RHO-1: Not All Tokens Are What You Need

24.01.10

박희선

❖ 연구 배경

- Next-token prediction에 있어서 모델 파라미터와 데이터 사이즈를 키우는 것이 성능 향상의 방법이었음
- 이때 데이터는 필터링 작업을 걸쳐서 만들어짐에도 노이즈는 여전히 존재를 함
- 즉, 모든 토큰들이 응용 작업에 있어서 전부 이상적인 분포를 따른다고 할 수 없음
- 그러나 기존의 모델은 **모든 토큰들에 있어서** 동일하게 loss를 계산하여 반영함

- 따라서 **필요없는 토큰들은 미리 제거시켜** 모델을 훈련시킨다는 것이 논문의 주아이디어

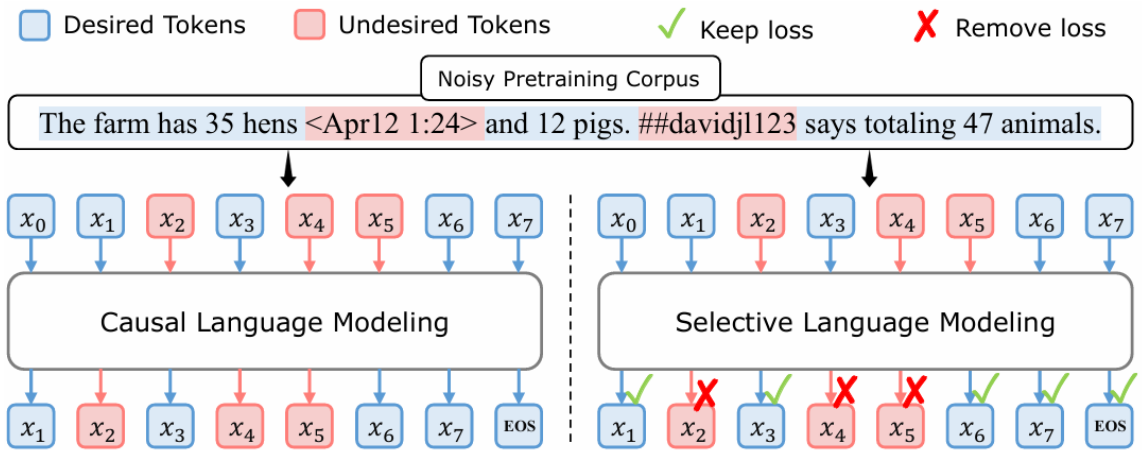
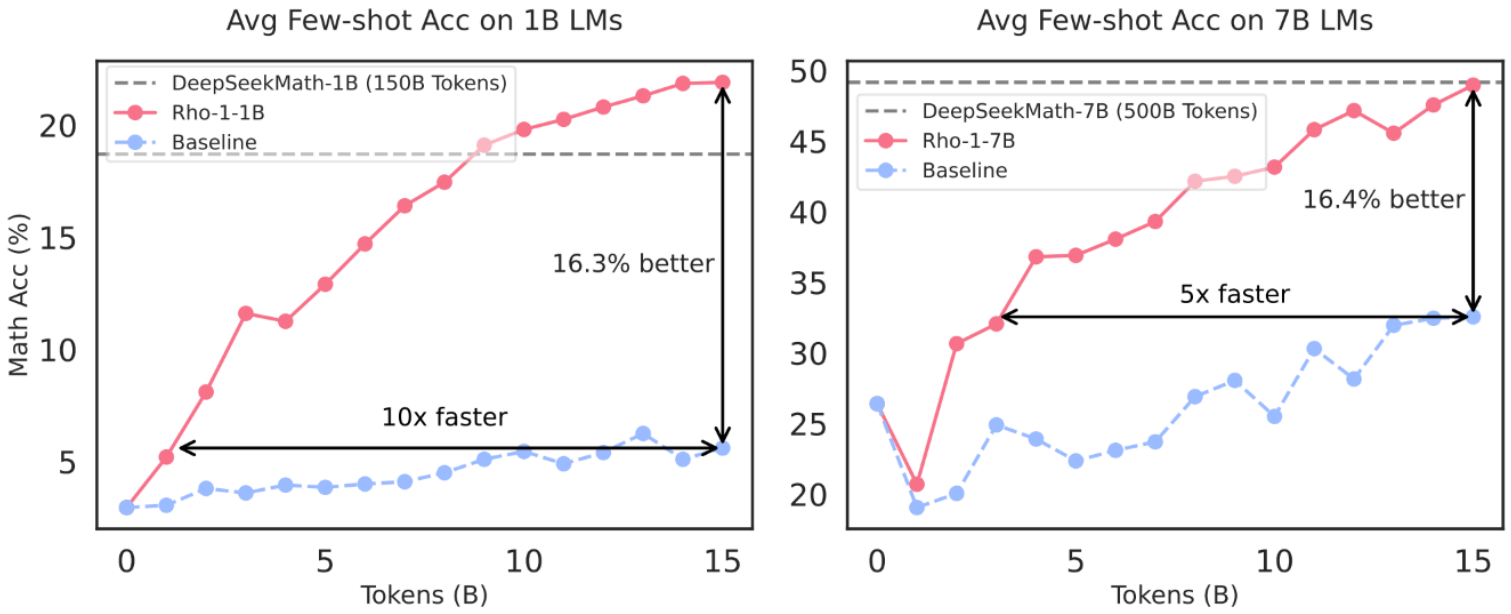


Figure 2: **Upper:** Even an extensively filtered pretraining corpus contains token-level noise. **Left:** Previous Causal Language Modeling (CLM) trains on all tokens. **Right:** Our proposed Selective Language Modeling (SLM) selectively applies loss on those useful and clean tokens.

I Introduction

- 1B 모델에서는 baseline model에 비해서 10배 빠르고, 16.3% 더 나은 성능을 보임
- 7B 모델에서는 baseline model에 비해서 5배 빠르고, 16.4% 더 나은 성능을 보임



I Introduction

❖ 연구 배경

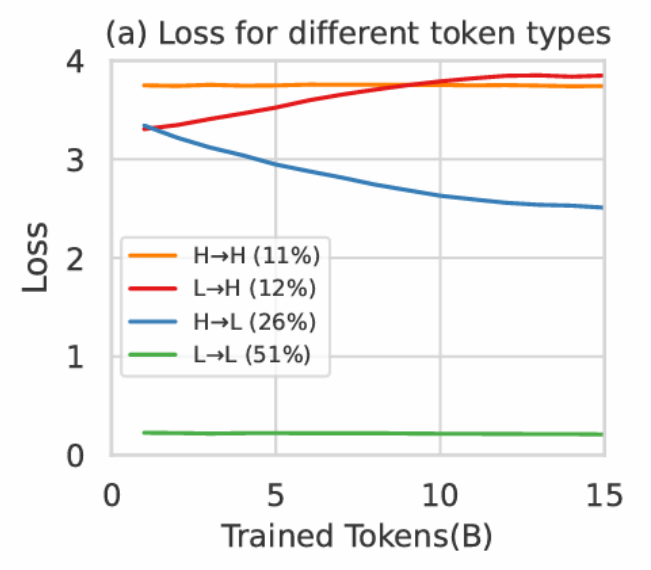
- 기존의 data selection은 Sample level selection과 Token level selection으로 나뉨
- Sample level selection:
  - Online Batch Selection for Faster Training of Neural Networks (2015)
    loss가 큰 batch를 선택하여 훈련

- Token level selection:
  - Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
    랜덤으로 token을 마스킹 처리한 후 해당 단어를 예측
  - Revisiting Token Dropping Strategy in Efficient BERT Pretraining (2023)
    일정비율로 의미기여도가 낮은 토큰을 drop

Selective Language Modeling

❖ Not all tokens are equal: Training dynamics of token loss

- 우선 사전 훈련을 통해 훈련 시 토큰의 loss를 추적함
- OpenWebMath의 15B 토큰 훈련에서 1B마다 체크 포인트를 생성한 후, 약 320,000개의 토큰으로 구성된 검증데이터로 loss를 계산함
- 그 결과 4가지 종류로 나눌 수 있음:



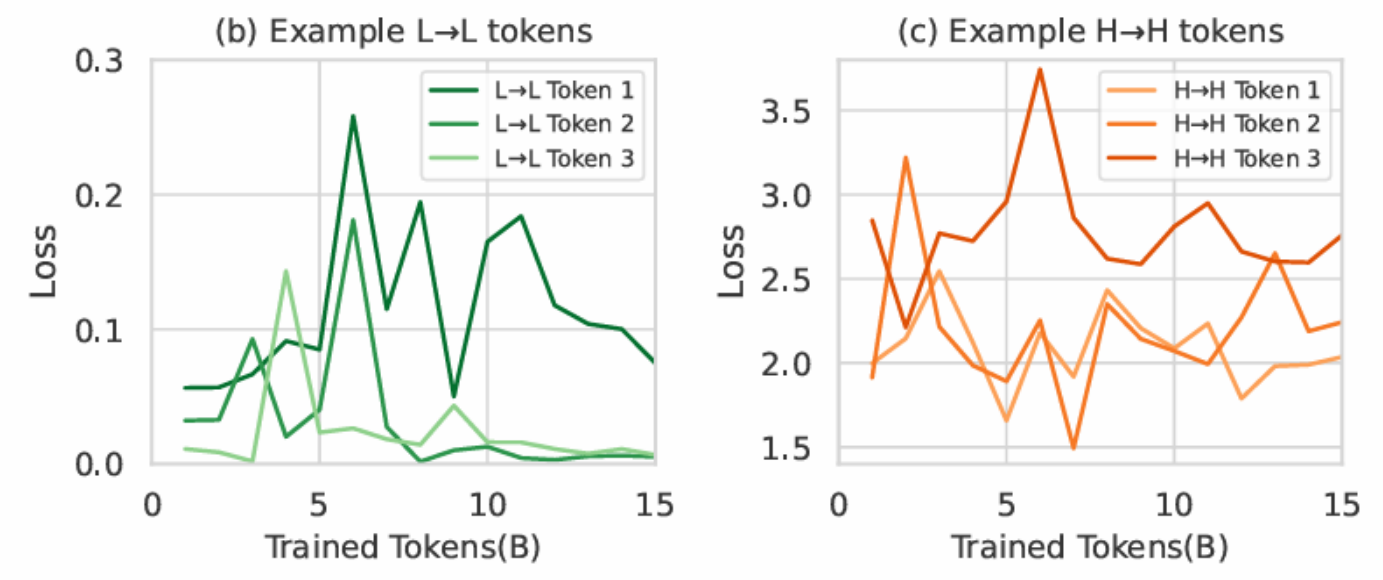
- H→H (11%): 일관되게 높은 손실을 기록한 토큰
- L→H (12%): 손실이 증가된 토큰
- H→L (26%): 손실이 감소된 토큰
- L→L (51%): 일관되게 낮은 손실을 기록한 토큰, 훈련이 됨

단, 이때 '→L'는 학습이 되었음을 의미

Selective Language Modeling

❖ Not all tokens are equal: Training dynamics of token loss

- 다음 그림을 통해 토큰 단위의 loss는 전체 loss와 같이 지속적으로 감소하며 수렴하지 않음을 보여줌



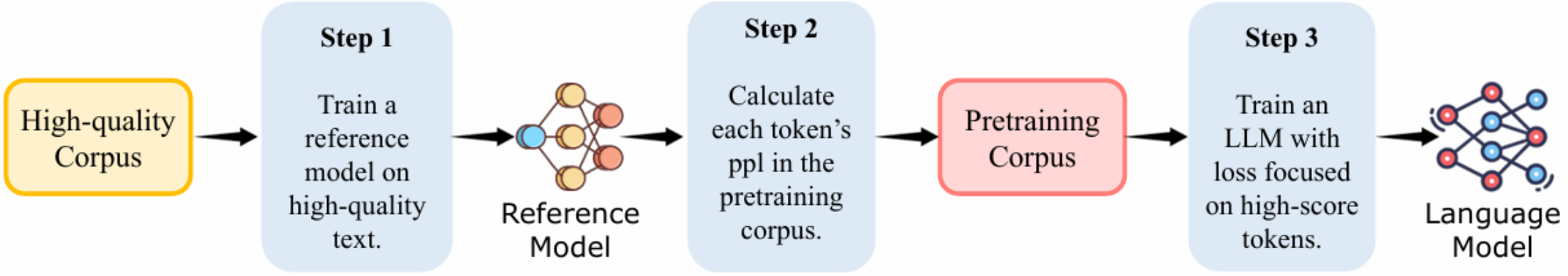
- 모델 훈련에 있어서 **적절한 토큰만을 선택하게** 된다면 안정적인 모델 훈련, 데이터 효율성을 높일 수 있음

Selective Language Modeling

❖ Selective language modeling

- 전체 과정:

1. 선별된 고품질의 데이터를 활용하여 reference model을 훈련시킨다.
2. 사전 훈련된 corpus(말뭉치)에서 각 **토큰의 손실**을 평가한다
3. 계산된 손실 중 **높은 손실을 기록한 토큰을 선택**하여 언어모델을 훈련시킨다



Selective Language Modeling

❖ Selective language modeling

- 기존의 LLM의 loss 계산 방법: cross-entropy loss

$$\mathcal{L}_{\text{CLM}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log P(x_i|x_{<i};\theta)$$

- 본 논문의 loss 계산 방법:
  1. reference model에서 cross-entropy를 사용해 모델 훈련

$$\mathcal{L}_{\text{RM}}(x_i) = -\log P(x_i|x_{<i})$$

  2. 현재 학습 모델과 reference model의 loss차이를 계산

$$\mathcal{L}_{\Delta}(x_i) = \mathcal{L}_{\theta}(x_i) - \mathcal{L}_{\text{RM}}(x_i)$$

  3. loss 차이가 큰 토큰 상위 k%를 선택해 현재 학습 모델의 loss 계산

$$\mathcal{L}_{\text{SLM}}(\theta) = -\frac{1}{N * k\%}\sum_{i=1}^{N} I_{k\%}(x_i)\cdot\log P(x_i|x_{<i};\theta)$$

상위 k%의 토큰 개수 — 토큰 선택 여부

$$I_{k\%}(x_i) = \begin{cases} 1 & \text{if } x_i \text{ ranks in the top } k\% \text{ by } S(x_i) \\ 0 & \text{otherwise} \end{cases}$$

Selective Language Modeling

After Training 0% Checkpoint

Item Type: Journal Article Copyright of this article belongs to Elsevier. Division of Mechanical Sciences > Mechanical Engineering 28 May 2007 19 Sep 2010 04:36 http://eprints.iisc.ernet.in/id/eprint/10277 # Question #8de97 \n \n Dec 10, 2016 \n \n That is not an identity. \n \n #### Explanation: \n \n Recall that \n \n ${\cot}^{2} x + 1 = {\csc}^{2} x$. \n \n So, we can write \n \n $\frac{1 - {\csc}^{2} x}{\csc} ^ 2 x = \frac{1 - \left({\cot}^{2} x + 1\right)}{\csc} ^ 2 x$ \n \n $= {\cot}^{2} \frac{x}{\csc} ^ 2 x$ \n \n Recall also that $\cot x = \cos \frac{x}{\sin} x$ and $\csc x = \frac{1}{\sin} x$. \n \n This allows us to continue \n \n $= \frac{{\cos}^{2} \frac{x}{\sin} ^ 2 x}{\frac{1}{\sin} ^ 2 x}$ \n \n $= {\cos}^{2} \frac{x}{\sin} ^ 2 x \cdot {\sin}^{2} \frac{x}{1}$ \n \n $= {\cos}^{2} x$ \n \n Which is not identically $\cos x$. \n \n (${\cos}^{2} x = \cos x$ only when $\cos x = 1$ or 0) \n \n Dec 10, 2016 \n \n No. It is equal to ${\sin}^{2} x - 1$. \n \n #### Explanation: \n \n If we have $\frac{1 - x}{x}$, we can write it as $\frac{1}{x} - \frac{x}{x}$. \n \n The same way, $\frac{1 - {\csc}^{2} x}{\csc} ^ 2$ can be written as $\frac{1}{\csc} ^ 2 x - \frac{{\csc}^{2} x}{{\csc}^{2} x}$. \n \n This is equal to ${\sin}^{2} x - 1$. SMS newsy item created by Hannah Bryant at Wed 25 May 2022 1227 \n Type: Seminar \n Distribution: World \n Expiry: 31 May 2022 \n Calendar1: 31 May 2022 1500-1600 \n CalLoc1: Quad S224 & via Zoom \n CalTitle1: SMRI 'What is ...a virtual knot?' Hans Boden (McMaster University) \n Auth: hannahb@w1d4n6z2.staff.sydney.edu.au (hbry8683)

After Training 33% Checkpoint

Item Type: Journal Article Copyright of this article belongs to Elsevier. Division of Mechanical Sciences > Mechanical Engineering 28 May 2007 19 Sep 2010 04:36 http://eprints.iisc.ernet.in/id/eprint/10277 # Question #8de97 \n \n Dec 10, 2016 \n \n That is not an identity. \n \n #### Explanation: \n \n Recall that \n \n ${\cot}^{2} x + 1 = {\csc}^{2} x$. \n \n So, we can write \n \n $\frac{1 - {\csc}^{2} x}{\csc} ^ 2 x = \frac{1 - \left({\cot}^{2} x + 1\right)}{\csc} ^ 2 x$ \n \n $= {\cot}^{2} \frac{x}{\csc} ^ 2 x$ \n \n Recall also that $\cot x = \cos \frac{x}{\sin} x$ and $\csc x = \frac{1}{\sin} x$. \n \n This allows us to continue \n \n $= \frac{{\cos}^{2} \frac{x}{\sin} ^ 2 x}{\frac{1}{\sin} ^ 2 x}$ \n \n $= {\cos}^{2} \frac{x}{\sin} ^ 2 x \cdot {\sin}^{2} \frac{x}{1}$ \n \n $= {\cos}^{2} x$ \n \n Which is not identically $\cos x$. \n \n (${\cos}^{2} x = \cos x$ only when $\cos x = 1$ or 0) \n \n Dec 10, 2016 \n \n No. It is equal to ${\sin}^{2} x - 1$. \n \n #### Explanation: \n \n If we have $\frac{1 - x}{x}$, we can write it as $\frac{1}{x} - \frac{x}{x}$. \n \n The same way, $\frac{1 - {\csc}^{2} x}{\csc} ^ 2$ can be written as $\frac{1}{\csc} ^ 2 x - \frac{{\csc}^{2} x}{{\csc}^{2} x}$. \n \n This is equal to ${\sin}^{2} x - 1$. SMS newsy item created by Hannah Bryant at Wed 25 May 2022 1227 \n Type: Seminar \n Distribution: World \n Expiry: 31 May 2022 \n Calendar1: 31 May 2022 1500-1600 \n CalLoc1: Quad S224 & via Zoom \n CalTitle1: SMRI 'What is ...a virtual knot?' Hans Boden (McMaster University) \n Auth: hannahb@w1d4n6z2.staff.sydney.edu.au (hbry8683)

After Training 66% Checkpoint

Item Type: Journal Article Copyright of this article belongs to Elsevier. Division of Mechanical Sciences > Mechanical Engineering 28 May 2007 19 Sep 2010 04:36 http://eprints.iisc.ernet.in/id/eprint/10277 # Question #8de97 \n \n Dec 10, 2016 \n \n That is not an identity. \n \n #### Explanation: \n \n Recall that \n \n ${\cot}^{2} x + 1 = {\csc}^{2} x$. \n \n So, we can write \n \n $\frac{1 - {\csc}^{2} x}{\csc} ^ 2 x = \frac{1 - \left({\cot}^{2} x + 1\right)}{\csc} ^ 2 x$ \n \n $= {\cot}^{2} \frac{x}{\csc} ^ 2 x$ \n \n Recall also that $\cot x = \cos \frac{x}{\sin} x$ and $\csc x = \frac{1}{\sin} x$. \n \n This allows us to continue \n \n $= \frac{{\cos}^{2} \frac{x}{\sin} ^ 2 x}{\frac{1}{\sin} ^ 2 x}$ \n \n $= {\cos}^{2} \frac{x}{\sin} ^ 2 x \cdot {\sin}^{2} \frac{x}{1}$ \n \n $= {\cos}^{2} x$ \n \n Which is not identically $\cos x$. \n \n (${\cos}^{2} x = \cos x$ only when $\cos x = 1$ or 0) \n \n Dec 10, 2016 \n \n No. It is equal to ${\sin}^{2} x - 1$. \n \n #### Explanation: \n \n If we have $\frac{1 - x}{x}$, we can write it as $\frac{1}{x} - \frac{x}{x}$. \n \n The same way, $\frac{1 - {\csc}^{2} x}{\csc} ^ 2$ can be written as $\frac{1}{\csc} ^ 2 x - \frac{{\csc}^{2} x}{{\csc}^{2} x}$. \n \n This is equal to ${\sin}^{2} x - 1$. SMS newsy item created by Hannah Bryant at Wed 25 May 2022 1227 \n Type: Seminar \n Distribution: World \n Expiry: 31 May 2022 \n Calendar1: 31 May 2022 1500-1600 \n CalLoc1: Quad S224 & via Zoom \n CalTitle1: SMRI 'What is ...a virtual knot?' Hans Boden (McMaster University) \n Auth: hannahb@w1d4n6z2.staff.sydney.edu.au (hbry8683)

After Training 100% Checkpoint

Item Type: Journal Article Copyright of this article belongs to Elsevier. Division of Mechanical Sciences > Mechanical Engineering 28 May 2007 19 Sep 2010 04:36 http://eprints.iisc.ernet.in/id/eprint/10277 # Question #8de97 \n \n Dec 10, 2016 \n \n That is not an identity. \n \n #### Explanation: \n \n Recall that \n \n ${\cot}^{2} x + 1 = {\csc}^{2} x$. \n \n So, we can write \n \n $\frac{1 - {\csc}^{2} x}{\csc} ^ 2 x = \frac{1 - \left({\cot}^{2} x + 1\right)}{\csc} ^ 2 x$ \n \n $= {\cot}^{2} \frac{x}{\csc} ^ 2 x$ \n \n Recall also that $\cot x = \cos \frac{x}{\sin} x$ and $\csc x = \frac{1}{\sin} x$. \n \n This allows us to continue \n \n $= \frac{{\cos}^{2} \frac{x}{\sin} ^ 2 x}{\frac{1}{\sin} ^ 2 x}$ \n \n $= {\cos}^{2} \frac{x}{\sin} ^ 2 x \cdot {\sin}^{2} \frac{x}{1}$ \n \n $= {\cos}^{2} x$ \n \n Which is not identically $\cos x$. \n \n (${\cos}^{2} x = \cos x$ only when $\cos x = 1$ or 0) \n \n Dec 10, 2016 \n \n No. It is equal to ${\sin}^{2} x - 1$. \n \n #### Explanation: \n \n If we have $\frac{1 - x}{x}$, we can write it as $\frac{1}{x} - \frac{x}{x}$. \n \n The same way, $\frac{1 - {\csc}^{2} x}{\csc} ^ 2$ can be written as $\frac{1}{\csc} ^ 2 x - \frac{{\csc}^{2} x}{{\csc}^{2} x}$. \n \n This is equal to ${\sin}^{2} x - 1$. SMS newsy item created by Hannah Bryant at Wed 25 May 2022 1227 \n Type: Seminar \n Distribution: World \n Expiry: 31 May 2022 \n Calendar1: 31 May 2022 1500-1600 \n CalLoc1: Quad S224 & via Zoom \n CalTitle1: SMRI 'What is ...a virtual knot?' Hans Boden (McMaster University) \n Auth: hannahb@w1d4n6z2.staff.sydney.edu.au (hbry8683)

토큰 선택 확률 낮음 토큰 선택 확률 높음

(hbry8683)

(hbry8683)

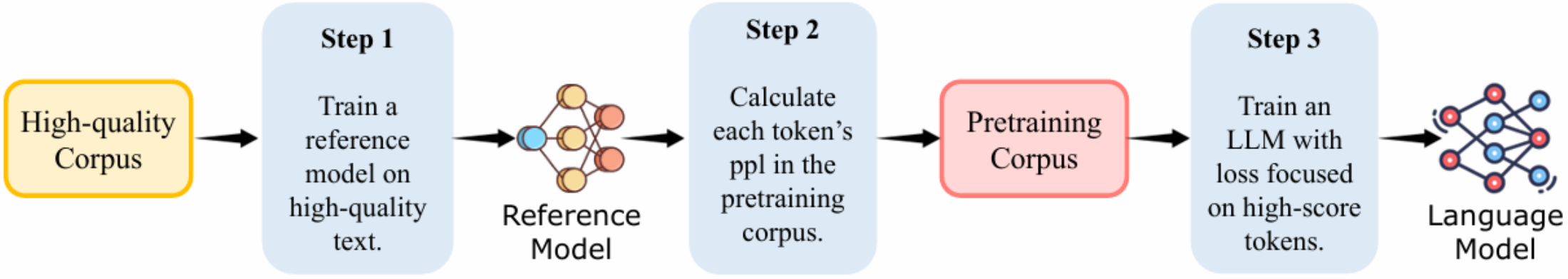
(hbry8683)

(hbry8683)

❖ Experimental Setup

- 1B 모델의 reference model은 Tinyllama-1.1Bmodel (Zhangetal.,2024)
  7B 모델의 reference model은 Mistral-7Bmodel (Jiangetal.,2023)을 사용

1. High-quality text를 얻는 방법은 GPT 또는 수동 선별 합성 데이터를 혼합한다
2. pretraining corpus는 OpenWebMath(OWM) dataset (Pasteretal.,2023)을 사용
   → OpenWebMath (OWM) dataset (Pasteretal.,2023), SlimPajam (Dariaetal.,2023), StarCoderData (Lietal.,2023a)를 혼합하여 일반화
3. reference model과 language model의 학습데이터를 동일하게 사용 (high-quality text를 얻을 수 없는 경우)



III Experiments

Table 1: **Few-shot CoT reasoning results of math pretraining.** All models are tested with few-shot prompting. Previous best results are highlighted in blue, while our best results are in purple. *Only unique math-related tokens are calculated. For RHO-1, we calculate only the selected tokens that are used for training. †We use OpenAI's MATH subset [Lightman et al., 2023] for evaluation, since some original test samples have been used in public training sets such as PRM800k. ‡The SAT only has 32 four-choice problems, so we average our results over the last three checkpoints, if available.

Model	$\|\theta\|$	Data	Uniq. Toks*	Train Toks	GSM8K	MATH†	SVAMP	ASDiv	MAWPS	TAB	MQA	MMLU STEM	SAT‡	AVG
1-2B Base Models														
Tinyllama	1.1B	-	-	-	2.9	3.2	11.0	18.1	20.4	12.5	14.6	16.1	21.9	13.4
Phi-1.5	1.3B	-	-	-	32.4	4.2	43.4	53.1	66.2	24.4	14.3	21.8	18.8	31.0
Qwen1.5	1.8B	-	-	-	36.1	6.8	48.5	63.6	79.0	29.2	25.1	31.3	40.6	40.0
Gemma	2.0B	-	-	-	18.8	11.4	38.0	56.6	72.5	36.9	26.8	34.4	50.0	38.4
DeepSeekLLM	1.3B	OWM	14B	150B	11.5	8.9	-	-	-	-	-	29.6	31.3	-
DeepSeekMath	1.3B	-	120B	150B	23.8	13.6	-	-	-	-	-	33.1	56.3	-
Continual Pretraining on Tinyllama-1B														
Tinyllama-CT	1.1B	OWM	14B	15B	6.4	2.4	21.7	36.7	47.7	17.9	13.9	23.0	25.0	21.6
RHO-1-Math	1.1B	OWM	14B	9B	29.8	14.0	49.2	61.4	79.8	25.8	30.4	24.7	28.1	38.1
Δ				-40%	+23.4	+11.6	+27.5	+24.7	+32.1	+7.9	+16.5	+1.7	+3.1	**+16.5**
RHO-1-Math	1.1B	OWM	14B	30B	36.2	15.6	52.1	67.0	83.9	29.0	32.5	23.3	28.1	40.9
≥ 7B Base Models														
LLaMA-2	7B		-	-	14.0	3.6	39.5	51.7	63.5	30.9	12.4	32.7	34.4	31.4
Mistral	7B		-	-	41.2	11.6	64.7	68.5	87.5	52.9	33.0	49.5	59.4	52.0
Minerva	8B	-	39B	164B	16.2	14.1	-	-	-	-	-	35.6	-	-
Minerva	62B	-	39B	109B	52.4	27.6	-	-	-	-	-	53.9	-	-
Minerva	540B	-	39B	26B	58.8	33.6	-	-	-	-	-	63.9	-	-
LLemma	7B	PPile	55B	200B	38.8	17.2	56.1	69.1	82.4	48.7	41.0	45.4	59.4	50.9
LLemma	34B	PPile	55B	50B	54.2	23.0	67.9	75.7	90.1	57.0	49.8	54.7	68.8	60.1
Intern-Math	7B	-	31B	125B	41.8	14.4	61.6	66.8	83.7	50.0	57.3	24.8	37.5	48.7
Intern-Math	20B	-	31B	125B	65.4	30.0	75.7	79.3	94.0	50.9	38.5	53.1	71.9	62.1
DeepSeekMath	7B	-	120B	500B	64.1	34.2	74.0	83.9	92.4	63.4	62.4	56.4	84.4	68.4
Continual Pretraining on Mistral-7B														
Mistral-CT	7B	OWM	14B	15B	42.9	22.2	68.6	71.0	86.1	45.1	47.7	52.6	65.6	55.8
RHO-1-Math	7B	OWM	14B	10.5B	66.9	31.0	77.8	79.0	93.9	49.9	58.7	54.6	84.4	66.2
Δ				-30%	+24.0	+8.8	+9.2	+8.0	+7.8	+4.8	+11.0	+2.0	+18.8	**+10.4**

Experiments

Table 2: **Tool-integrated reasoning results of math pretraining.**

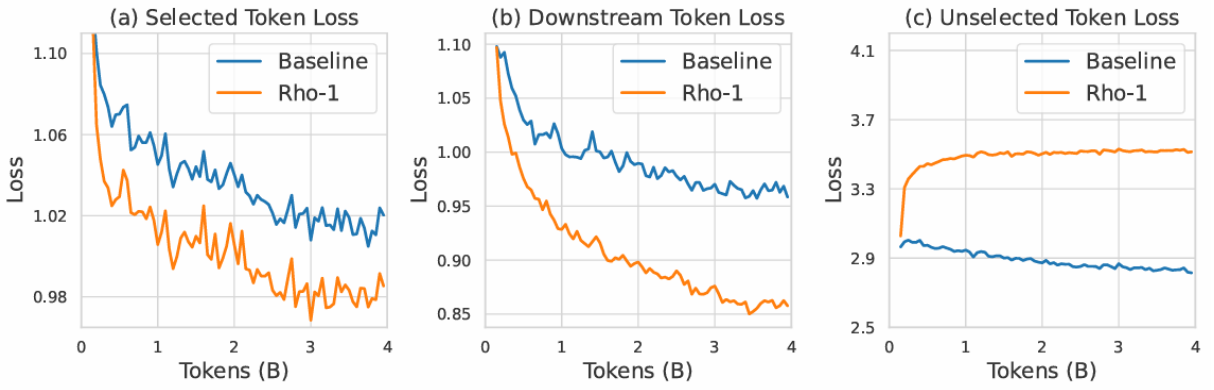
Model	Size	Tools	SFT Data	GSM8k	MATH	SVAMP	ASDiv	MAWPS	TAB	GSM-H	AVG
Used for SFT?				✓	✓	✗	✗	✗	✗	✗	
Previous Models											
GPT4-0314	-	✗	-	92.0	42.5	93.1	91.3	97.6	67.1	64.7	78.3
GPT4-0314 (PAL)	-	✓	-	94.2	51.8	94.8	92.6	97.7	95.9	77.6	86.4
MAmmoTH	70B	✓	MI-260k	76.9	41.8	82.4	-	-	-	-	-
ToRA	7B	✓	ToRA-69k	68.8	40.1	68.2	73.9	88.8	42.4	54.6	62.4
ToRA	70B	✓	ToRA-69k	84.3	49.7	82.7	86.8	93.8	74.0	67.2	76.9
DeepSeekMath	7B	✓	ToRA-69k	79.8	52.0	80.1	87.1	93.8	85.8	63.1	77.4
Our Pretrained Models											
TinyLlama-CT	1B	✓	ToRA-69k	51.4	38.4	53.4	66.7	81.7	20.5	42.8	50.7
RHO-1-Math	1B	✓	ToRA-69k	59.4	40.6	60.7	74.2	88.6	26.7	48.1	56.9
Δ				+8.0	+2.2	+7.3	+7.5	+6.9	+6.2	+5.3	**+6.2**
Mistral-CT	7B	✓	ToRA-69k	77.5	48.4	76.9	83.8	93.4	67.5	60.4	72.6
RHO-1-Math	7B	✓	ToRA-69k	81.3	51.8	80.8	85.5	94.5	70.1	63.1	75.3
Δ				+3.8	+3.4	+3.9	+1.7	+1.1	+2.6	+2.7	**+2.7**

Table 3: **Self-Reference results.** We use OpenWebMath (OWM) to train the reference model.

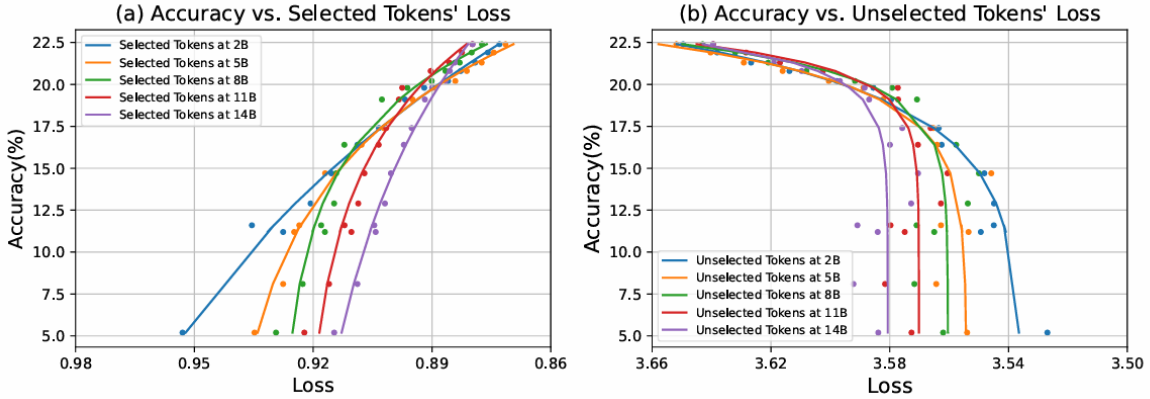
Model	Score Function	Data	Uniq. Toks	Train Toks	GSM8K	MATH	SVAMP	ASDiv	MAWPS	MQA	AVG
Tinyllama-CT (RM)	-	OWM	14B	15B	6.3	2.6	21.7	36.7	47.7	13.9	21.5
Tinyllama-SLM	$\mathcal{L}_{\text{RM}}$	OWM	14B	10.5B	6.7	4.6	23.3	40.0	54.5	14.3	23.9
Tinyllama-SLM	$\mathcal{H}_{\text{RM}}$	OWM	14B	10.5B	7.0	4.8	23.0	39.3	50.5	13.5	23.0
Tinyllama-SLM	$\mathcal{L}_{\text{RM}} \cap \mathcal{H}_{\text{RM}}$	OWM	14B	9B	7.1	5.0	23.5	41.2	53.8	18.0	24.8
Tinyllama-CT	-	PPile	55B	52B	8.0	6.6	23.8	41.0	54.7	14.2	24.7
Tinyllama-SLM	$\mathcal{L}_{\text{RM}} \cap \mathcal{H}_{\text{RM}}$	PPile	55B	36B	8.6	8.4	24.4	43.6	57.9	16.1	26.5

❖ Selected Token Loss Aligns Better with Downstream Performance

- token loss를 관찰해보면 selected token의 loss는 줄어드는 반면, unselected token의 loss는 오히려 증가한다
- 하지만 unselected loss의 증가에도 불구하고 downstream token의 loss가 줄어듦



- selected token의 loss가 거듭제곱 법칙을 따르는 것으로 보아 이것이 미치는 효과가 더 큰 것으로 판단할 수 있음



❖ What Tokens are Selected with SLM?

- ppl (perplexity): 초과 손실이 높을수록 perplexity가 높음
- 선택된 토큰에서 "double descent" 현상 발생
- 초반에 perplexity가 낮을 수록 후반 perplexity가 높음

- → perplexity가 높은 토큰들을 먼저 선택하기 때문인 것으로 추론

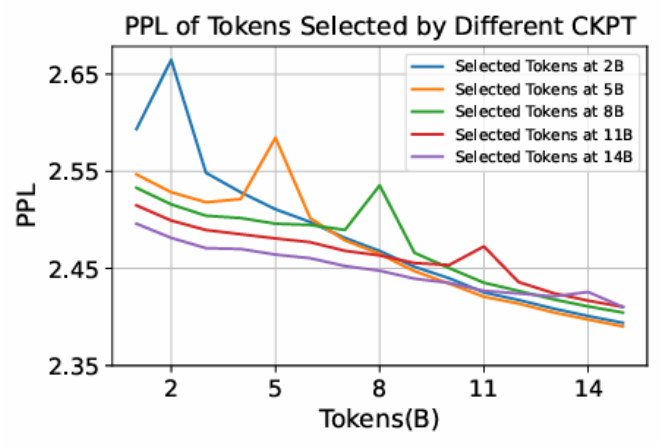


Figure 8: **The PPL of tokens selected by different checkpoint.** We test the PPL of the tokens selected at 2B, 5B, 8B, 11B, and 14B.

❖ Effect of Token Select Ratio

- 가장 효과적인 선택비율은 60%

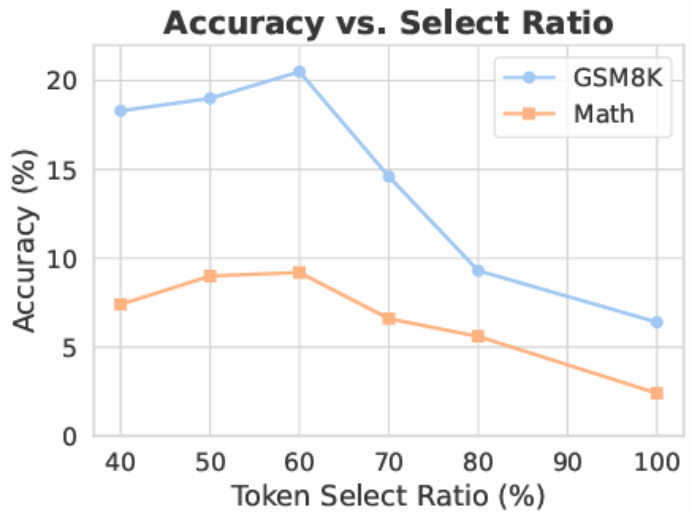


Figure 9: **Effect of token select ratio.** We train 1B LM with SLM objective on 5B tokens.

IV Conclusion

❖ Contribution

- 사전 학습에 있어서 모든 토큰들이 학습에 동일하게 기여하지 않으므로 중요한 토큰을 선택하여 학습을 하는 것이 학습효율을 높였다
- 학습 속도, 학습 성능면에서 향상을 이뤄냈다

❖ Limitation

- 예산 문제로 작은 모델과 작은 데이터셋으로 훈련시켰다
  추후 사이즈를 늘려 성능을 확인해볼 필요가 있다
- 토큰을 계산하기 위해 성능이 좋은 reference model이 필요하다