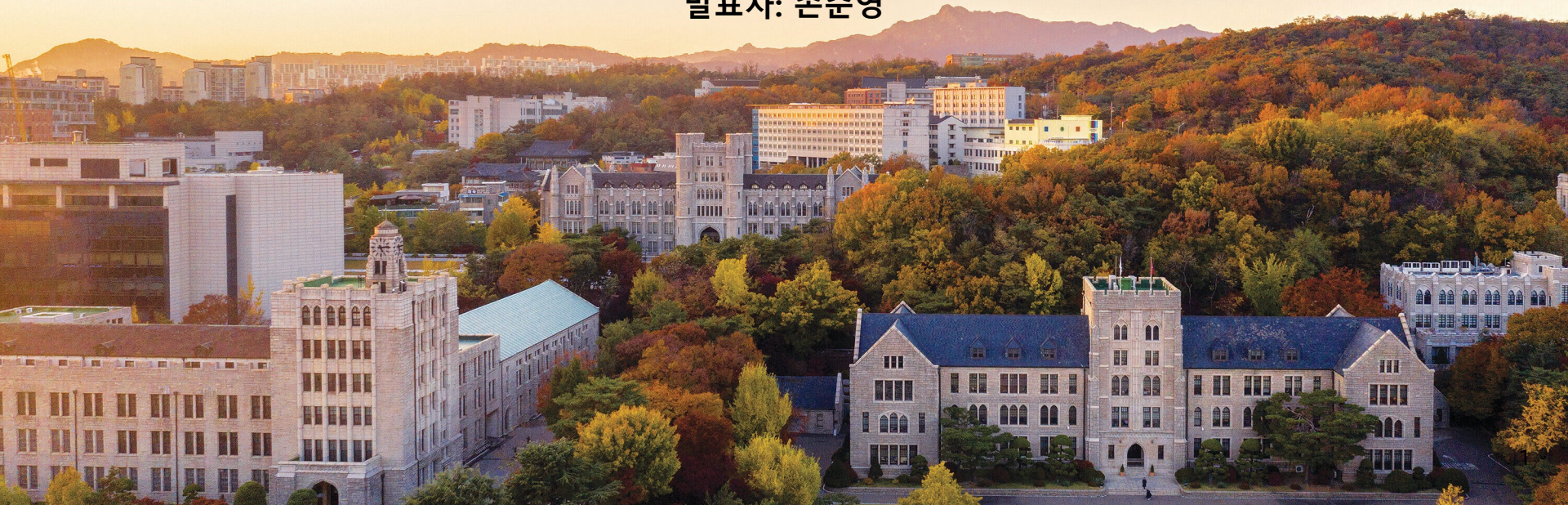
고려대학교 KOREA UNIVERSITY

Natural Language Processing & Artificial Intelligence

여름방학 세미나

고려대학교 NLP&AI 연구실
발표자: 손준영



◆ Retrieval meets Long Context Large Language Models

Published as a conference paper at ICLR 2024

RETRIEVAL MEETS LONG CONTEXT LARGE LANGUAGE MODELS

Peng Xu[†], **Wei Ping**[†], **Xianchao Wu, Lawrence McAfee**
Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina
Mohammad Shoeybi, Bryan Catanzaro
NVIDIA
[†]{pengx, wping}@nvidia.com

LLM의 **"Long context extension"** versus "**retrieval-augmentation**" 선택을 위한 general insights 제공

LLMs을 활용하여 9 downstream **long context tasks**에 대한 comprehensive study 수행 및 분석

◆ Retrieval meets Long Context Large Language Models

	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
# of samples	200	1,726	2,000	2,000	200	200	150
avg doc length	14,140	4,912	84,770	6,592	16,198	13,319	7,185
avg top-5 chunks	2,066	2,071	2,549	2,172	2,352	2,322	2,385
avg top-10 chunks	4,137	3,716	5,125	4,018	4,644	4,554	4,305
avg top-20 chunks	8,160	4,658	10,251	5,890	9,133	8,635	6,570

Table 1: Statistics of seven datasets used for zero-shot evaluation. All lengths are counted by the number of tokens using Llama2-70B tokenizer, and "avg top N chunks" denotes the average number of tokens from the top N retrieved chunks. Figure 2 gives more details.

- **Tasks**: Single, multi document QA, Summarization **(Zero-shot evaluation)**
- **Models**: LLaMa2 (7B, 70B), GPT-43B
- **Context Window Extension**: Position interpolation method 적용하여 finetuning on the Pile dataset
  - **LLaMa2**: 16K, 32K까지 확장
  - **GPT-43B**: 16K까지 확장

◆ Retrieval meets Long Context Large Language Models

Token length distributions

- 300 words로 chunking 했을 때 Token length distribution
- Top 5 chunks의 경우 4K LLMs에서 어느정도 커버 가능하나, 그 이상은 한계가 있음
- 16K LLM 기준 top-10 또는 top-20 chunks 커버 가능

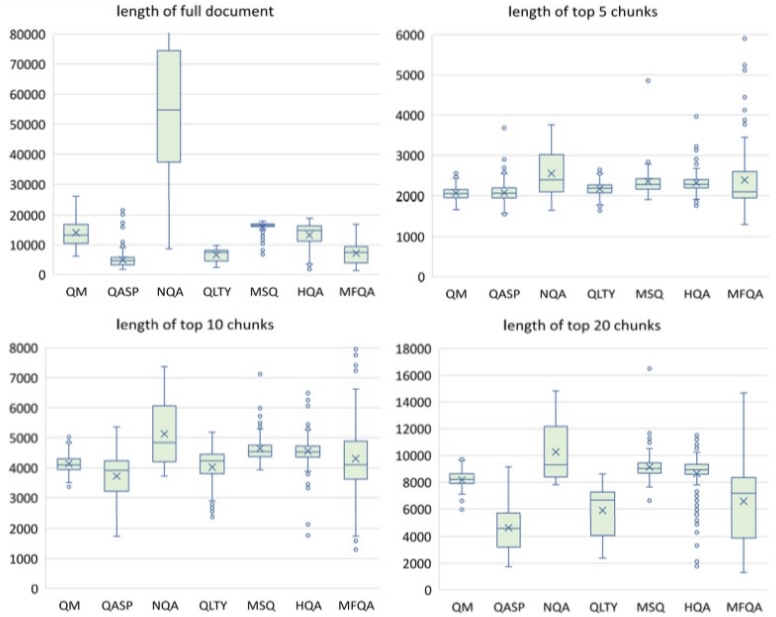


Figure 2: Token length distribution of the full document and the top-5, 10, 20 chunks of the seven datasets.

◆ Retrieval meets Long Context Large Language Models

Main Results

- Retrieval is especially helpful for 4K LLMs
  (Comparable to 16K long context LLMs, while being more efficient at inference)

Model	Seq len.	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
GPT-43B	4k	26.44	15.56	23.66	15.64	49.35	11.08	28.91	40.90
+ ret	4k	29.32	16.60	23.45	19.81	51.55	14.95	34.26	44.63
GPT-43B	16k	29.45	16.09	25.75	16.94	50.05	14.74	37.48	45.08
+ ret	16k	**29.65**	15.69	23.82	21.11	47.90	15.52	36.14	47.39
Llama2-70B	4k	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
+ ret	4k	36.02	17.41	28.74	23.41	70.15	21.39	42.06	48.96
Llama2-70B	16k	36.78	16.72	30.92	22.32	**76.10**	18.78	43.97	48.63
+ ret	16k	37.23	**18.70**	29.54	23.12	70.90	23.28	44.81	50.24
Llama2-70B	32k	37.36	15.37	**31.88**	23.59	73.80	19.07	49.49	48.35
+ ret	32k	**39.60**	18.34	31.27	**24.53**	69.55	**26.72**	**53.89**	**52.91**
Llama2-7B	4k	22.65	14.25	22.07	14.38	40.90	8.66	23.13	35.20
+ ret	4k	**26.04**	16.45	22.97	18.18	43.25	14.68	26.62	40.10
Llama2-7B	32k	**28.20**	16.09	23.66	19.07	44.50	15.74	31.63	46.71
+ ret	32k	27.63	17.11	23.25	19.12	43.70	15.67	29.55	45.03

◆ Retrieval meets Long Context Large Language Models

Main Results

- **Interesting point**: Long context LLMs w/ Retrieval (16K and 32K) 가 4K LLMs w/ Retrieval 보다 성능이 좋음
  : even same top 5 chunks of evidence가 입력되는데..
  ➔ "lost in the middle" phenomenon (U-shaped curve)

Model	Seq len.	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
GPT-43B	4k	26.44	15.56	23.66	15.64	49.35	11.08	28.91	40.90
+ ret	4k	29.32	16.60	23.45	19.81	51.55	14.95	34.26	44.63
GPT-43B	16k	29.45	16.09	25.75	16.94	50.05	14.74	37.48	45.08
+ ret	16k	**29.65**	15.69	23.82	21.11	47.90	15.52	36.14	47.39
Llama2-70B	4k	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
+ ret	4k	36.02	17.41	28.74	23.41	70.15	21.39	42.06	48.96
Llama2-70B	16k	36.78	16.72	30.92	22.32	**76.10**	18.78	43.97	48.63
+ ret	16k	37.23	**18.70**	29.54	23.12	70.90	23.28	44.81	50.24
Llama2-70B	32k	37.36	15.37	**31.88**	23.59	73.80	19.07	49.49	48.35
+ ret	32k	**39.60**	18.34	31.27	**24.53**	69.55	**26.72**	**53.89**	**52.91**
Llama2-7B	4k	22.65	14.25	22.07	14.38	40.90	8.66	23.13	35.20
+ ret	4k	**26.04**	16.45	22.97	18.18	43.25	14.68	26.62	40.10
Llama2-7B	32k	**28.20**	16.09	23.66	19.07	44.50	15.74	31.63	46.71
+ ret	32k	27.63	17.11	23.25	19.12	43.70	15.67	29.55	45.03

◆ Retrieval meets Long Context Large Language Models

Main Results

- **Interesting point**: Long context LLMs w/ Retrieval (16K and 32K) 가 4K LLMs w/ Retrieval 보다 성능이 좋음
  : even same top 5 chunks of evidence가 입력되는데..
  ➔ "lost in the middle" phenomenon (U-shaped curve)

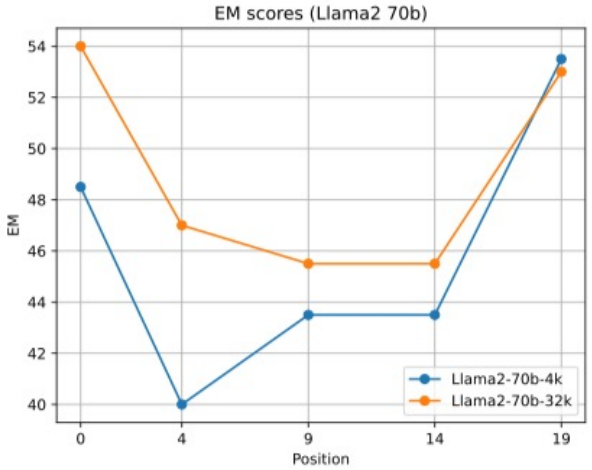


Figure 1: Llama2-70B also displays lost-in-the-middle phenomenon

◆ Retrieval meets Long Context Large Language Models

Main Results

- Long context capability와 model size 사이의 상관관계

  이전 연구에서는 Retrieval이 4k LLM (Long context X) 모델에게는 효과적이나, 이미 Strong Long Context capability가 있는 LLM 에게는 효과적이지 않았다는 분석 결과
  ➔ Model size에 따른 변인을 고려하지 않음

- Llama2-70B-32k-ret이 32k baseline을 크게 압도 (37.36 vs. 39.60)

  Retrieval is not helpful for Llama2-7B-32k, But further boost for larger models (Llama2-70B-32k)

  ➔ 32K 이상의 seq len에선 모델 사이즈가 따라줘야 retrieval 부착시 성능이 따라옴 (zero-shot capability, in-context learning capability, instruction following capability 등..)

Model	Seq len.	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
GPT-43B	4k	26.44	15.56	23.66	15.64	49.35	11.08	28.91	40.90
+ ret	4k	29.32	16.60	23.45	19.81	51.55	14.95	34.26	44.63
GPT-43B	16k	29.45	16.09	25.75	16.94	50.05	14.74	37.48	45.08
+ ret	16k	**29.65**	15.69	23.82	21.11	47.90	15.52	36.14	47.39
Llama2-70B	4k	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
+ ret	4k	36.02	17.41	28.74	23.41	70.15	21.39	42.06	48.96
Llama2-70B	16k	36.78	16.72	30.92	22.32	**76.10**	18.78	43.97	48.63
+ ret	16k	37.23	**18.70**	29.54	23.12	70.90	23.28	44.81	50.24
Llama2-70B	32k	37.36	15.37	**31.88**	23.59	73.80	19.07	49.49	48.35
+ ret	32k	**39.60**	18.34	31.27	**24.53**	69.55	**26.72**	**53.89**	**52.91**
Llama2-7B	4k	22.65	14.25	22.07	14.38	40.90	8.66	23.13	35.20
+ ret	4k	**26.04**	16.45	22.97	18.18	43.25	14.68	26.62	40.10
Llama2-7B	32k	**28.20**	16.09	23.66	19.07	44.50	15.74	31.63	46.71
+ ret	32k	27.63	17.11	23.25	19.12	43.70	15.67	29.55	45.03

◆ Retrieval meets Long Context Large Language Models

More contexts get more improvements?

- 전반적으로 Top-5 or top-10 에서 가장 좋은 성능을 보임
- 20 chunks or more의 경우 not helpful and sometime hurt the performance
  ➔ "believe" this is related to the "lost in the middle" phenomenon or the model is getting distracted by irrelevant information and therefore needs further research.

Seq len	Setting	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
4k	base	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
	top-5	**35.73**	18.14	29.20	23.39	70.30	20.09	41.54	47.45
	top-10	34.62	16.54	28.67	24.38	68.70	19.00	42.18	42.84
	top-20	34.61	16.52	28.67	24.38	68.70	19.00	42.18	42.84
16k	base	36.78	16.72	30.92	22.32	76.10	18.78	43.97	48.63
	top-5	37.23	18.70	29.54	23.12	70.90	23.28	44.81	50.24
	top-10	**38.31**	18.41	30.20	25.53	73.60	22.78	47.72	49.91
	top-20	36.61	17.26	29.60	25.81	72.30	22.69	41.36	47.23
32k	base	37.36	15.37	31.88	23.59	73.80	19.07	49.49	48.35
	top-5	**39.60**	18.34	31.27	24.53	69.55	26.72	53.89	52.91
	top-10	38.98	17.71	30.34	25.94	70.45	22.80	55.73	49.88
	top-20	38.38	16.36	30.42	24.42	69.60	24.51	54.67	48.65

Table 5: Comparisons of adding top-5/10/20 retrieved chunks to the context under 4k, 16k, and 32k input sequence lengths using Llama2-70B. More context does not always give better results.

◆ Retrieval meets Long Context Large Language Models

VS. OpenAI models

- [Context Window Extension + Retrieval Augmentation]을 모두 적용한 Llama2-70B-32k-ret이 GPT-3.5-turbo-16k-ret보다 성능이 좋음

Model	Avg-7	Avg-4*	QM*	QASP*	NQA*	QLTY*	MSQ	HQA	MFQA
Davinci003 (175B)	39.2	40.8*	16.9*	52.7*	24.6*	69.0*	22.1	41.2	47.8
GPT-3.5-turbo (4k)	38.4	39.2*	15.6*	49.3*	25.1*	66.6*	21.2	40.9	49.2
+ret							24.4	49.5	49.5
GPT-3.5-turbo-16k	42.8	42.4	17.6	50.5	28.8	72.6	26.9	51.6	52.3
+ret							30.4	46.6	52.8
Llama2-70B-32k	40.9	42.4	15.6	45.9	28.4	79.6	19.1	49.5	48.4
Llama2-70B-32k-ret	**43.6**	**43.0**	18.5	46.3	31.5	75.6	26.7	53.9	52.9

◆ Retrieval meets Long Context Large Language Models

Conclusion

- Retrieval largely boosts the performance of both 4K short context LLM and 16/32K long context LLMs.
- The 4K context LLMs with simple retrieval-augmentation can perform comparable to 16K long context LLMs, while being more efficient at inference.
- After **context window extension** and **retrieval-augmentation**, the best model Llama2-70B-32k-ret can outperform GPT-3.5-turbo-16k and Davinci003.

◆ Understanding Finetuning for Factual Knowledge Extraction

ICML 2024

Understanding Finetuning for Factual Knowledge Extraction

Gaurav Ghosal [1] Tatsunori Hashimoto [2] Aditi Raghunathan [1]

"QA finetuning data"가 downstream factuality에 주는 영향 분석

◆ Understanding Finetuning for Factual Knowledge Extraction

- Factuality 관점에서 모델이 올바른 정답을 알고 있는데도 잘못된 대답을 하는 경우 존재:
  - What factors determine the performance of fine-tuning?
  - What is the mechanism by which fine-tuning improves factuality?

◆ Understanding Finetuning for Factual Knowledge Extraction

Conceptual Mechanism of Finetuning on Popular versus Unpopular Knowledge

- A Factual question이 주어졌을 때, Language model은 relevant memorized knowledge를 사용하여 답변할 수도 있고, more general "shortcut"을 사용하여 plausible but incorrect response를 뱉을 수도 있음

  Asking about Person's occupation → take the shortcut of responding with a word that is generally associated with occupations (i.e. actor)

  이러한 shortcut 사용이 fine-tuning 에서 학습이 된다면 pre-training 단계에서 학습한 지식에 악영향을 주고, test data에서 모델이 less factually 행동하도록 할 수 있음

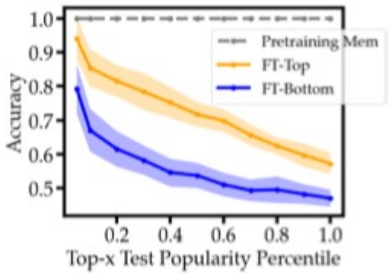
◆ Understanding Finetuning for Factual Knowledge Extraction

Conceptual Mechanism of Finetuning on Popular versus Unpopular Knowledge

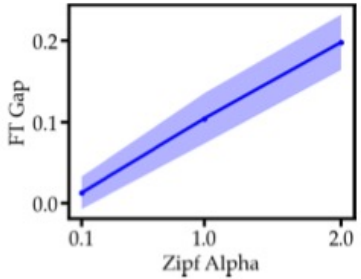
- Popular knowledge로 Fine-tuning하는 경우, Subject entity에 대한 attend를 하는 경향성이 강해지며
- Less popular knowledge로 Fine-tuning하는 경우, subject attention은 감소하고 relation과 같은 다른 토큰에 attend하는 경향성이 강해짐

◆ Understanding Finetuning for Factual Knowledge Extraction

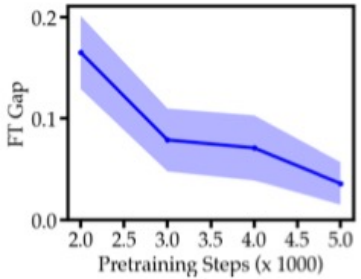
Simulation Study of Finetuning for Knowledge Extraction



(a) Impact of Finetuning Dataset



(b) Effect of Zipf Alpha

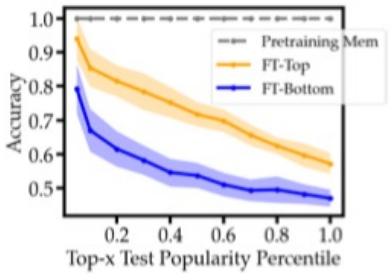


(c) Effect of Pretraining Steps

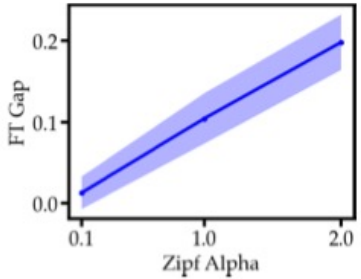
- (a) Fine-tuning Fact Popularity Impacts Downstream Performance
  - FT-Top (Fine-tuning via more popular facts) / FT-Bottom (Fine-tuning via less popular facts): both are present in pre-training corpus (its not new knowledge)
  - fine-tuning FT-Top results in 10% improvement in factuality
  - Test set에 더 많은 less popular facts를 포함시켜도 성능 차이가 유지, 특히 0.05 -> 0.1로 확장 시 FT-Top과 FT-Bottom의 성능 차이가 2배까지 격차가 남

◆ Understanding Finetuning for Factual Knowledge Extraction

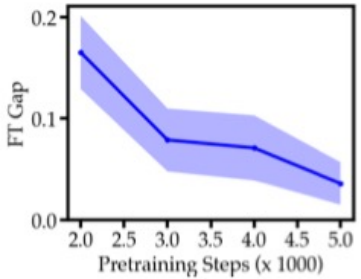
Simulation Study of Finetuning for Knowledge Extraction



(a) Impact of Finetuning Dataset



(b) Effect of Zipf Alpha



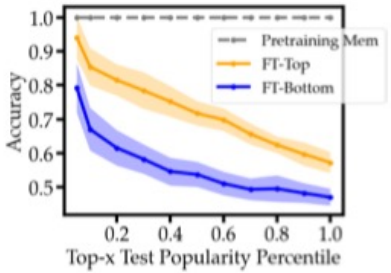
(c) Effect of Pretraining Steps

- (b) Impact of Long-Tailedness in Pretraining Corpus
  - Zipf's law: corpus에서 특정 단어들을 frequency가 높은 순서로 나열하였을 때, 모든 단어의 frequency는 해당 단어의 ranking에 반비례한다는 경험적 법칙
  - $word\ frequency \propto \frac{1}{(rank+b)^a}$
  - Zipf Alpha 값이 높으면 pre-training 단계에서 more popular facts를 위주로 학습, 낮으면 uniformly 학습
  - Pre-training 단계에서 facts를 uniformly 학습할 수록 (lowering a) FT-Top과 FT-Bottom의 성능 격차가 줄어듦
  - Pre-training 단계에서 more popular facts를 위주로 학습할 수록 FT-Top과 FT-Bottom의 성능 격차가 커짐

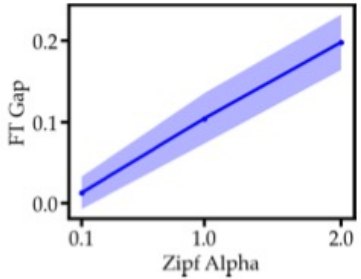
➔ fine-tuning dataset의 구성과 pre-training 단계에 학습한 fact frequency에 따른 상관 관계가 있다

◆ Understanding Finetuning for Factual Knowledge Extraction

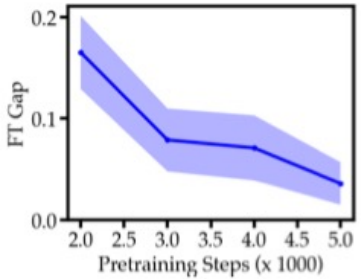
Simulation Study of Finetuning for Knowledge Extraction



(a) Impact of Finetuning Dataset



(b) Effect of Zipf Alpha



(c) Effect of Pretraining Steps

- (c) Impact of the Number of Pretraining Steps
  - 충분히 많은 반복 학습 → 차이를 좁히지만, impractical due to the large scale of pretraining data (and also model size)

➔ Pretraining corpus의 fact frequency와 fact salience가 관련이 있음

◆ Understanding Finetuning for Factual Knowledge Extraction

“Attention Imbalance”

- A one-layer transformer 를 활용한 분석을 통해, 모델이 실제로 알고있는 지식임에도, “Attention Imbalance”로 인하여 incorrectly respond 할 수 있음을 증명

- Assumptions:
  1) Answer Diversity: 모든 relation r에 대한 fact a는 적어도 한 번 pre-training dataset에 포함됨

  2) Non-Uniform Relation Marginal: relation r에 속한 fact a는 모두 동일하지 않음

  3) All Facts Memorized: 모델이 관계 r과 개체 s에 대하여 최적의 fact a를 uniquely 결정할 수 있도록 보장

  ➔ Value Matrix (W_V) 만 활용해도 이론적으로 100% accuracy 달성 가능

Synthetic Setting:

Orthogonal token embeddings

$$\phi(t)\ for\ token\ t \in T$$

Two learnable parameters

$$W_V, W_{KQ} \in R^{|T|\times|T|}$$

Output head as Identity operation

$$X = [\phi(s)\ \ \phi(r)] \rightarrow a \in T$$

$$f(s, r; W_{\mathrm{V}}, W_{\mathrm{KQ}}) = \sigma(\text{Self-Att}(X; W_{\mathrm{V}}, W_{\mathrm{KQ}}))$$

◆ Understanding Finetuning for Factual Knowledge Extraction

“Attention Imbalance”

- 이전 세팅 및 assumptions에 아래에서 $f(s, p_r; W_V, 0)$는 100% accuracy 달성 가능하나, $f(s, p_r; W_V, W_{KQ})$는 100%를 달성할 수 없는 경우가 존재함을 이론적으로 증명하였음
- Incorrect Prediction이 발생하는 조건을 도출함
- 즉, subject token에 대한 attention weight가 0에 가까워질 수록 incorrectly predicted 되나, 이러한 현상은 Factual Salience에 영향을 받음을 증명

➔ 즉, 모델은 실제로 모든 facts를 알고 있음에도 (value matrix), attention imbalance 로 인하여 잘못 예측할 수 있음

$$Att_s \leq \frac{d}{S(s,r,a)} \; for\ a\ constant\ d$$

◆ Understanding Finetuning for Factual Knowledge Extraction

“Finetuning Attention Dynamics”

- Low-salience facts는 모델의 attention imbalance를 악화시키도록 학습하는 데 기여함
- $s_{rel} - p_{rel} < 0$인 경우, s에 대한 attention이 감소
- $s_{rel} - p_{rel} > 0$인 경우 subject attention이 증가함
- 즉, Pre-training 단계에서 subject entity 에 충분한 학습을 수행한 fact라면 Fine-tuning 에서 imbalance를 발생시키지 않고, pre-training 단계에서 거의 학습하지 못한 fact를 Fine-tuning 단계에서 학습하는 경우 attention imbalance를 증폭

$$s_{rel} = \left(\phi(a) - f\left(s, p_r; W_V, W_{KQ}\right)\right)^T (W_V \phi(s))$$

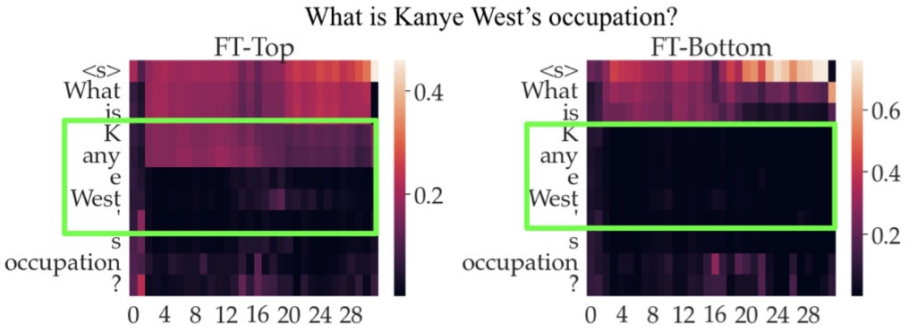
$$p_{rel} = \left(\phi(a) - f\left(r, p_r; W_V, W_{KQ}\right)\right)^T (W_V \phi(p_r))$$

$$-\frac{\partial L}{\partial W_{\mathrm{KQ}}} \propto (s_{\mathrm{rel}} - p_{\mathrm{rel}})(\phi(s)\phi(p_r)^\top - \phi(p_r)\phi(p_r)^\top).$$

◆ Understanding Finetuning for Factual Knowledge Extraction

Analysis of Attention Patterns

- FT-Top에 학습한 모델이 FT-Bottom에 학습한 모델보다 subject entity에 더 attend하는 경향이 있음
- FT-Bottom에 학습한 모델은 subject tokens에 거의 attend하지 않음



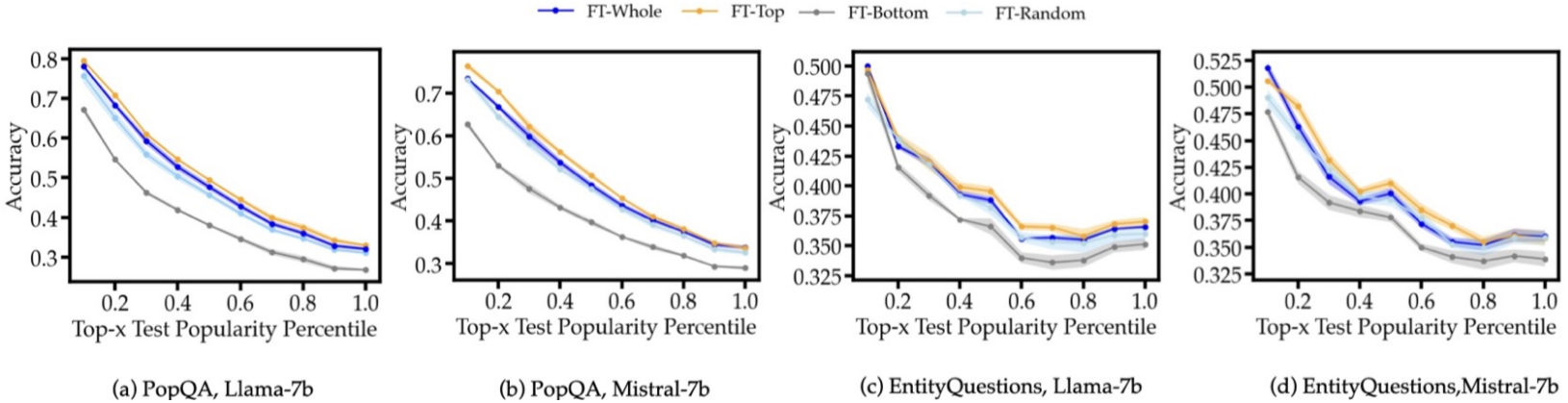
(b) Attention Pattern on FT-Top versus FT-Bottom

◆ Understanding Finetuning for Factual Knowledge Extraction

Experiments on Real QA Datasets

1. Unpopular Facts Harm Downstream Factuality (Popular Facts Mitigate Unpopular)
   Both QA datasets (PopQA, Entity Questions) and models (Llama, Mistral)에서 동일한 경향성

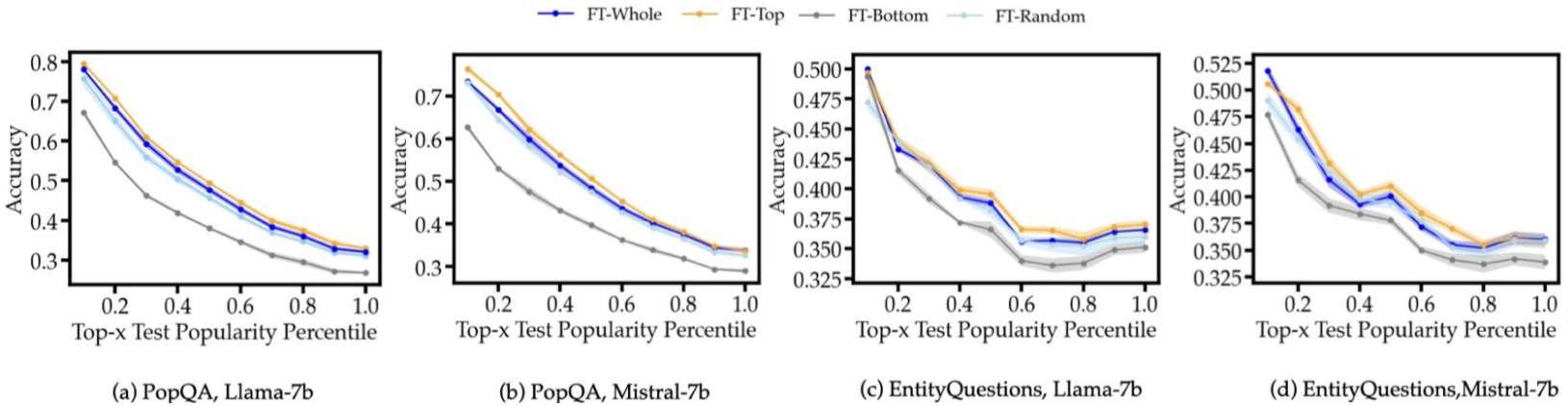
2. Impact Relative to Test Popularity:
   . Top-x Test popularity = 10%에서 FT-Top과 FT-Bottom의 성능 격차가 가장 커짐
   . Unpopular examples를 더 포함시키더라도 more popular examples로 학습한 경우 더 성능이 좋음



◆ Understanding Finetuning for Factual Knowledge Extraction

Experiments on Real QA Datasets

3. (FT-Top vs. FT-Whole): 전체 데이터를 사용해서 학습하는 것보다, 일부 popular subset을 학습한 모델이 더 성능이 좋음
   (a) only a subset of the most popular facts로 학습하는 것이 factual QA에 더 유용
   (b) Additional QA examples를 학습 데이터에 포함시키는 것은 harmful 할 수 있음



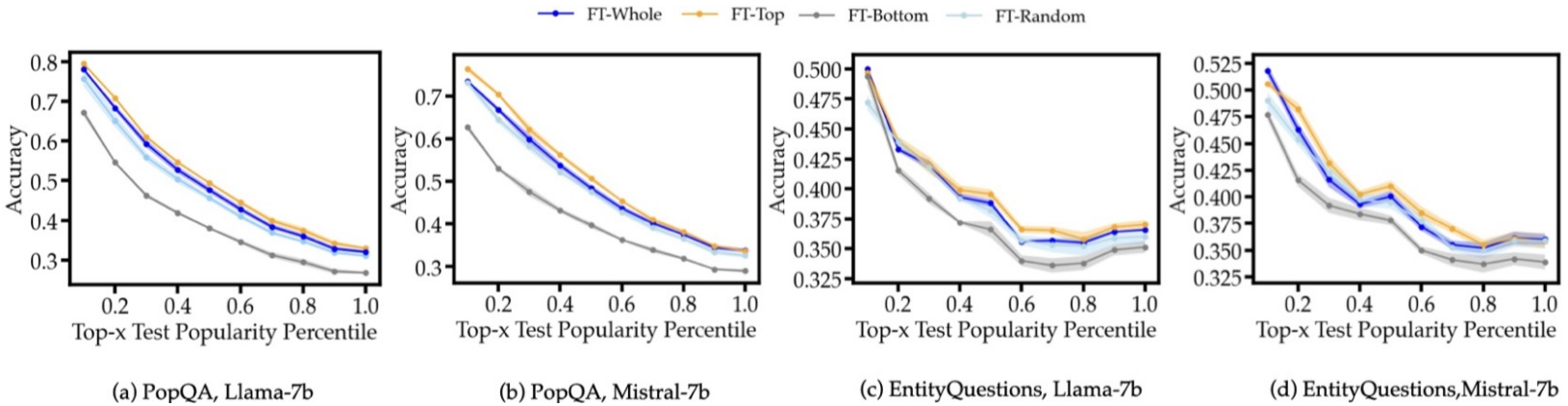
(a) PopQA, Llama-7b (b) PopQA, Mistral-7b (c) EntityQuestions, Llama-7b (d) EntityQuestions,Mistral-7b

◆ Understanding Finetuning for Factual Knowledge Extraction

Experiments on Real QA Datasets

4. (FT-Random vs. FT-Bottom): Randomly sampled subset은 FT-Bottom보다 popular examples를 더 포함
   FT-Random의 some popular points로 인하여 FT-Bottom 보다 성능이 높게 나타난 것

→ more popular examples가 attention imbalance를 해소하는 데 도움이 된다는 이론적 분석을 지지하는 실험 결과



◆ Understanding Finetuning for Factual Knowledge Extraction

Conclusion

Fine-tuning 단계에서 QA finetuning의 factuality를 고려할 때, pre-training 단계에서 많이 봤을 법한 popular facts가 도움이 됨 <-> less popular facts는 attention imbalance를 초래할 수 있어 factuality를 감소시킬 수 있음

이론적+실험적으로 증명

감사합니다.