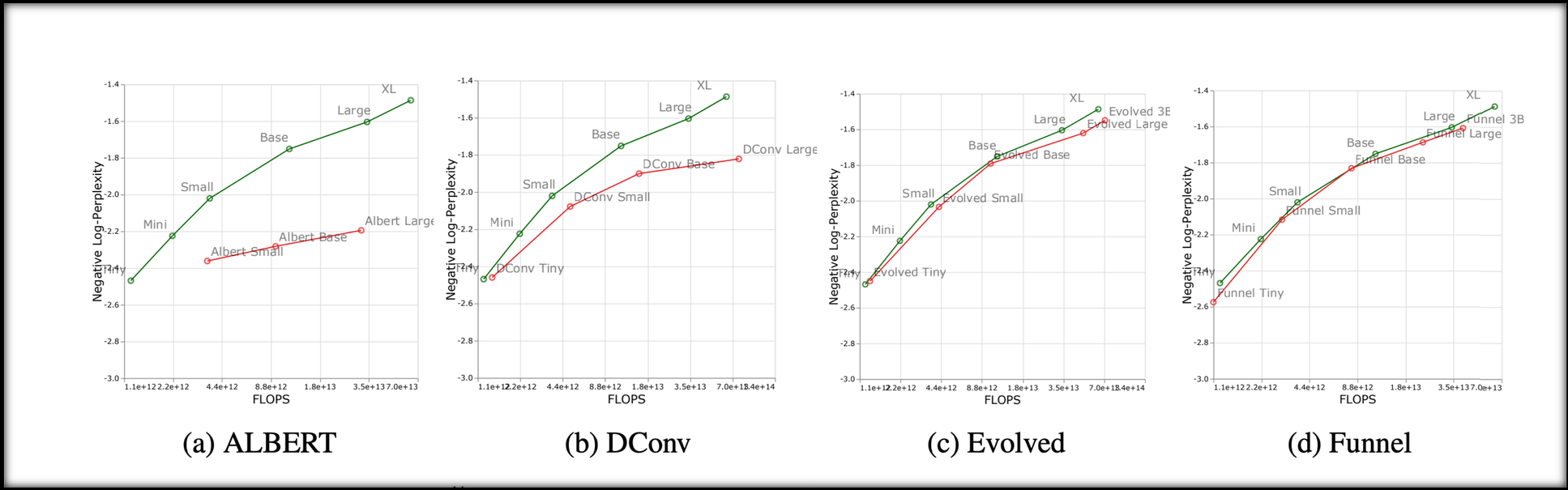


T5 (Raffel et al., 2019)

1.4. Seq2Seq을 선택해야하는 이유

NAVER
DEVIEW
2023

- 두 번째 이유: Information Bottleneck을 통해 효율적인 Scaling Law 달성
- Seq2Seq 계열 언어모델에서도 Scaling Law 발현 (Tay et al., 2022)
- 저연산 구간에서 GPT모델 대비 우세한 Scaling Law 나타냄 [후속 참조]



Scaling Laws vs Model Architectures (Tay et al., 2022)

1.4. Seq2Seq을 선택해야하는 이유

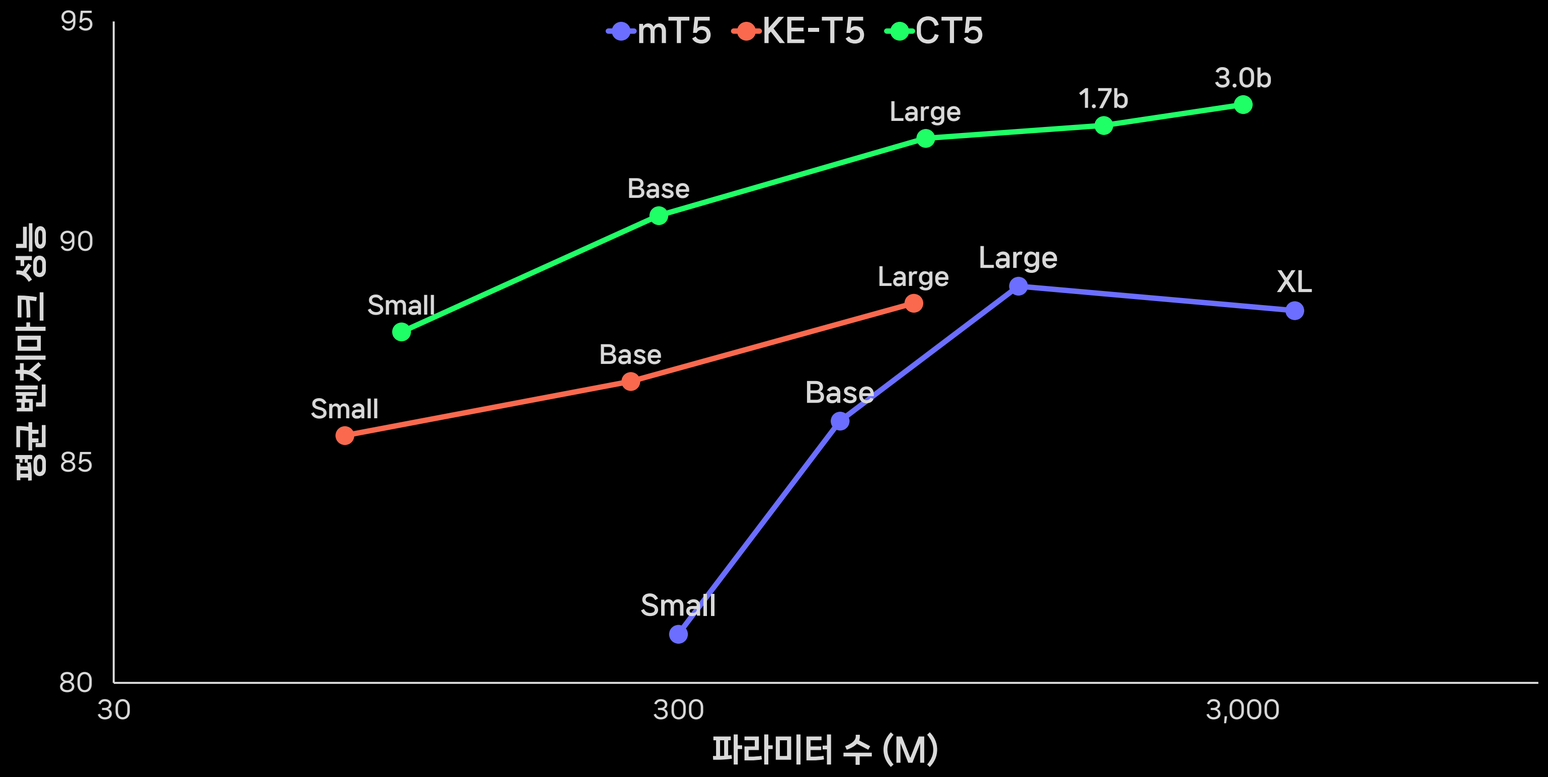
NAVER
DEVIEW
2023

- 세 번째 이유: 공짜 점심이 있다?! 쓸만한 Text Encoder의 자연 창발
- Encoder는 텍스트를 실수 형태(Embedding)로 변환하는데 사용
- Information Bottleneck의 부속 현상

1.5. 한국어 Seq2Seq, 그리고 HyperCLOVA

NAVER
DEVIEW
2023

- 그래서 HyperCLOVA 코퍼스 기반으로 CT5를 만들었습니다.
- 최대 크기 3B, 한국어 이해 및 생성 SOTA 수준



1.5. 한국어 Seq2Seq, 그리고 HyperCLOVA

NAVER
DEVIEW
2023

- HyperCLOVA와 CT5는 **상생 관계**
- HyperCLOVA와 CT5는 Text기반 Transformer로서 기술 상호호환
- 기민한 CT5 모델 연구를 통해 HyperCLOVA의 적용 가능한 요소 선행 발굴

1.6. 본 발표를 보시며

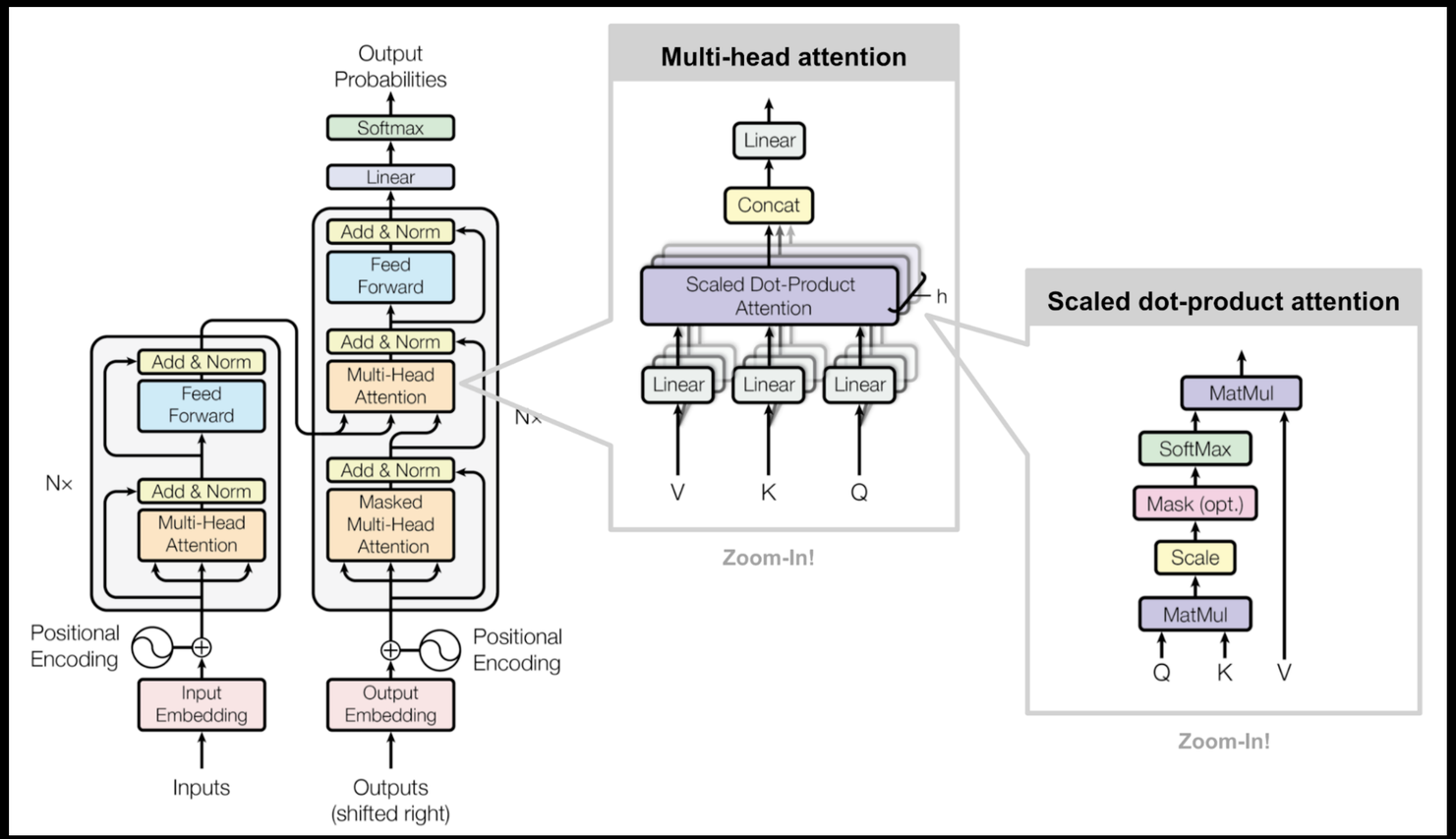
NAVER
DEVIEW
2023

- Compute-Optimality: 같은 연산 자원 수준 내에서 디코더 언어모델보다 더 효율적일 수 있는가?
- Domain Adaptation: Fine-Tuning의 데이터 효율을 높이는 방법에 관한 고찰

2. CT5: 현존 가장 강력한 한국어 Seq2Seq 언어모델

2.1. CT5 구조

NAVER
DEVIEW
2023



Transformers 구조 기반

하지만...

- No bias
- RMS Norm
- Relative position embedding
- Residual after layer norm
- ...

https://deepfrench.gitlab.io/deep-learning-project/

2.1. CT5 구조

NAVER
DEVIEW
2023

New T5 model : T5.1.1

- 기본적인 구조는 T5 구조와 동일하지만 모델을 이루는 컴포넌트들의 하이퍼 파라미터 변경

- 전반적으로 성능 향상

	Activation	Dropout (pre-training)	Dataset	Embedding parameter sharing
T5	ReLU	On	C4 w/ downstream task	On
T5.1.1	GEGLU	Off	C4	Off

https://github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md

2.1. CT5 구조

NAVER
DEVIEW
2023

T5를 학습하는 방법

- 주어진 코퍼스를 가장 효율적으로 학습할 수 있는 Unsupervised objective는 무엇일까?
- Transformers 구조에서 사용할 수 있는 다양한 Objective 후보

- 원본 텍스트 : **지난 주 너의 파티에 나를 초대해줘서 고마워**

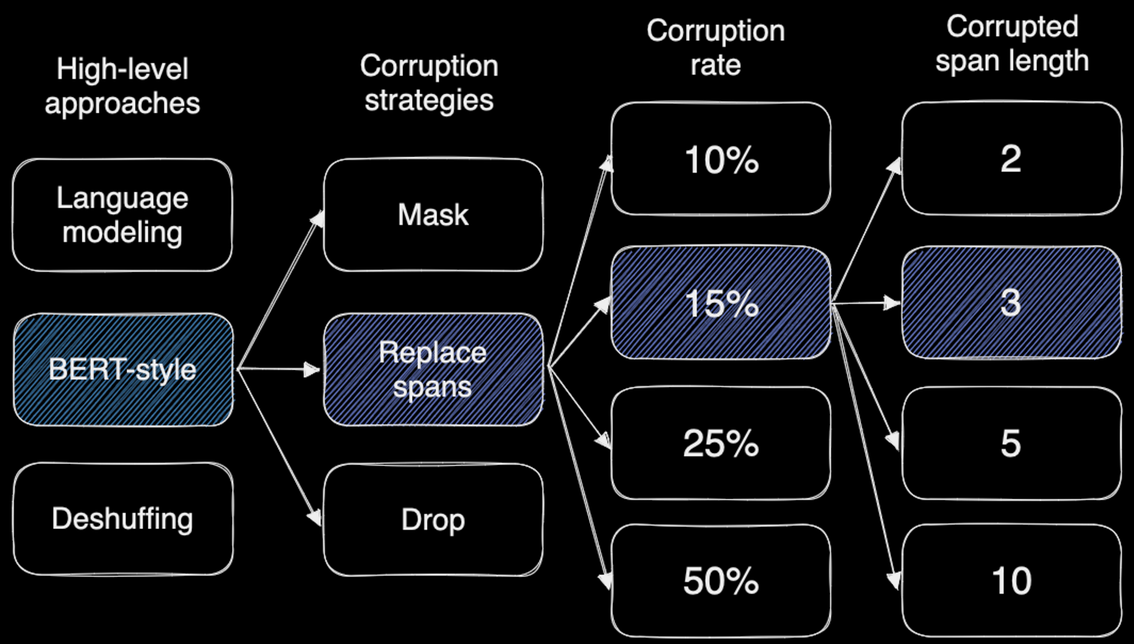
Objective	Inputs	Targets
Prefix language modeling	지난 주 너의 파티에	나를 초대해줘서 고마워
BERT-style	지난 주 <M> <M> 나를 초청해줘서 고마워	*(원본 텍스트)*
Deshuffling	나를 초대해줘서 고마워 지난 주 너의 파티에	*(원본 텍스트)*
I.i.d. noise, replace span	지난 주 <X> 파티에 <Y> 고마워	<X> 너의 <Y> 초대해줘서 <Z>
I.i.d. noise, drop tokens	지난 주 너의 고마워	파티에 초대해줘서

2.1. CT5 구조

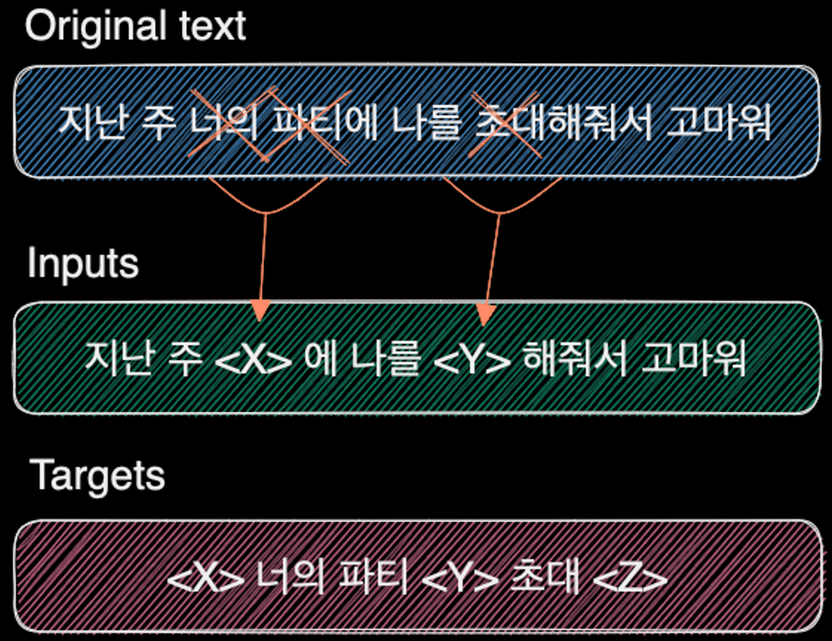
NAVER
DEVIEW
2023

T5를 학습하는 방법

- 수 많은 실험을 통해 얻은 답 : BERT-style Span Corruption



A flow chart of exploration of unsupervised objective

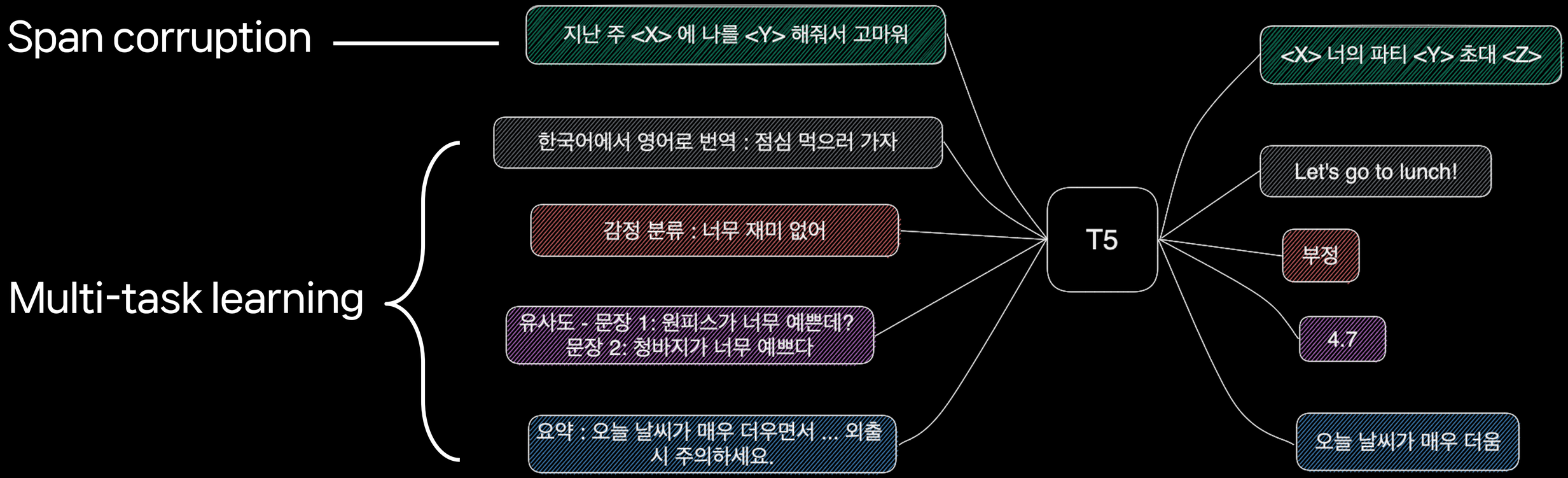


Schematic of the objective

2.1. CT5 구조

NAVER
DEVIEW
2023

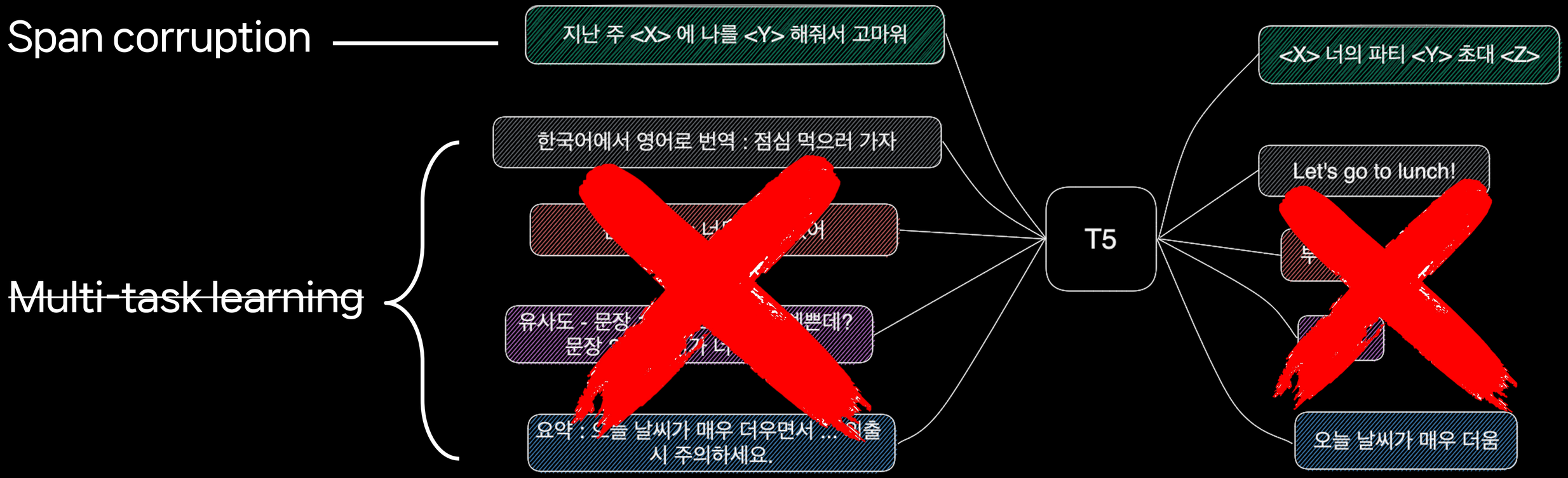
T5의 학습 방법



2.1. CT5 구조

NAVER
DEVIEW
2023

CT5의 학습 방법



2.2. 학습 세팅

NAVER
DEVIEW
2023

학습 데이터

- 서로 다른 두가지 코퍼스에 대하여 사전 학습
- Downstream task 에 대한 Fine-tuning 성능을 바탕으로 코퍼스 결정

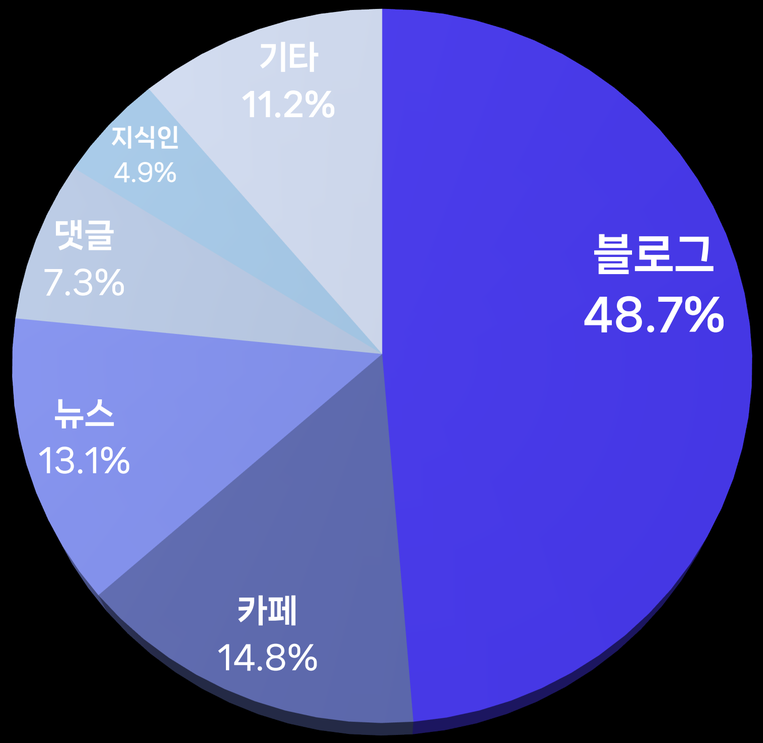
2.2. 학습 세팅

NAVER
DEVIEW
2023

학습 데이터

1. HyperCLOVA 코퍼스
   - 블로그 / 카페 / 뉴스 데이터 위주
   - 구어체 다수 포함
   - Noisy data 다수 존재 가능

- 코퍼스 내 561.8B 토큰 중 300B 토큰만 사용



데이터 비율

What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers (Kim et al., 2021)

2.2. 학습 세팅

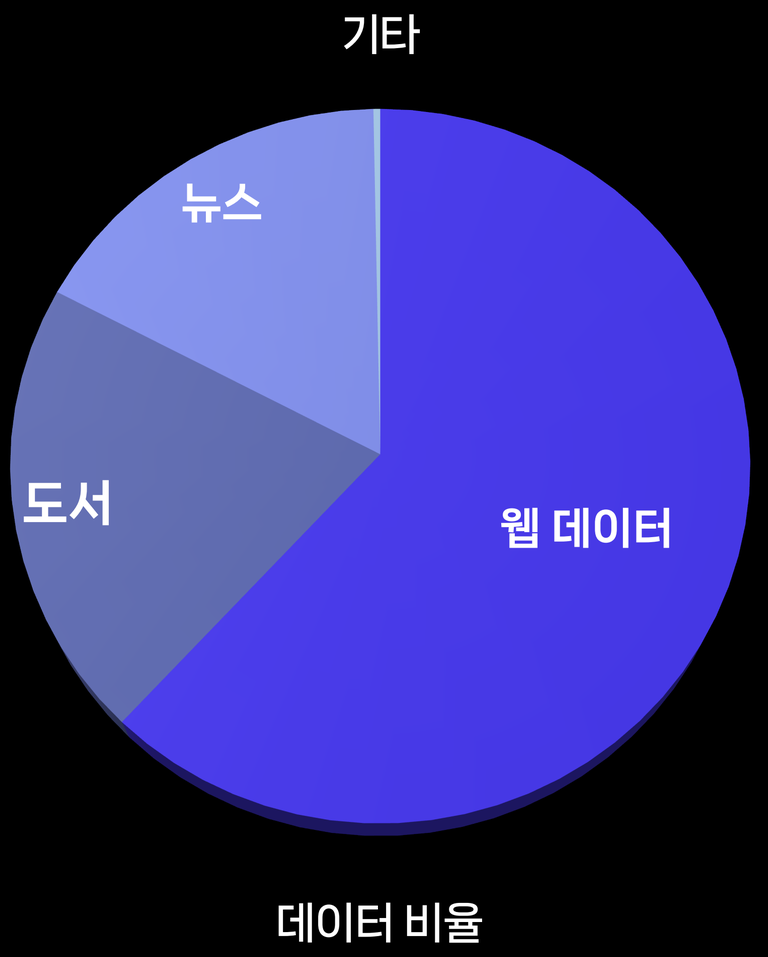
NAVER
DEVIEW
2023

학습 데이터

2. 지식 기반 코퍼스

- HyperCLOVA 코퍼스 기반 리샘플링
- 지식 관련 코퍼스 비율 증가 (Book, News)
- 영어 데이터 비율 증가 (3% -> 20%)

- Unique tokens 는 감소하지만 300B에 맞춰서 학습



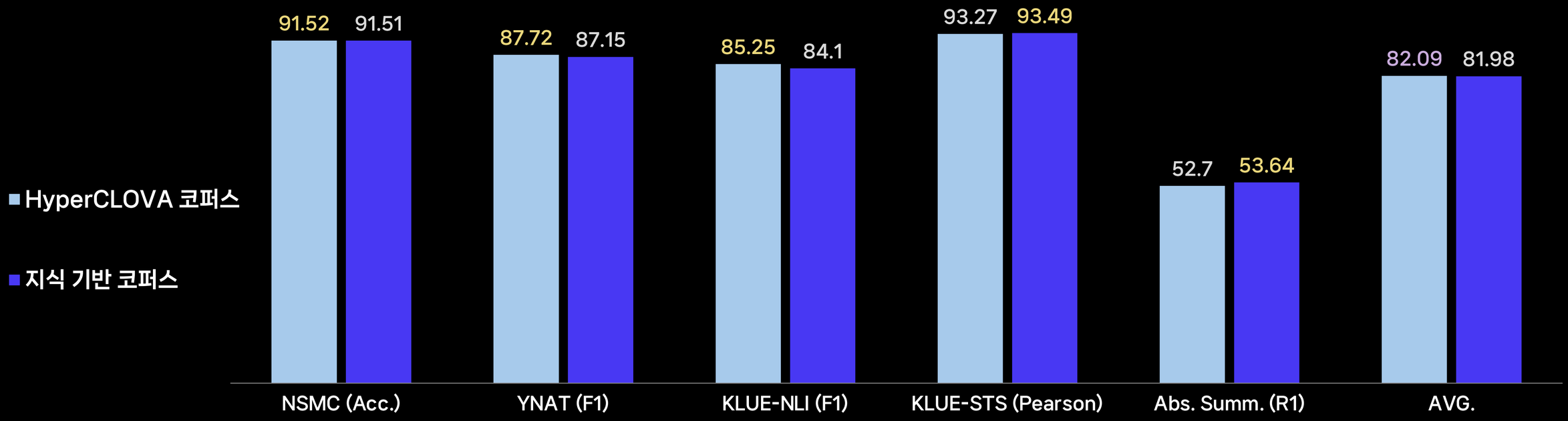
2.2. 학습 세팅

NAVER
DEVIEW
2023

학습 데이터

실험결과

- 언어 이해 테스크 : HyperCLOVA 코퍼스 우위
- 언어 생성 테스크 : 지식 기반 코퍼스 우위



2.2. 학습 세팅

NAVER
DEVIEW
2023

안정적인 학습을 위한 열쇠

- Mixed precision training with BFLOAT16

- From AdamW optimizer to Adafactor optimizer

2.2. 학습 세팅

NAVER
DEVIEW
2023

안정적인 학습을 위한 열쇠

- Mixed precision training
- 다양한 정밀도 중 무엇을 써야 하나?

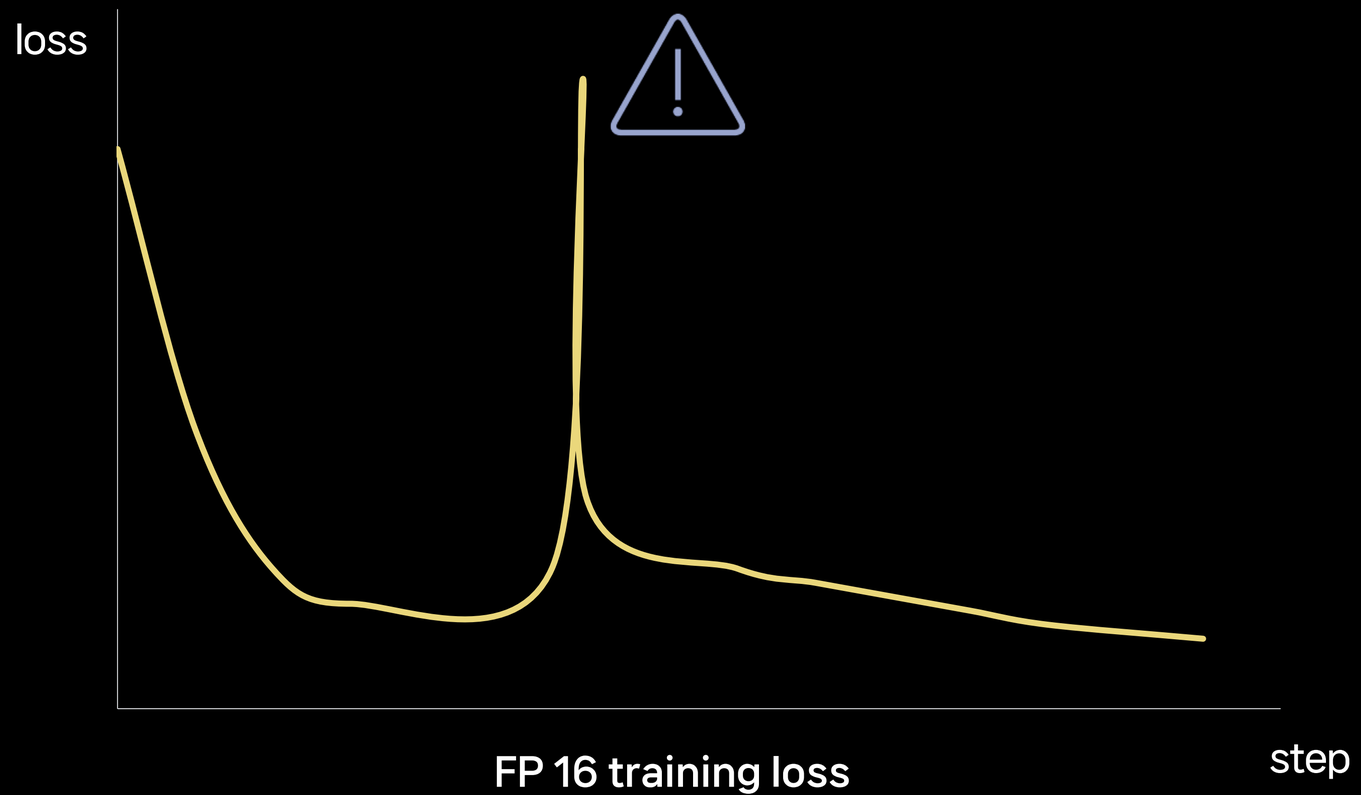
정밀도 별 지수와 가수

2.2. 학습 세팅

NAVER
DEVIEW
2023

안정적인 학습을 위한 열쇠

- Mixed precision training with FP16



해결을 위한 다양한 시도

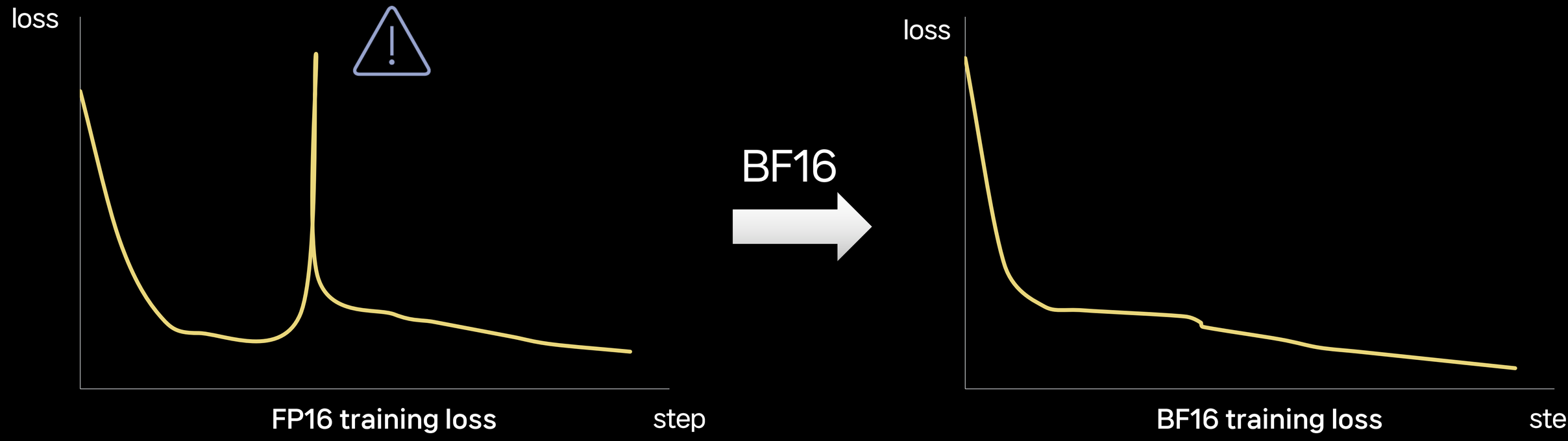
- Learning rate
- Optimizer
- Gradient clipping
- Loss 코드 수정
- …

2.2. 학습 세팅

NAVER
DEVIEW
2023

안정적인 학습을 위한 열쇠

- Mixed precision training with BF16
- FP16에 비해 더 넓은 범위의 수를 표현할 수 있는 BF16 으로 해결!

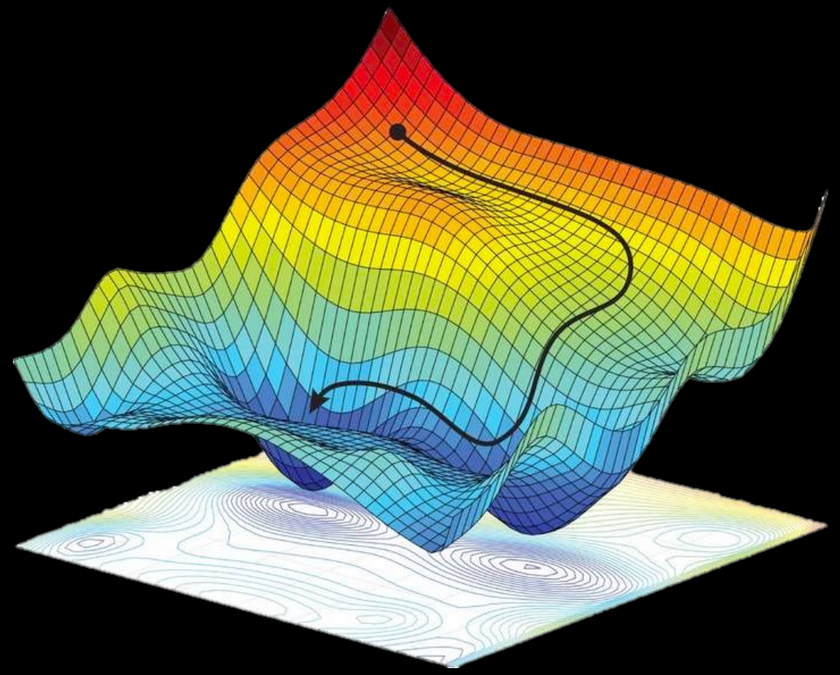


2.2. 학습 세팅

NAVER
DEVIEW
2023

안정적인 학습을 위한 열쇠

- 다양한 종류의 Optimizer
- 각 Optimizer 가 갖는 다양한 하이퍼파라미터



- 모든걸 고려하기엔 너무 많은 경우의 수

→ 대표적인 AdamW 와 Adafactor 만 비교!

2.2. 학습 세팅

NAVER
DEVIEW
2023

안정적인 학습을 위한 열쇠

실험 세팅

- Model : CT5-small

- Optimizer

1. AdamW-1 : Learning rate 1e-3 / Warmup 10k
2. AdamW-2 : Learning rate 6e-4 / Warmup 0.36k
3. Adafactor : Learning rate $\frac{1}{\sqrt{\max(n,k)}}$ (n=step, k=warmup step) / Warmup 10k

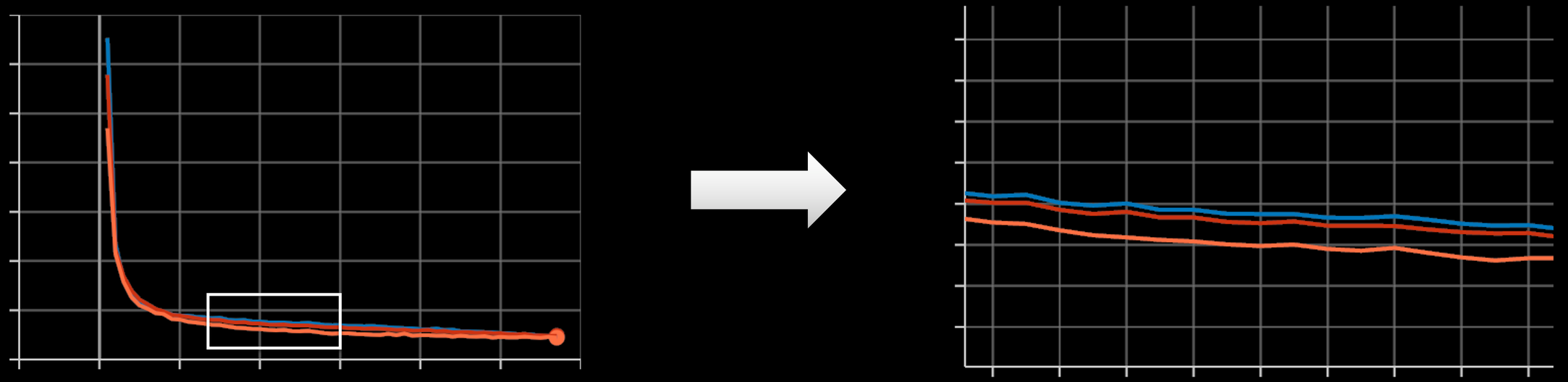
2.2. 학습 세팅

NAVER
DEVIEW
2023

안정적인 학습을 위한 열쇠

- 학습 중반에는 옵티마이저 별 loss 차이가 최대 2%

- 하지만 학습 후반에는 loss 차이 미미



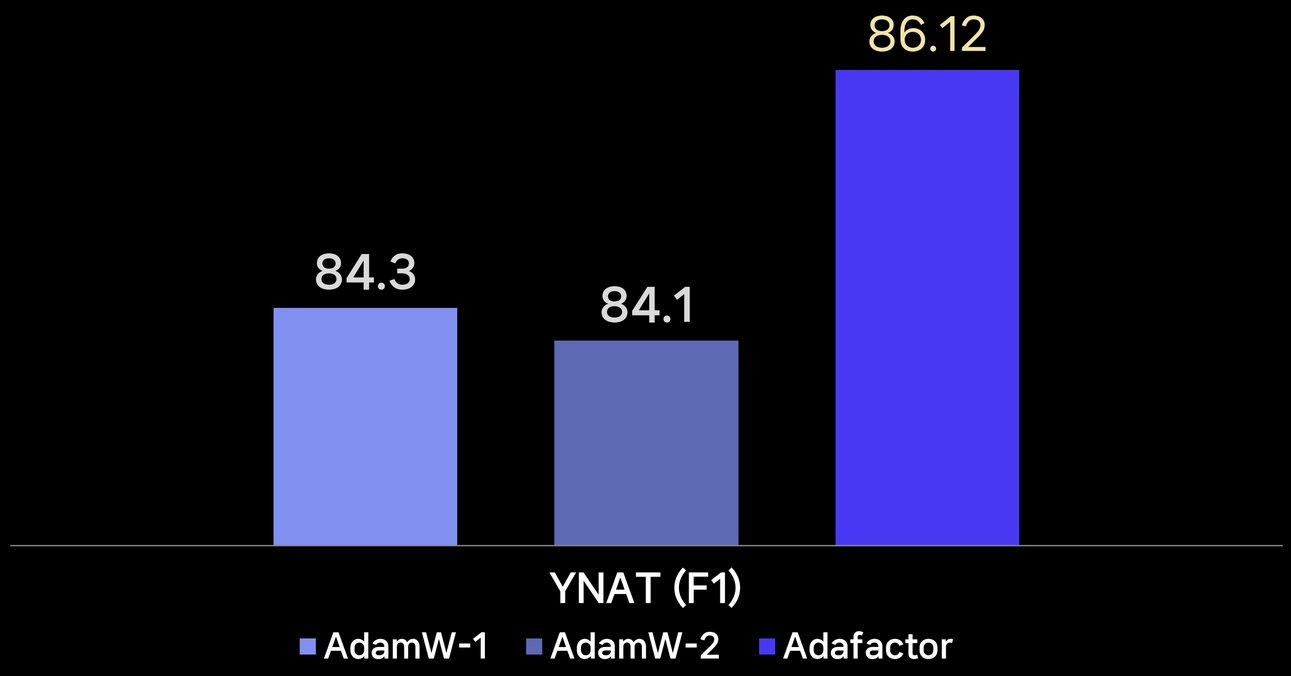
2.2. 학습 세팅

NAVER
DEVIEW
2023

안정적인 학습을 위한 열쇠

- T5 계열에서는 Pretraining loss 만으로 모델 성능 판별 불가능

- Downstream task (Topic classification) 의 성능이 Adafactor 가 우위



2.2. 학습 세팅

NAVER
DEVIEW
2023

Putting It All Together

- Architecture : T5.1.1 (Huggingface)
- Corpus : HyperCLOVA corpus w/o Multi-task learning
- Precision : BF16
- Optimizer : Adafactor

	Small	Base	Large	1.7B	3B
# of parameters	97M	277M	822M	1.7B	3B

모델 별 파라미터 수

2.3. 벤치마크

NAVER
DEVIEW
2023

다양한 테스크에 대한 벤치마크

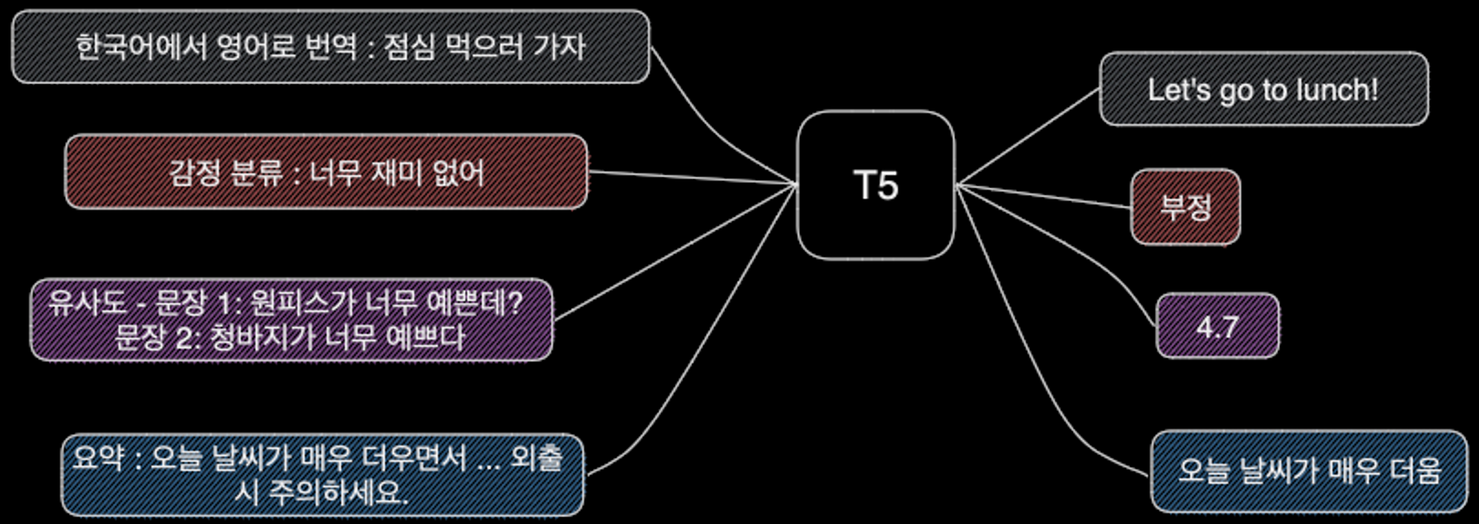
- Natural Language Understanding
- Natural Language Generation
- Parameter and Memory Efficient Finetuning

2.3. 벤치마크

NAVER
DEVIEW
2023

Natural Language Understanding

- NSMC : 영화 리뷰 감정 분류
- YNAT : 토픽 분류
- KLUE NLI : 자연어 추론
- KorQuAD : 기계 독해
- KLUE-STS : 유사도 예측

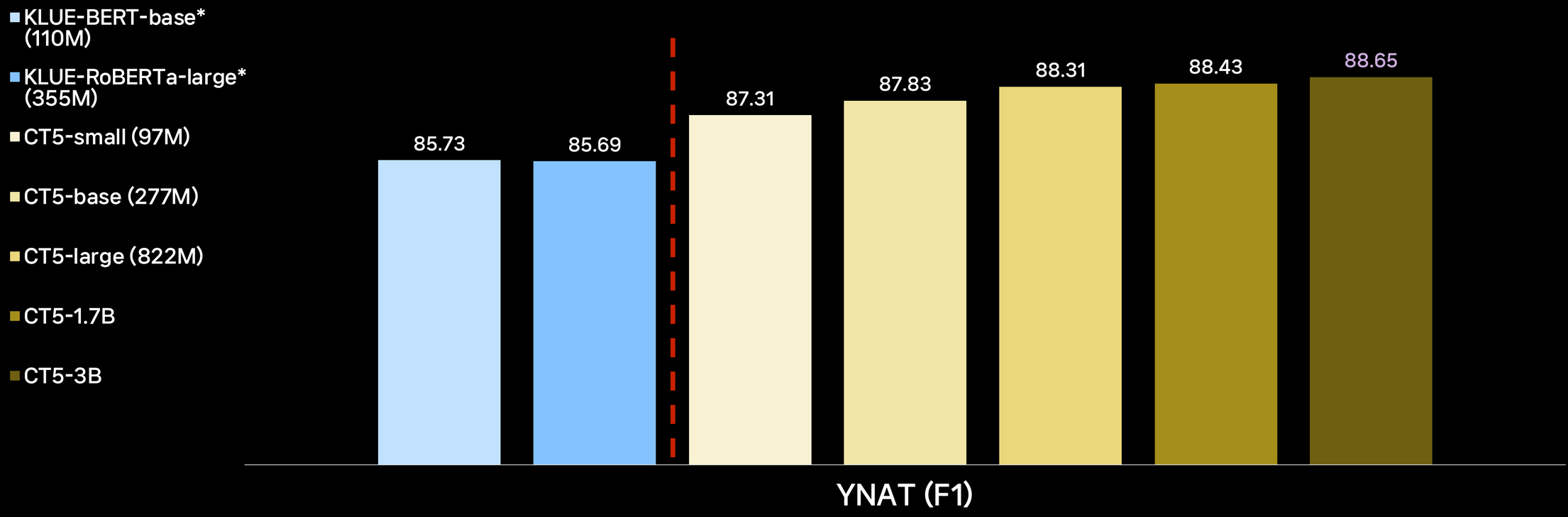


2.3. 벤치마크

NAVER
DEVIEW
2023

Natural Language Understanding

- Encoder 모델과 비교



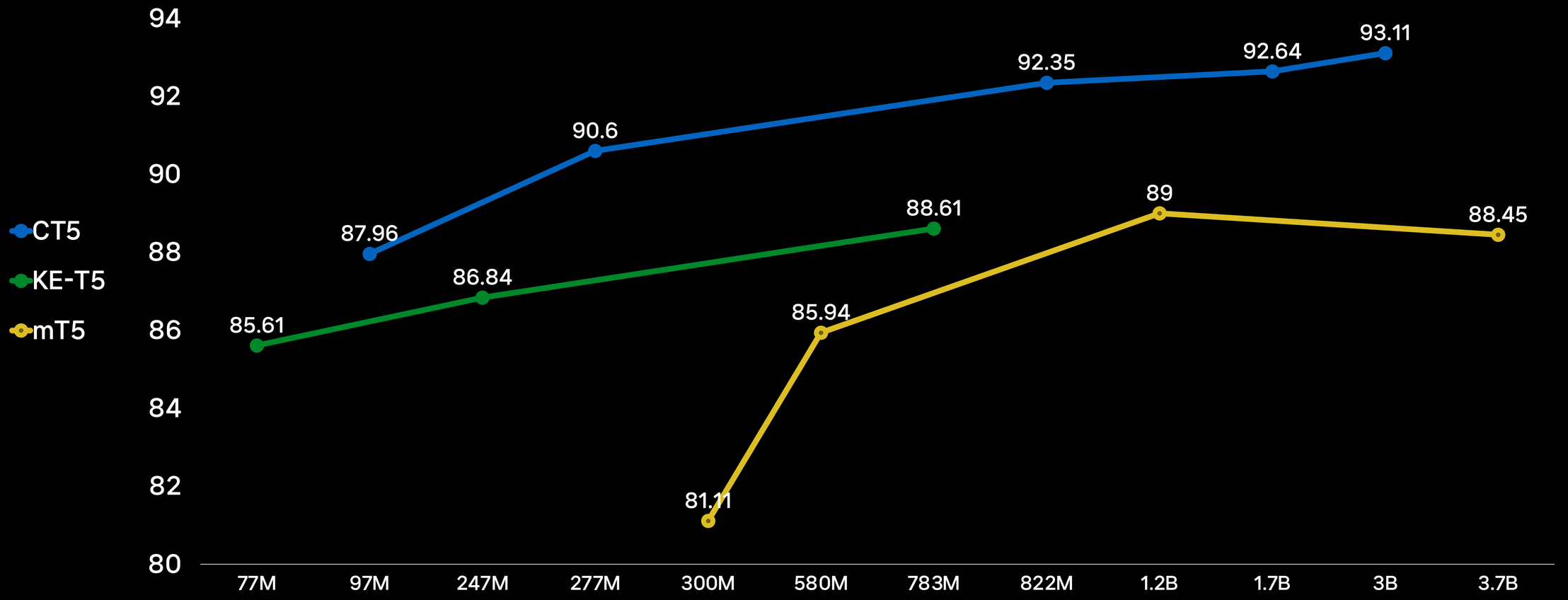
KLUE: Korean Language Understanding Evaluation (Park et al., 2021)

2.3. 벤치마크

NAVER
DEVIEW
2023

Natural Language Understanding

- Encoder-Decoder 모델과 비교 (NLU 테스크 평균)

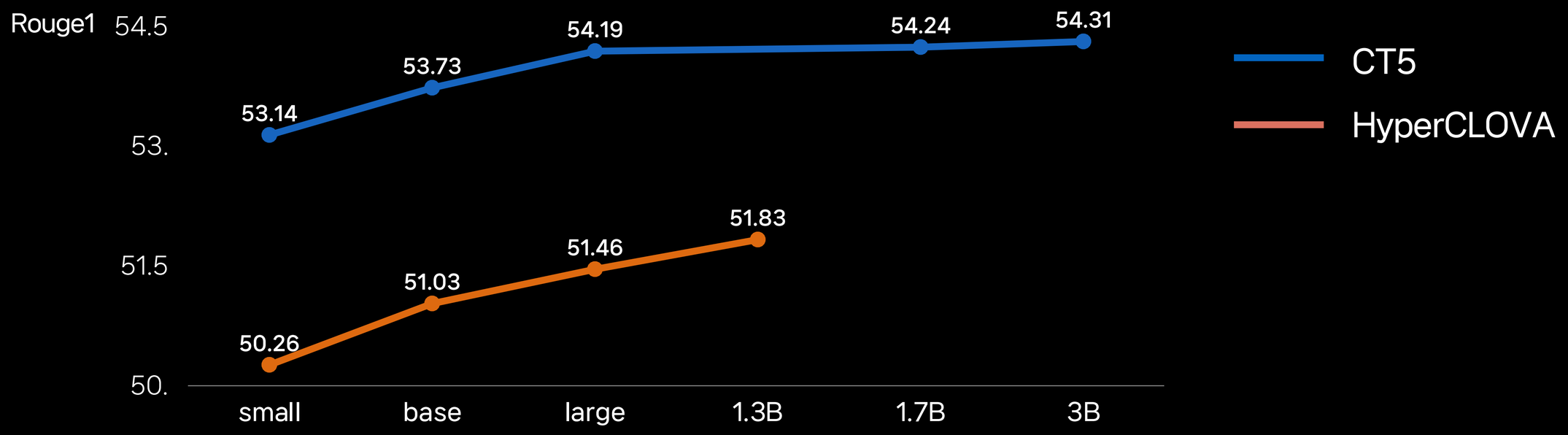


2.3. 벤치마크

NAVER
DEVIEW
2023

Natural Language Generation

- Summarization
- Input : "*요약: 우리나라는 사계절이 뚜렷하며 봄, 여름, 가을, 그리고 겨울... (생략)*"
- Output : "*우리나라는 사계절이 뚜렷하다*"



2.3. 벤치마크

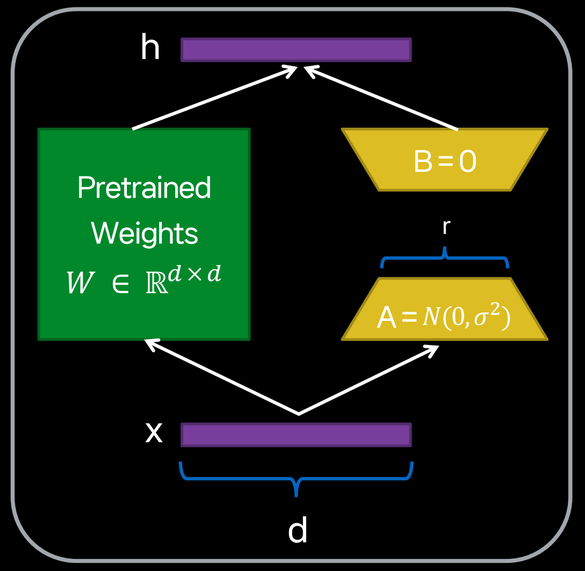
NAVER
DEVIEW
2023

Large-scale Language Model 의 효율적인 학습 방법

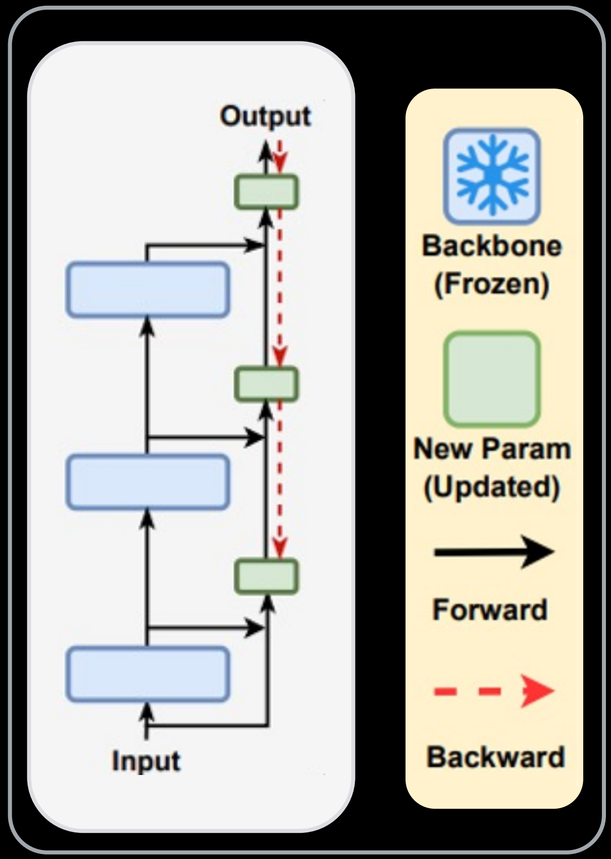
Parameter and Memory Efficient Finetuning

- 파라미터 효율 학습
  - LoRA: Low-Rank Adaption
  - Full fine-tuning 대비 최대 0.09% 의 파라미터만 학습

- 메모리 효율 학습
  - LST : Ladder Side Tuning
  - Full fine-tuning 대비 메모리 사용량 최대 69% 감소



LoRA



LST

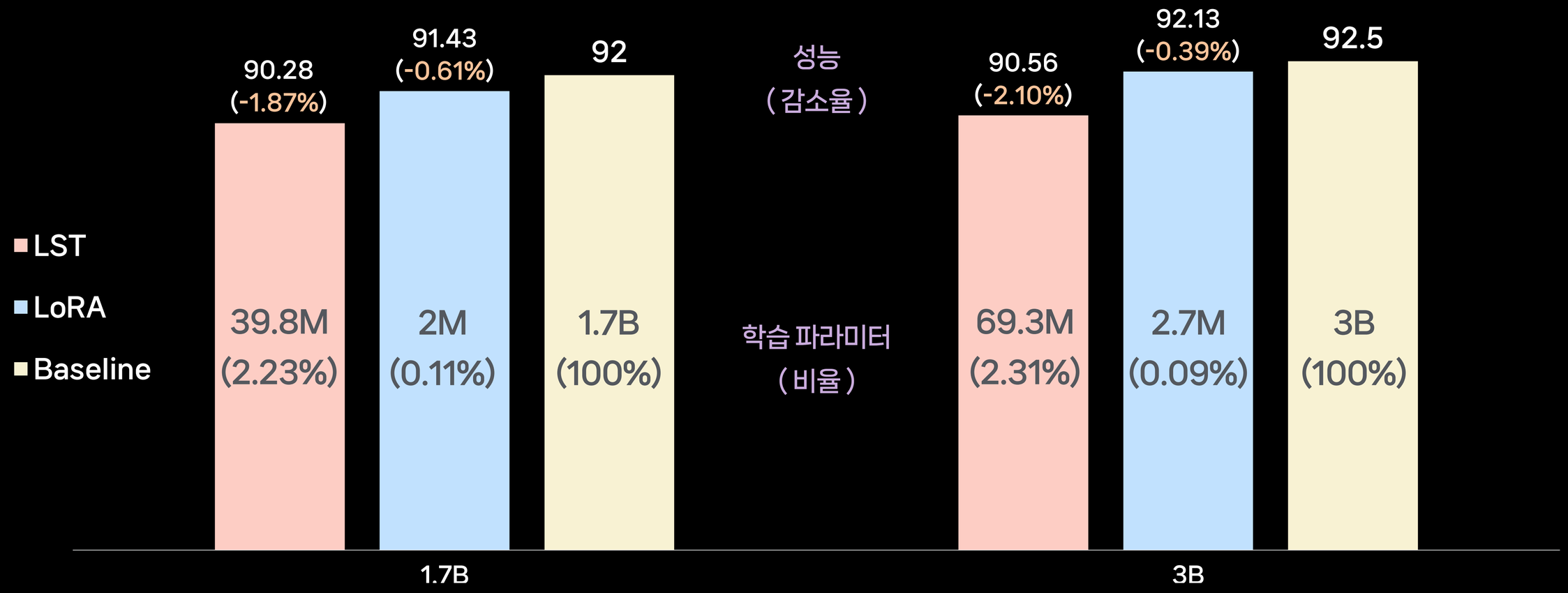
LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS (Hu et al., 2021)
LST: Ladder Side-Tuning for Parameter and Memory Efficient Transfer Learning (Sung et al., 2022)

2.3. 벤치마크

NAVER
DEVIEW
2023

Large-scale Language Model 의 효율적인 학습 방법

- NLU 성능 평균 (NSMC, YNAT, KLUE-NLI, KLUE-STS)



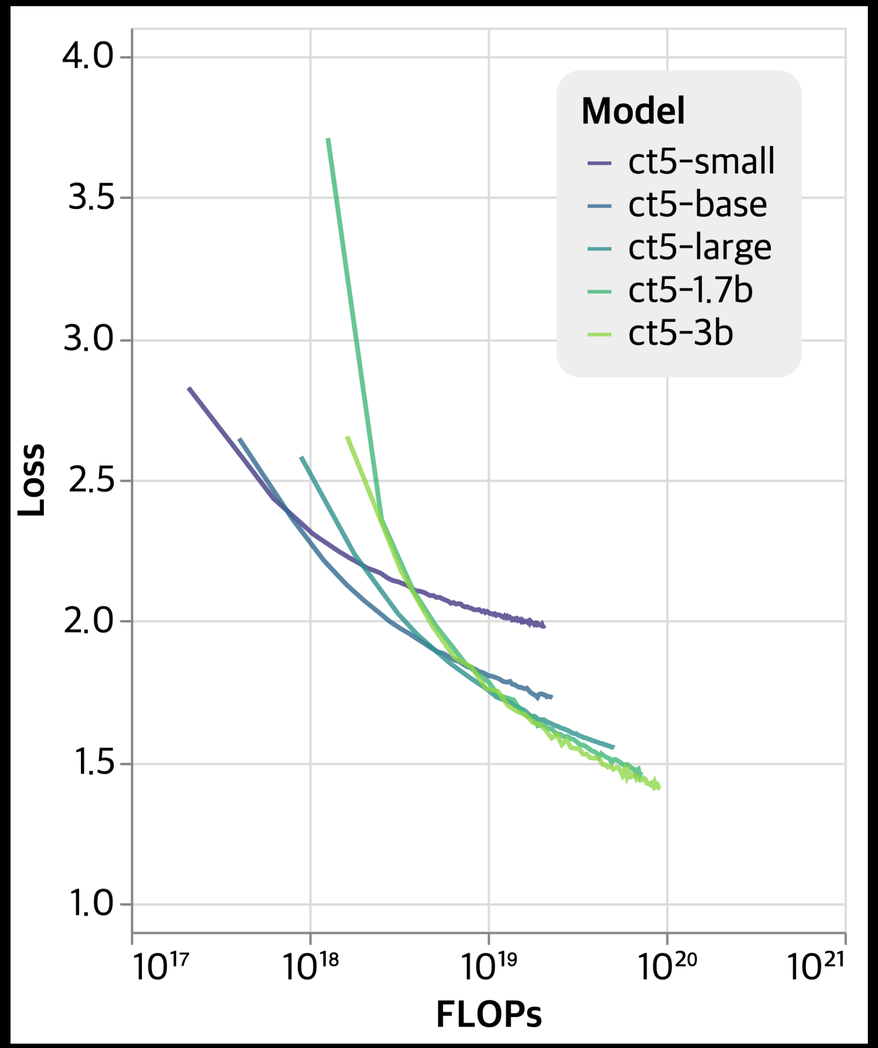
2.4. 연산자원 효율성

NAVER
DEVIEW
2023

Compute-Optimality 분석

- 벤치마크를 통해 모델 사이즈와 성능이 비례함을 확인
- 모델 크기가 커질 수록 도달할 수 있는 Loss 의 값이 작아짐
- 즉, 도달할 수 있는 Loss 와 모델의 성능은 상관관계가 존재

→ 연산 자원이 주어졌을 때 Loss 가 가장 낮은 모델이 가장 효율적인 모델



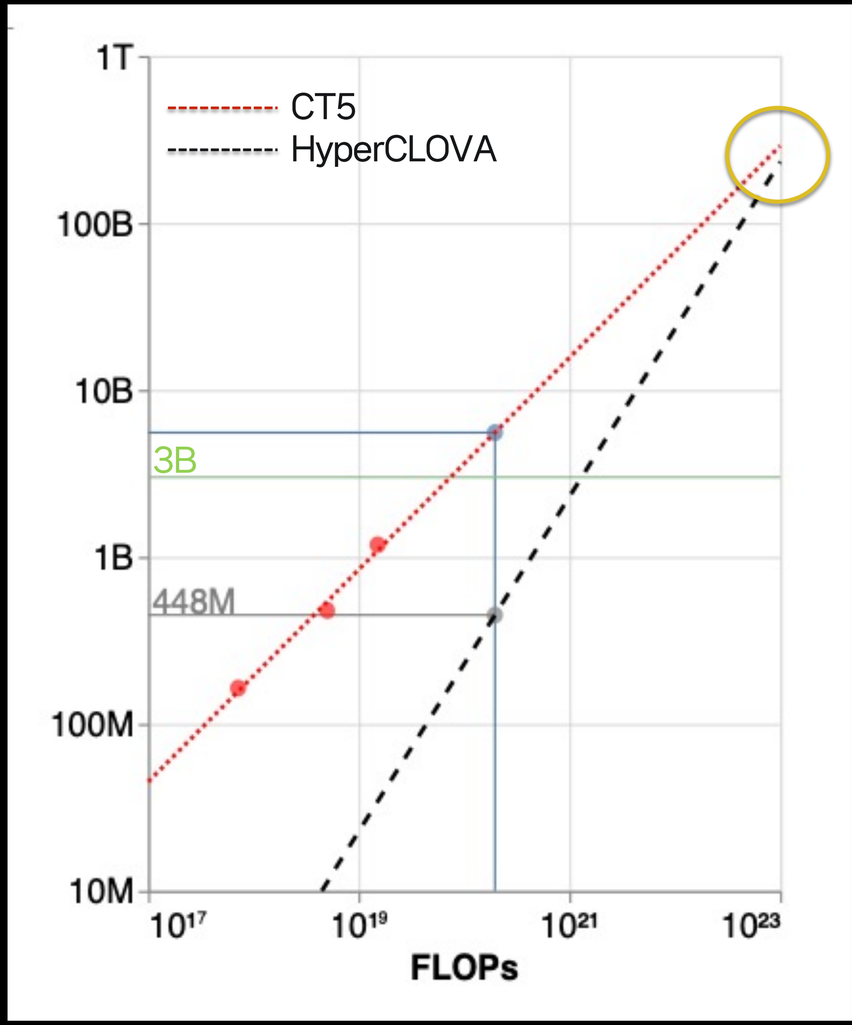
2.4. 연산자원 효율성

NAVER
DEVIEW
2023

Compute-Optimality 분석

- Small, Medium budget에서 Seq2Seq 구조가 효율적으로 모델 학습 가능

- 자연스럽게, Large budget에서는 Decoder only 구조가 유리



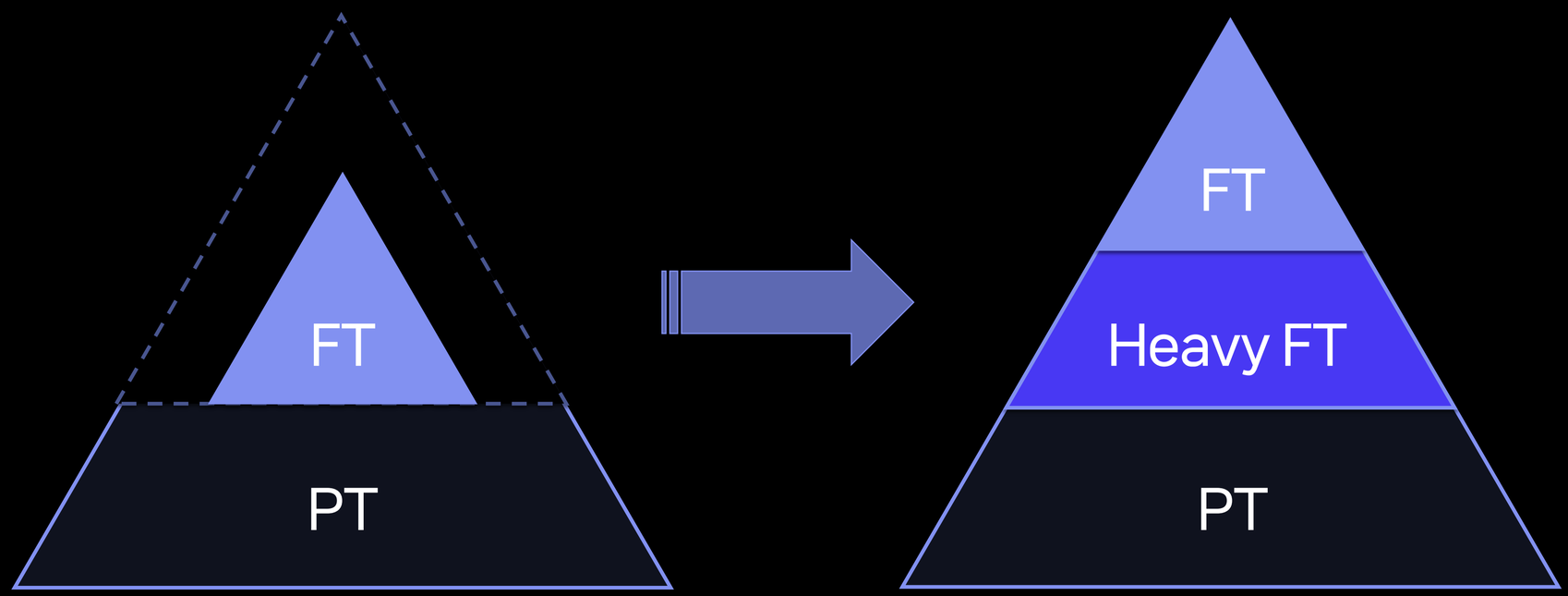
3. DialogCT5:
Domain Adaptation으로
데이터 효율 UP!

3.1. Why DialogCT5?

NAVER
DEVIEW
2023

Heavy Fine-Tuning

- 기존 Pretraining (PT) – Finetuning (FT) 두 단계 학습 패러다임에서 Generalization
- 관련 연구: Domain Adaptation (DialogZoo, ...), Extra Compute (U-PaLM)

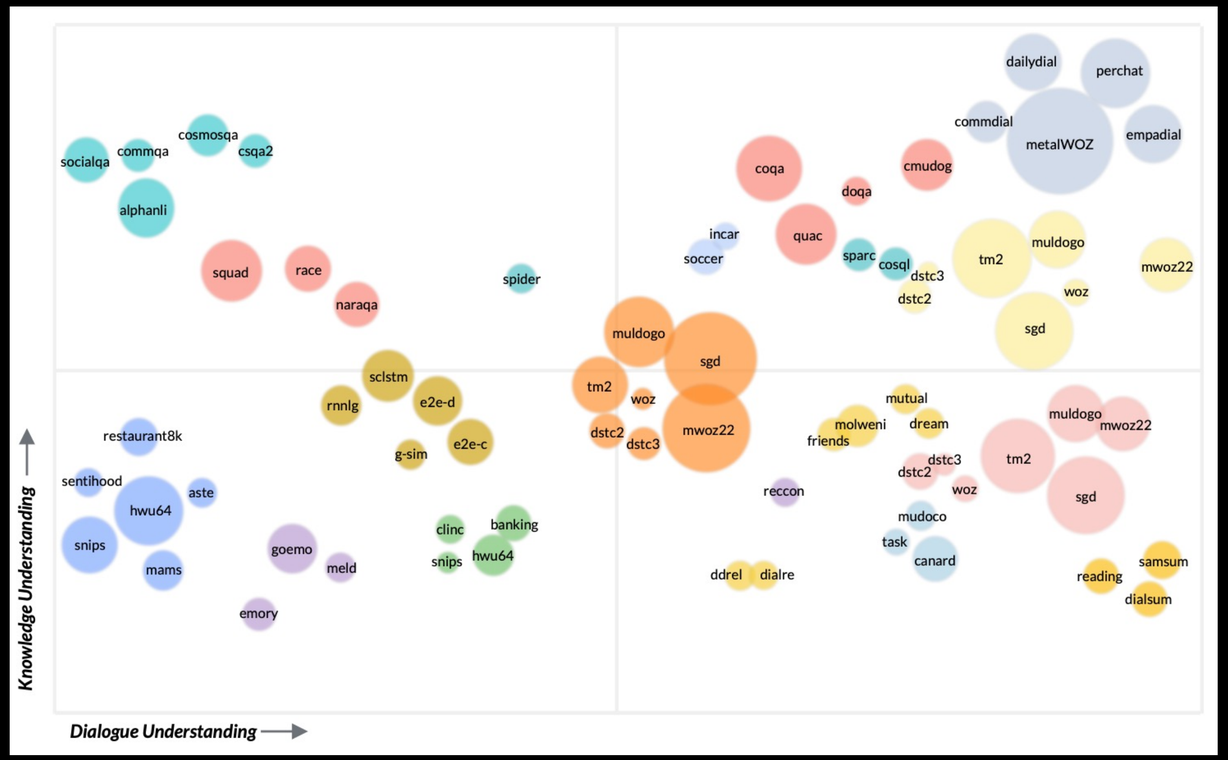


3.1. Why DialogCT5?

NAVER
DEVIEW
2023

Domain Adaptation

- 대량의 대화 데이터셋 기반 멀티태스크 러닝으로 대화 성능 전반 향상



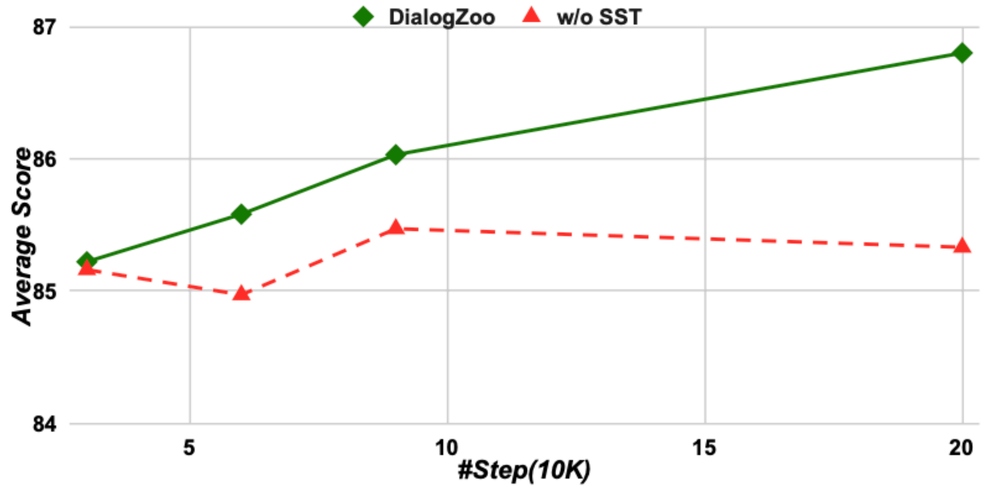


Figure 5: The average scores of DialogZoo models with and without two self-supervised denoising tasks.

DialogZoo (Chen et al., 2022)

3.1. Why DialogCT5?

NAVER
DEVIEW
2023

Extra Compute

- 상대적으로 작은 compute budget의 mixed denoising 추가 학습으로 일반화 성능 향상

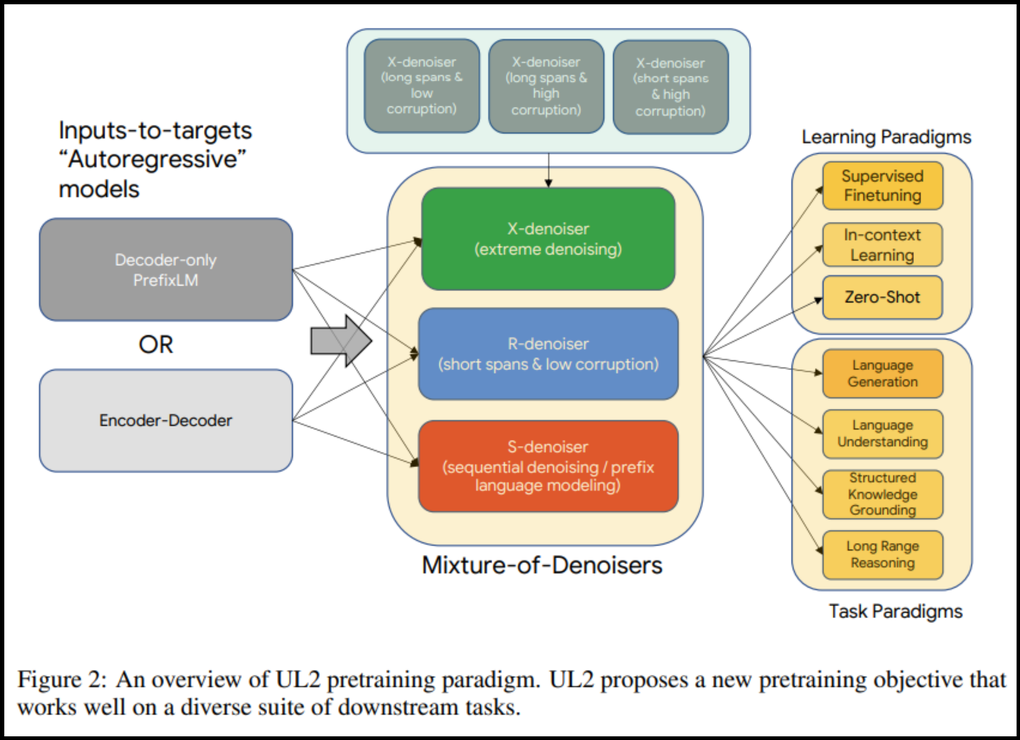
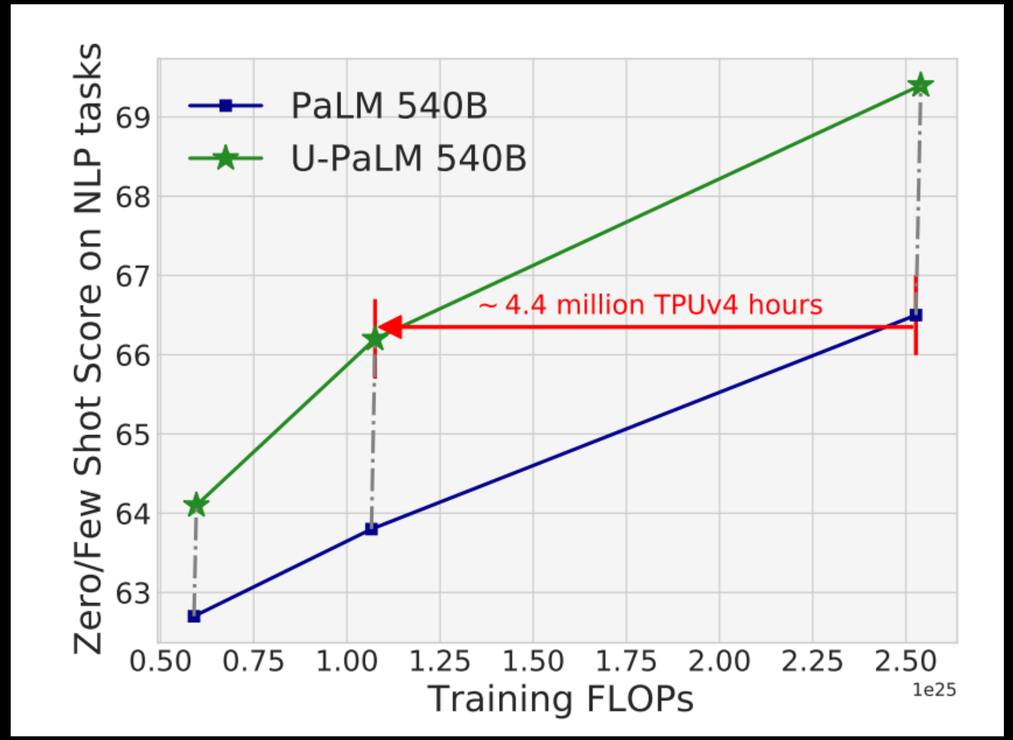


Figure 2: An overview of UL2 pretraining paradigm. UL2 proposes a new pretraining objective that works well on a diverse suite of downstream tasks.



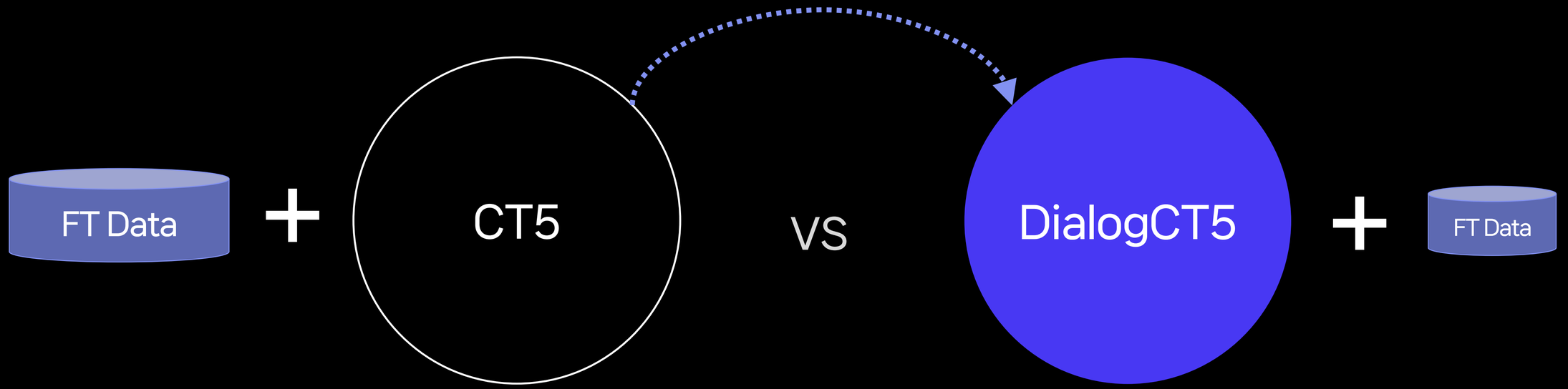
Transcending Scaling Laws with 0.1% Extra Compute [Google'22]

3.1. Why DialogCT5?

NAVER
DEVIEW
2023

CT5 towards Data efficiency and Dialog-domain adaptation

→ 대화 기반 태스크에서 적은 in-domain 데이터로 좋은 성능 내기



3.1. Why DialogCT5?

NAVER
DEVIEW
2023

CT5 towards Data efficiency and Dialog-domain adaptation

Dialogue infilling
Dialogue response generation
Dialogue summarization
...

3.1. Why DialogCT5?

NAVER
DEVIEW
2023

CT5 towards Data efficiency and Dialog-domain adaptation

Dialogue infilling
Dialogue response generation
Dialogue summarization
...

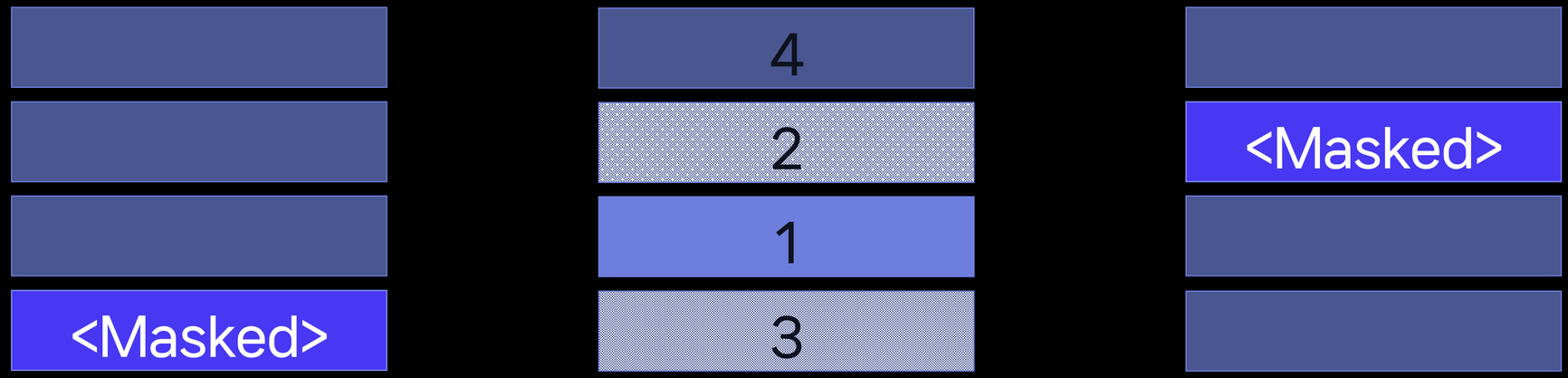
Scarce data performance

3.2. 학습 세팅

NAVER
DEVIEW
2023

대화 특화 사전학습 Objective

- 다음 발화 예측 (Next utterance prediction)
- 발화 순서 맞추기 (Utterance order reconstruction)
- 단일 발화 마스킹 (Single utterance masking)



3.2. 학습 세팅

NAVER
DEVIEW
2023

타겟 학습 Objective

- Partial Utterance

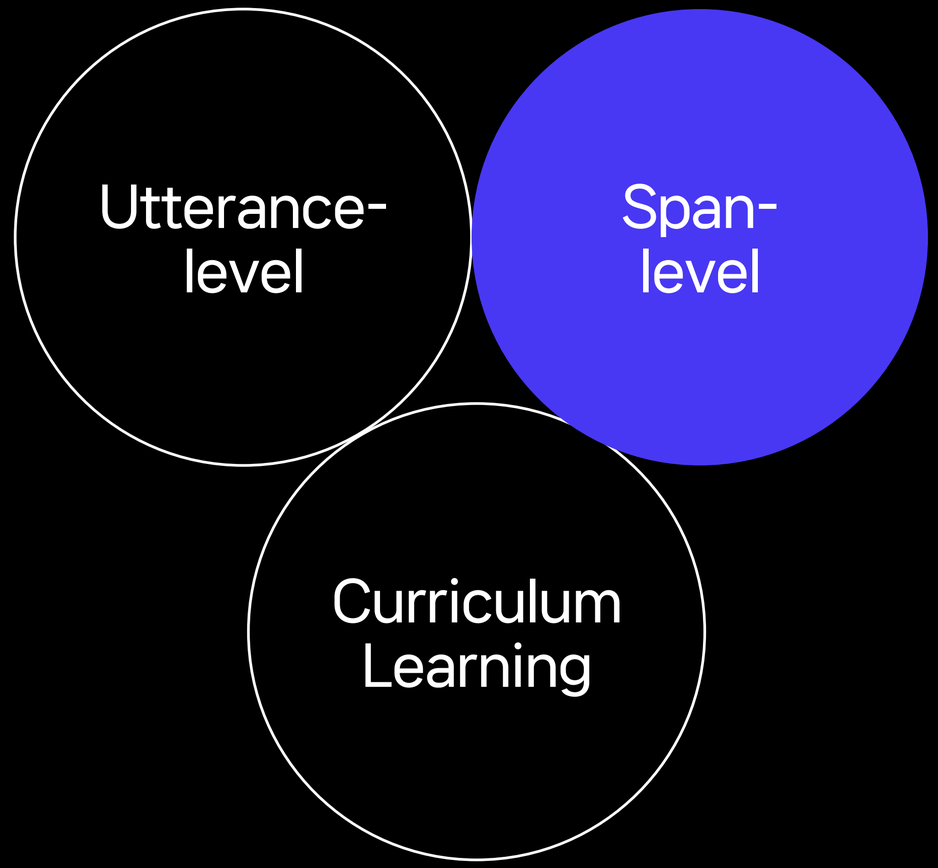
Partial Utterance Masking

A: 에고. <M0> 사고를 치세요?

B: 꼼꼼하지 않아서 자꾸 뭔가를 놓쳐. <M1>를 않나 엑셀 수식을 다 <M2> 않나 ㅋㅋ

A: 아이쿠야. 사소한 실수는 할 수 있지만 그건 정말 <M3>아요.

...

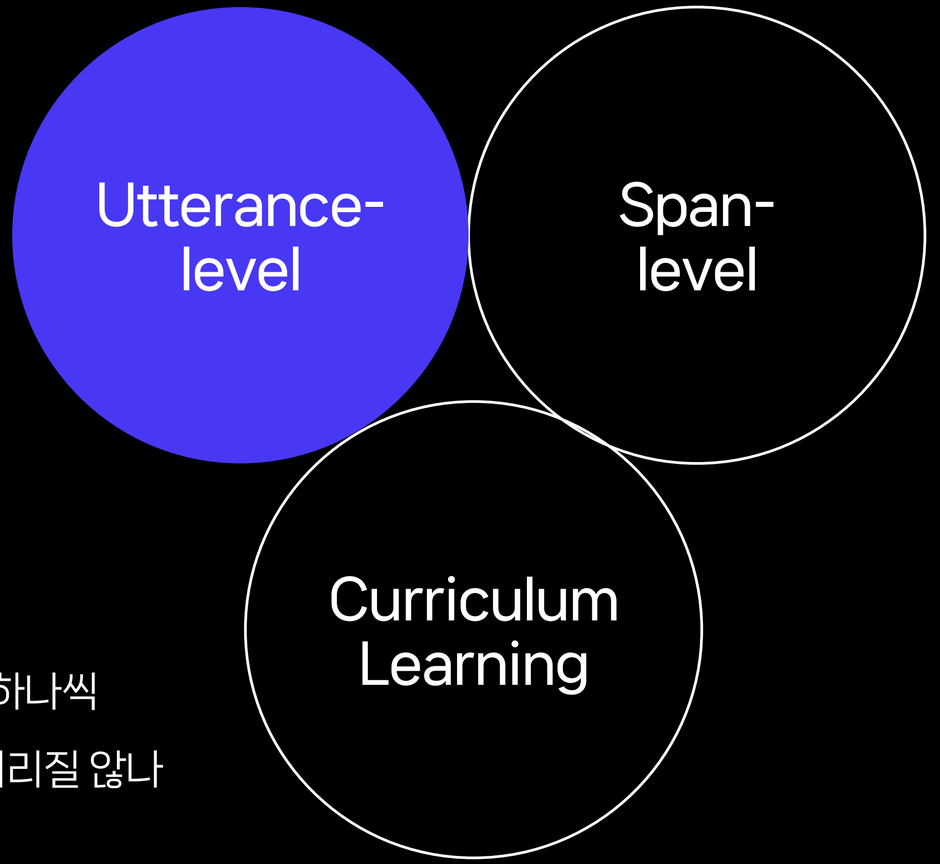


3.2. 학습 세팅

NAVER
DEVIEW
2023

타겟 학습 Objective

- Single Utterance
- Multiple Utterance ✓



Single Utterance Masking

A: 에고, 보통 어떤 사고를 치세요?

B: <M0>

A: 아이쿠야. 사소한 실수는 할 수 있지만 그건 정말 심각한 문제잖아요.

...

Multiple Utterance Masking

A: <M0>

B: 꼼꼼하지 않아서 자꾸 뭔가를 놓쳐. 하나씩 빠뜨리지를 않나 엑셀 수식을 다 바꿔버리질 않나 ㅋㅋ

A: <M1>

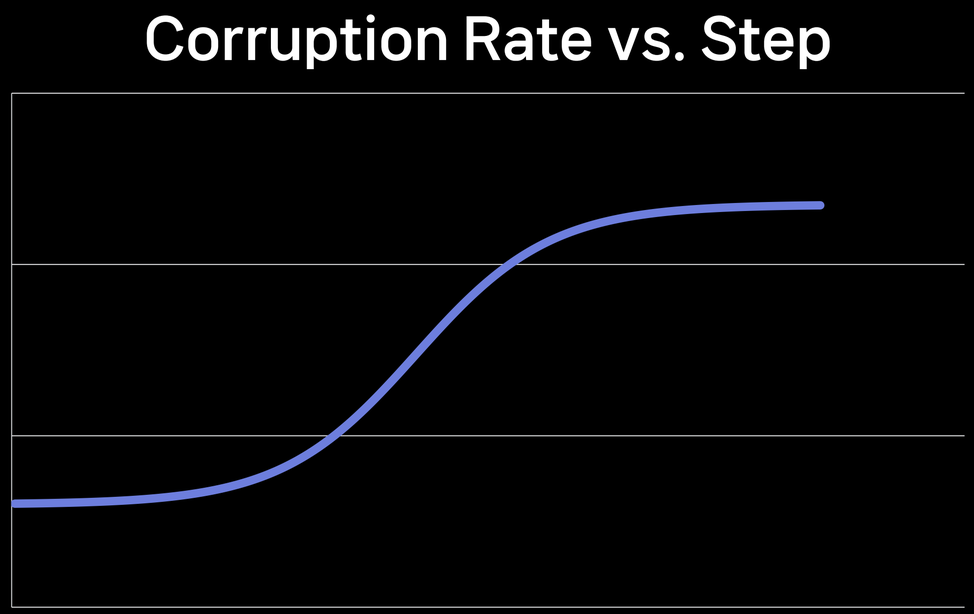
...

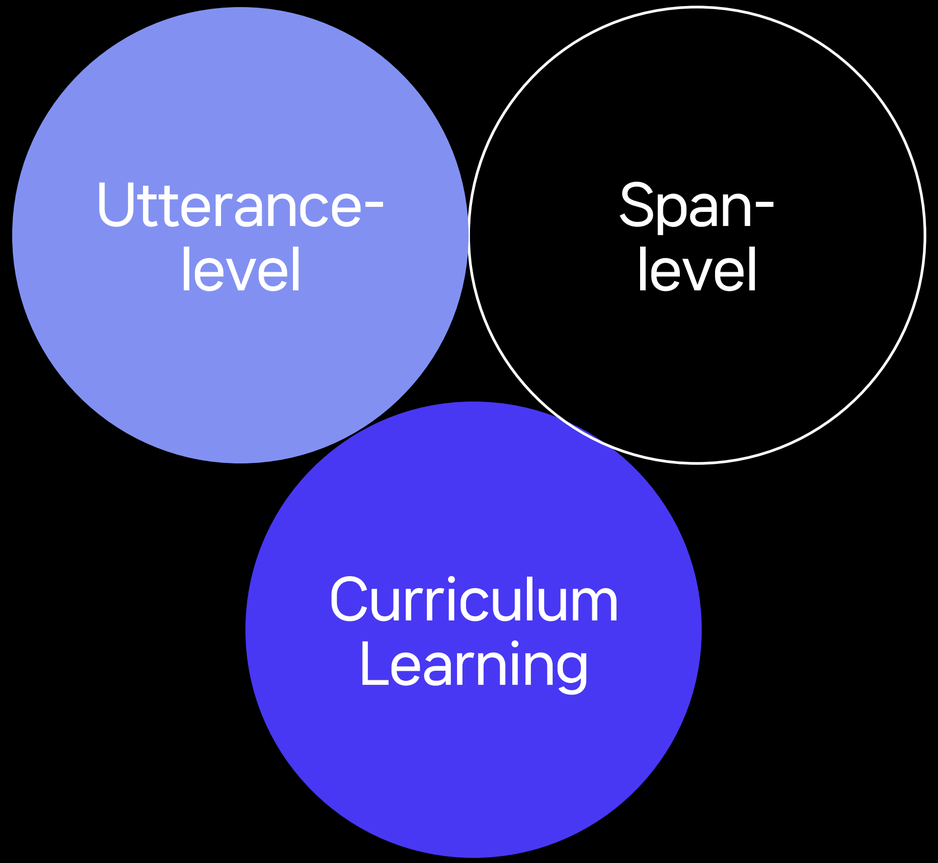
3.2. 학습 세팅

NAVER
DEVIEW
2023

타겟 학습 Objective

- Curriculum learning with corruption rate ✓





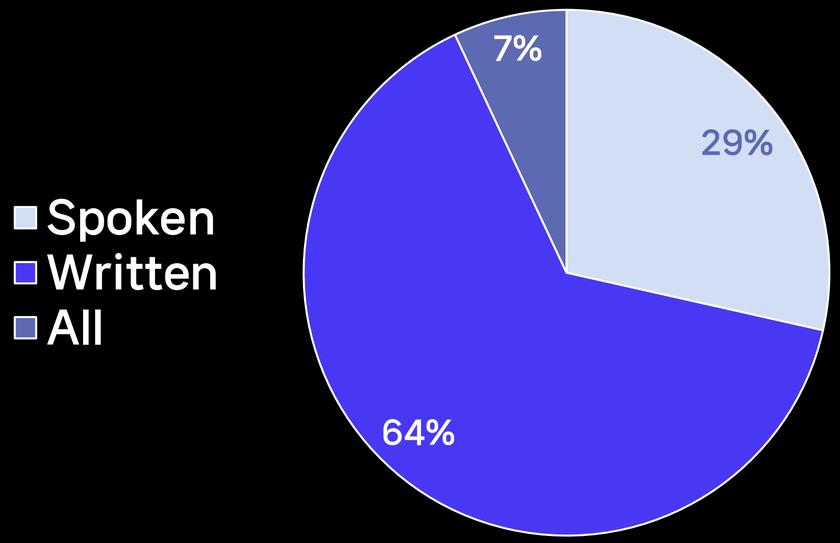
3.2. 학습 세팅

NAVER
DEVIEW
2023

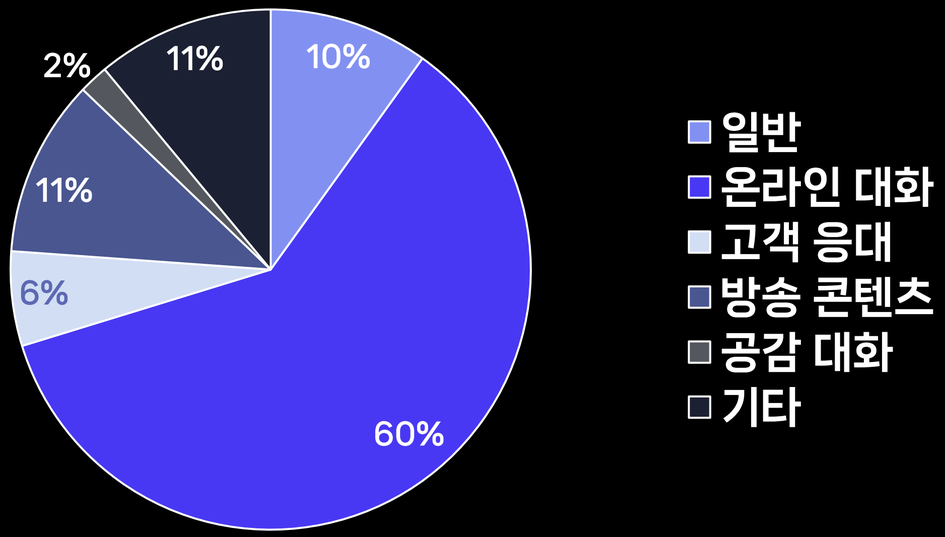
어떤 데이터로 학습했나?

- 대화 세션수: 3.3M Train / 10k Test (약 1B Tokens)
- 턴수: 50.2M

대화 타입



대화 코퍼스 종류



3.2. 학습 세팅

NAVER
DEVIEW
2023

모델 학습 파라미터

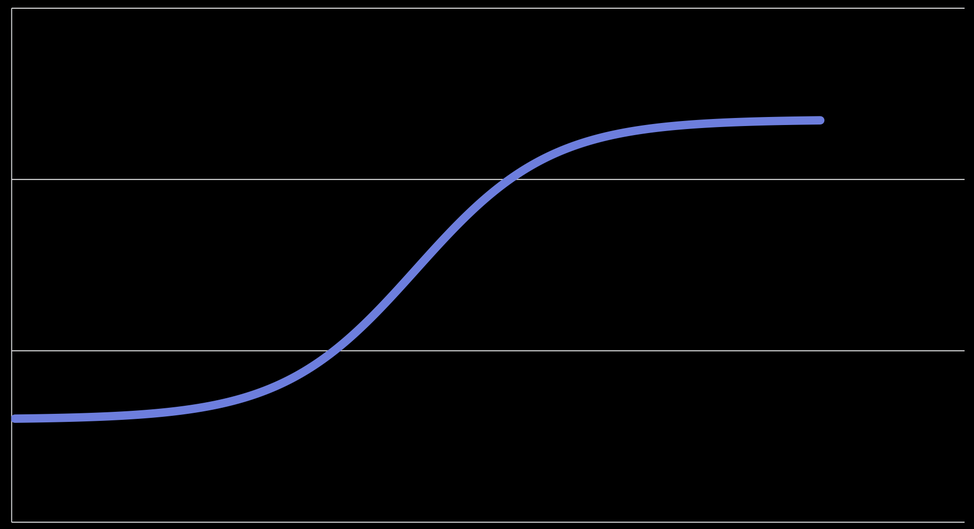
항목	설명
Backbone	Vanilla CT5 (small/base/large)
Epoch	5
Global Batch Size	64
Dropout	X
Optimizer	Adafactor
Learning Rate	5e-5
bf16	True

3.2. 학습 세팅

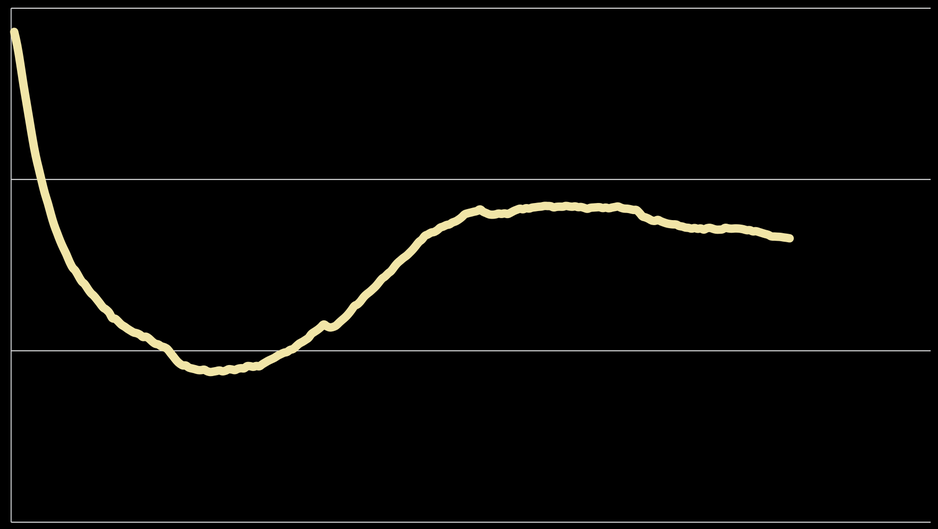
NAVER
DEVIEW
2023

Training Loss Graph w/ Curriculum Learning

Corruption Rate vs. Step



Training Loss vs. Step



3.3. 벤치마크

NAVER
DEVIEW
2023

AI Hub 데이터세트

- Task-oriented dialogues (ToD): 목적 지향 대화
- Open-domain dialogues (ODD): 자유 주제 대화
- Script summarization (Script-Summ): 방송 대본 요약

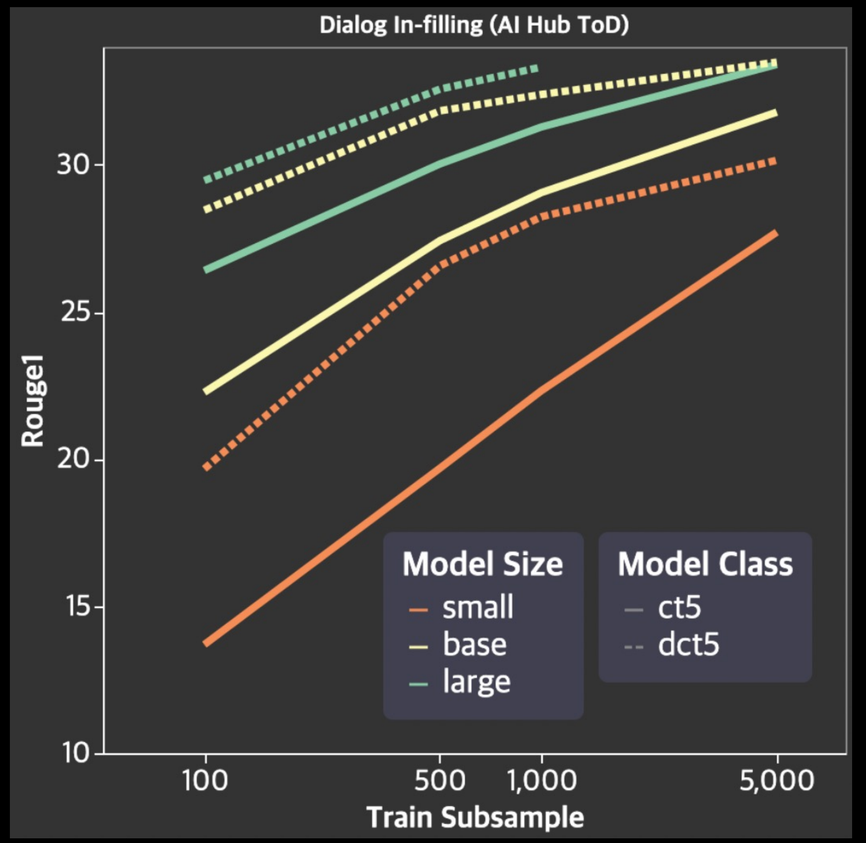
Task	설명	데이터세트
Dialog In-filling	대화 중간 발화 채우기	AI Hub ToD
		AI Hub ODD
Dialog Response Generation	대화 생성	AI Hub ToD
		AI Hub ODD
Dialog Summarization	대화 요약	AI Hub Script-Summ

3.3. 벤치마크

NAVER
DEVIEW
2023

데이터 효율성 실험

- DialogCT5가 CT5와 같은 성능을 내는 데에 약 10배의 데이터 효율까지 확인

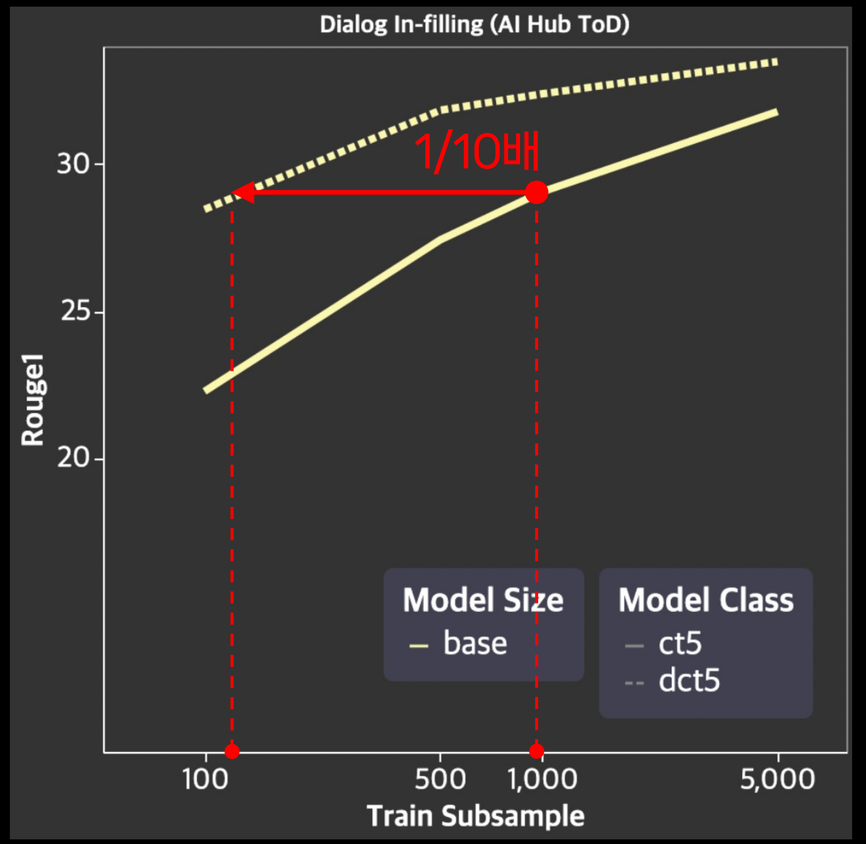


3.3. 벤치마크

NAVER
DEVIEW
2023

데이터 효율성 실험

- DialogCT5가 CT5와 같은 성능을 내는 데에 약 10배의 데이터 효율까지 확인

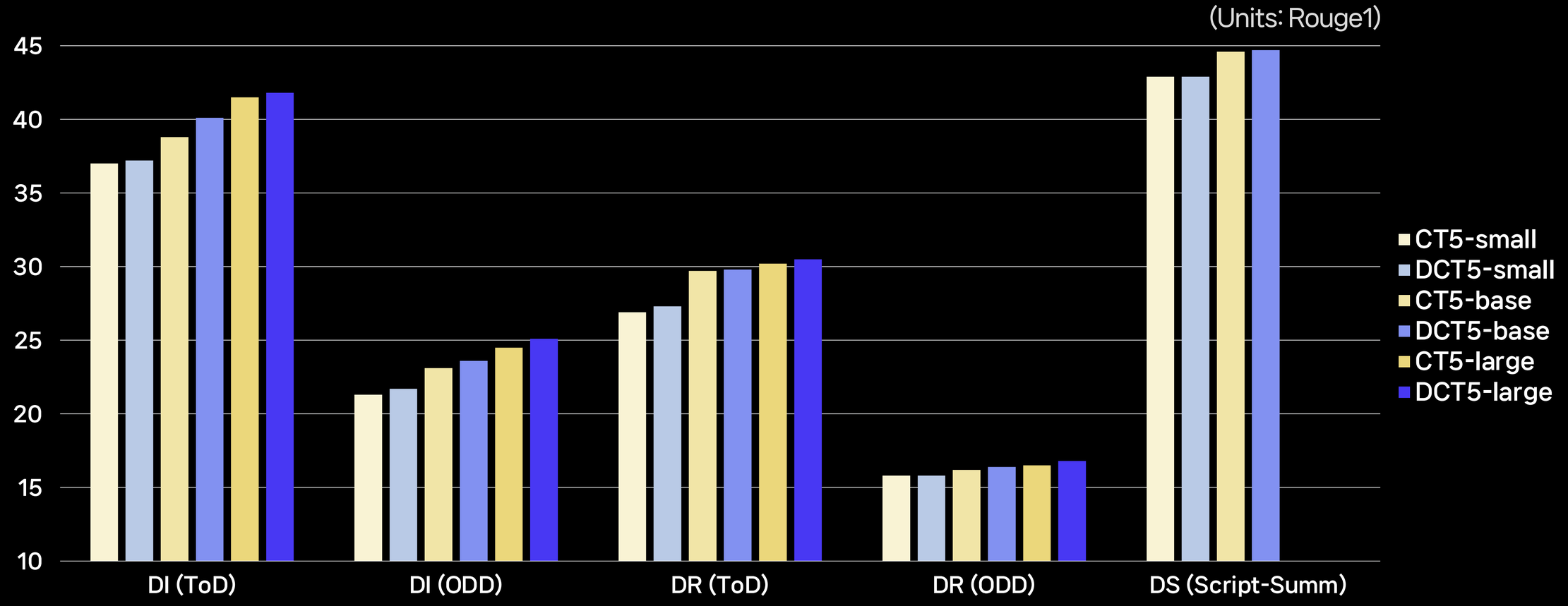


3.3. 벤치마크

NAVER
DEVIEW
2023

전체 데이터 fine tuning 실험

- 전체적으로 CT5보다 높은 점수 달성

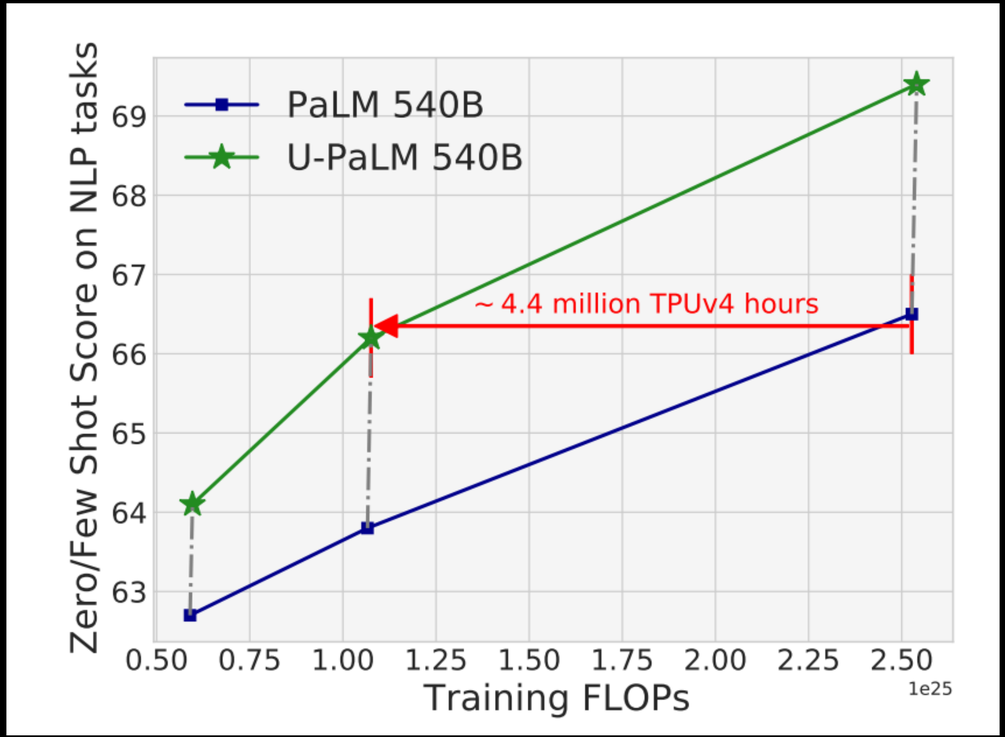


3.4. 정성적 평가 및 Discussion

NAVER
DEVIEW
2023

Heavy Fine-Tuning

- Little Extra Compute for Higher Data Efficiency



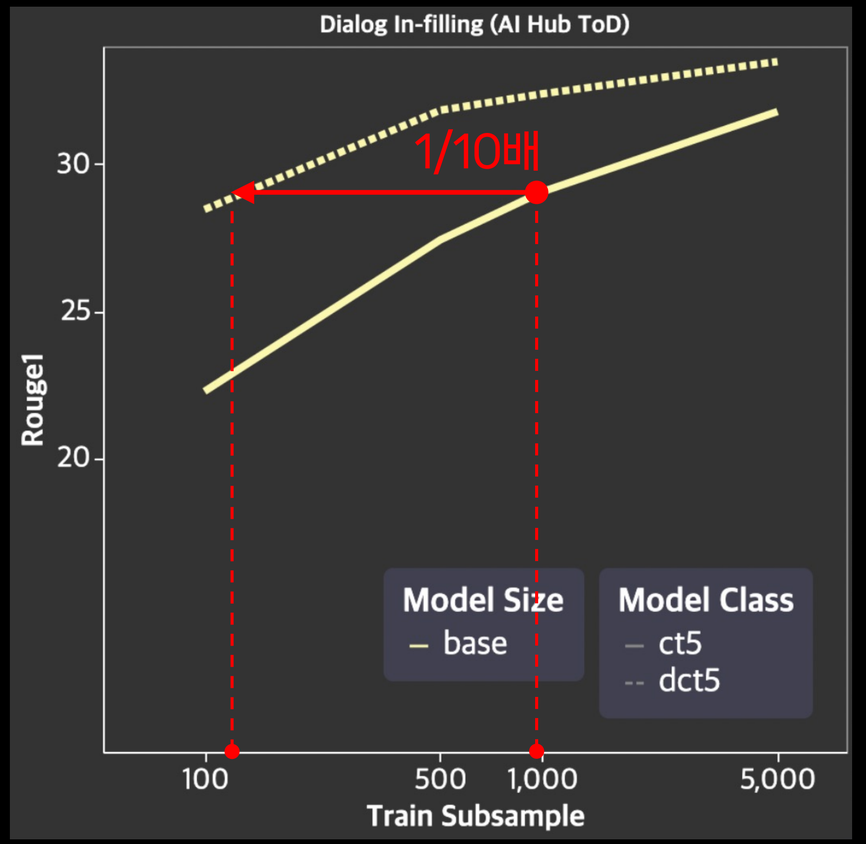
Transcending Scaling Laws with 0.1% Extra Compute [Google'22]

3.4. 정성적 평가 및 Discussion

NAVER
DEVIEW
2023

Heavy Fine-Tuning

- 상대적으로 작은 대화 데이터로 추가 학습 → 적은 데이터 환경에서의 성능 대폭 향상



4. 결론

NAVER
DEVIEW
2023

- 언어모델 세계에서 Seq2Seq이란: 이해 능력, 생성 능력, 효율 모두 겸비한 알찬 접근법
- 강력한 한국어 Seq2Seq 모델인 CT5 공개 및 학습 경험담 소개
- 대화 관련 문제에 완벽히 적응하여 우월한 데이터 활용 능력을 보여준 DialogCT5 공개
- Compute-Optimality 분석을 통해 학습 자원이 GPT급 미만일 경우 Seq2Seq의 장점 확인

감사합니다!