2024 하계 세미나

2024. 08. 22

발표자 장윤나

INJECTING NEW KNOWLEDGE INTO LARGE LANGUAGE MODELS VIA SUPERVISED FINE-TUNING

Microsoft
Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes,
Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, Todd Hendry

Arxiv 2024

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

1. Introduction

- LLM에 new & out-of-domain 지식에 적응 시키는 거은 여전히 challenging
  - 특히 모델의 ***knowledge cutoff*** (더 이상 update 되지 않는 시점) 이후 발생한 사건에 대해
- 기존에는 Few-shot learning, Prompt engineering, RAG, SFT, RLHF 혹은 이것들의 혼합
- Internet-scale corpora에 대한 PT를 통해 지식을 배우지만 시간의 흐름에 따라 제한되는 지식이 있음
- RAG는 외부 지식을 참조하여 모델 답변 증강이 가능하지만 모델 내부에 직접 넣지는 못했음
- **최근 스포츠 이벤트 지식**을 LLM 내부에 직접 주입시키는 Supervised Fine-Tuning (SFT) 방법 연구
- 모델의 새로운 정보 습득을 위한 다른 데이터셋 생성 방법 비교: **token-based** & **fact-based scaling**

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

2. Dataset Generation

- 6개의 Wikipedia articles:
  - 2023 Cricket World Cup
  - 2023 American Football Superbowl
  - 2023 FIFA Women's World Cup — Post-cutoff
  - 2023 PGA Championship
  - 2023 NCAA Men's Basketball Division I Tournament
  - 2018 FIFA World Cup — Pre-cutoff
- Fact/statistic 정보가 풍부, 이해 쉬움, 관련 정보를 날짜 기준으로 구분 가능
- 각 문서에 대해 plain text 추출, 섹션별 필터링, cleaning 후 데이터셋 작업

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

2. Dataset Generation – Token-based Scaling

- Article의 overview 섹션에 기반하여 Q&A 쌍을 manual하게 제작하고 question 리스트에 추가
- 각 섹션마다 token 카운트, GPT-4에 unique한 Q&A 쌍을 각 섹션 토큰 수의 10배가 넘을 때까지 반복해서 생성하도록 함 (10x)
- 10x의 subset을 이용하여 섹션마다 1x, 5x 데이터셋 제작 – FT 데이터로 사용
- 평가셋은 1x token scale, unique하며 train set에 없는 질문들로 뽑음

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

2. Dataset Generation – Fact-based Scaling

- 문서에 있는 atomic facts를 GPT-4 활용하여 리스팅
- Atomic facts를 iterate하며 10개의 unique Q&A 쌍 생성 (GPT-4)
- Question 목록에 없는 경우에만 추가가 되고, 있다면 새로운 question을 생성하도록 하여 10x 셋을 만들고, subset으로 1x, 5x 데이터셋 확보
- 만들 때 fact가 너무 broad, unclear한 경우 건너뛰며, topic과도 관련 없으면 패스
- 평가셋은 1x scale

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

3. Training Methodology & Evaluation

- GPT-4 - v0613 (cutoff 09.2021)을 FT에 활용
- LoRA 활용해서 FT (rank 16, batch 1, 3 epochs)
- 모든 training samples는 context length안에 다 들어온다 함
- Gradient updates는 user prompt 제외한 답변 prompt에만 진행

- 평가는 각 데이터셋에 GPT-4를 학습시킨 후 inference에 학습된 lora delta값 활용

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

4. Results – Token-based Scaling

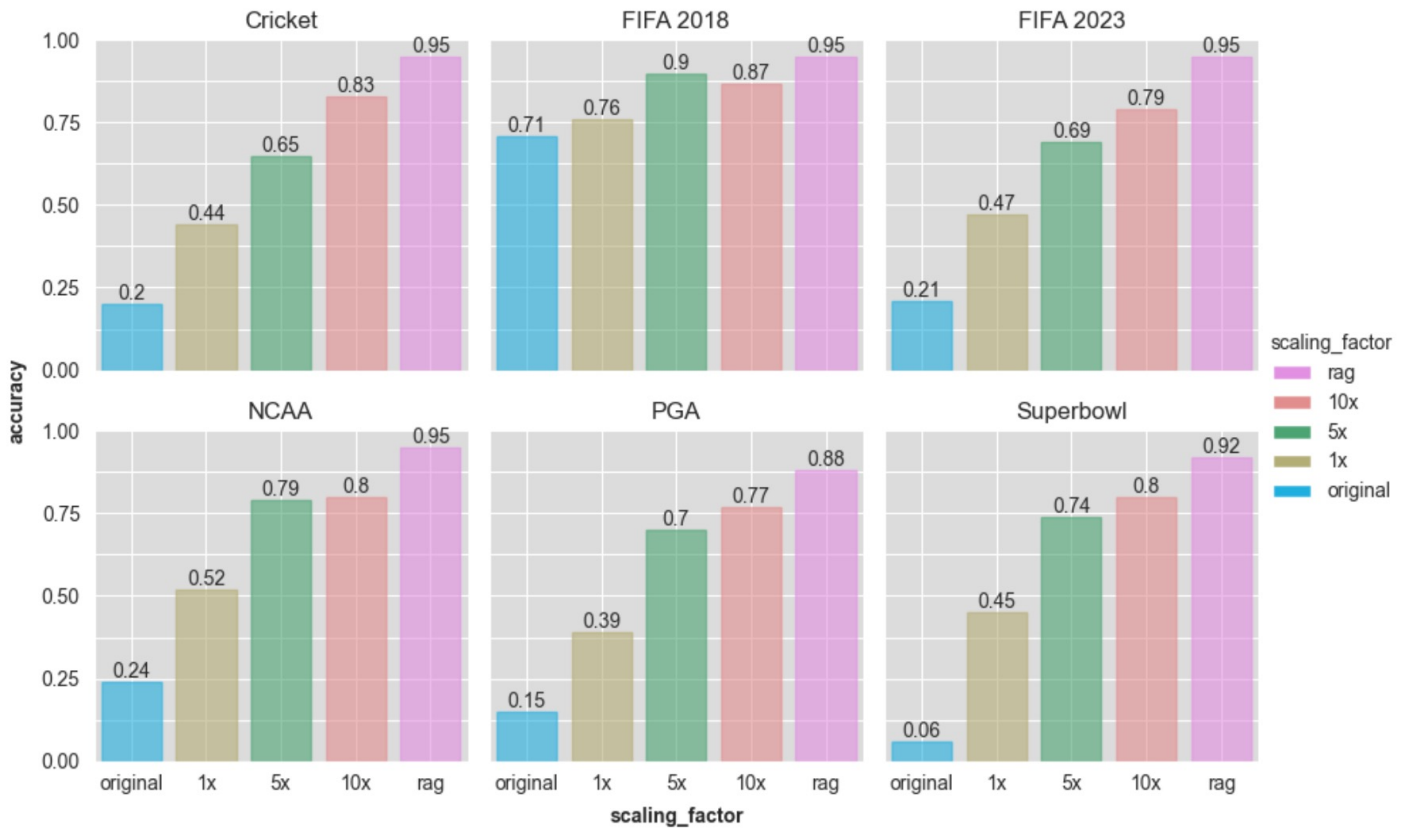


Figure 1: **Token-based evaluation set** accuracy for our six documents across 1x, 5x, and 10x scaling with models trained on **token-scaled datasets**. The base model results with no training are included under the bars annotated as "original," and we include a RAG baseline as well which leverages the cleaned document sections to answer the eval questions.

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

4. Results – Token-based Scaling

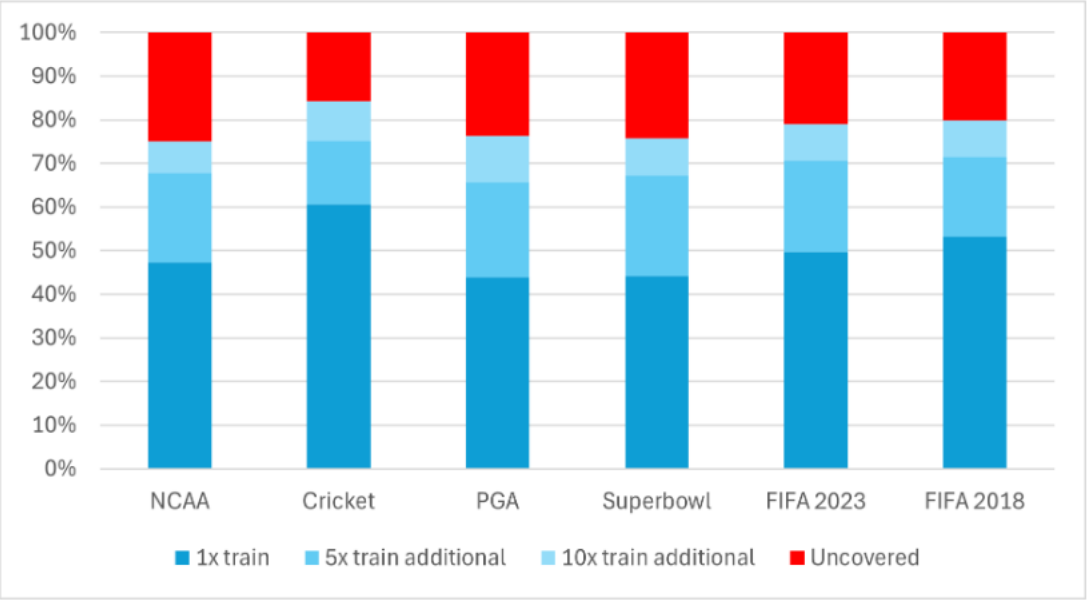


Figure 2: Fact coverage across token-based datasets.

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

4. Results – Fact-based Scaling

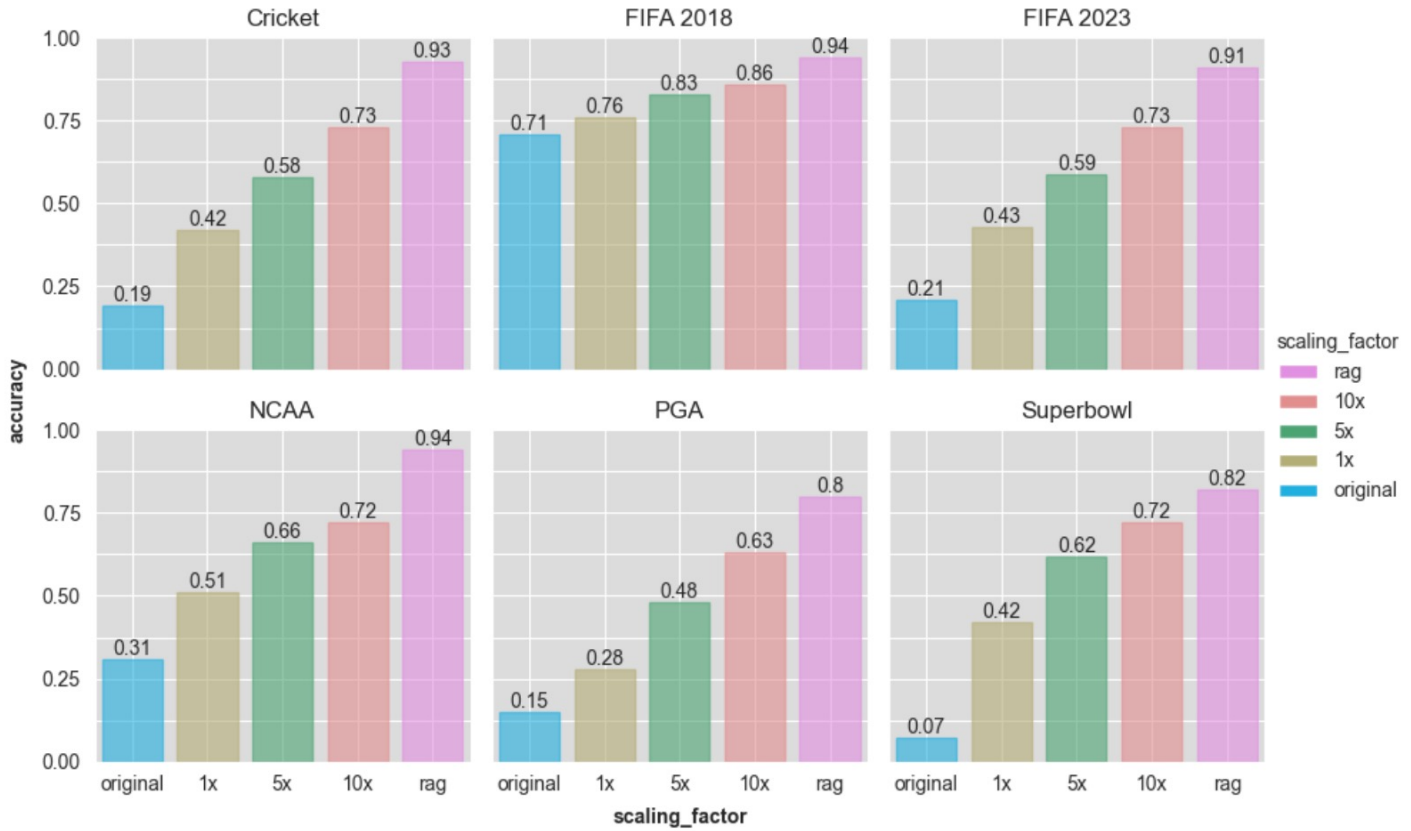


Figure 3: **Fact-based evaluation set** accuracy for our six documents across 1x, 5x, and 10x scaling with models trained on **fact-scaled datasets**. The base model results with no training are included under the bars annotated as original," and we include a RAG baseline as well which leverages the cleaned document sections to answer the eval questions.

Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning

4. Results – Token-based Scaling

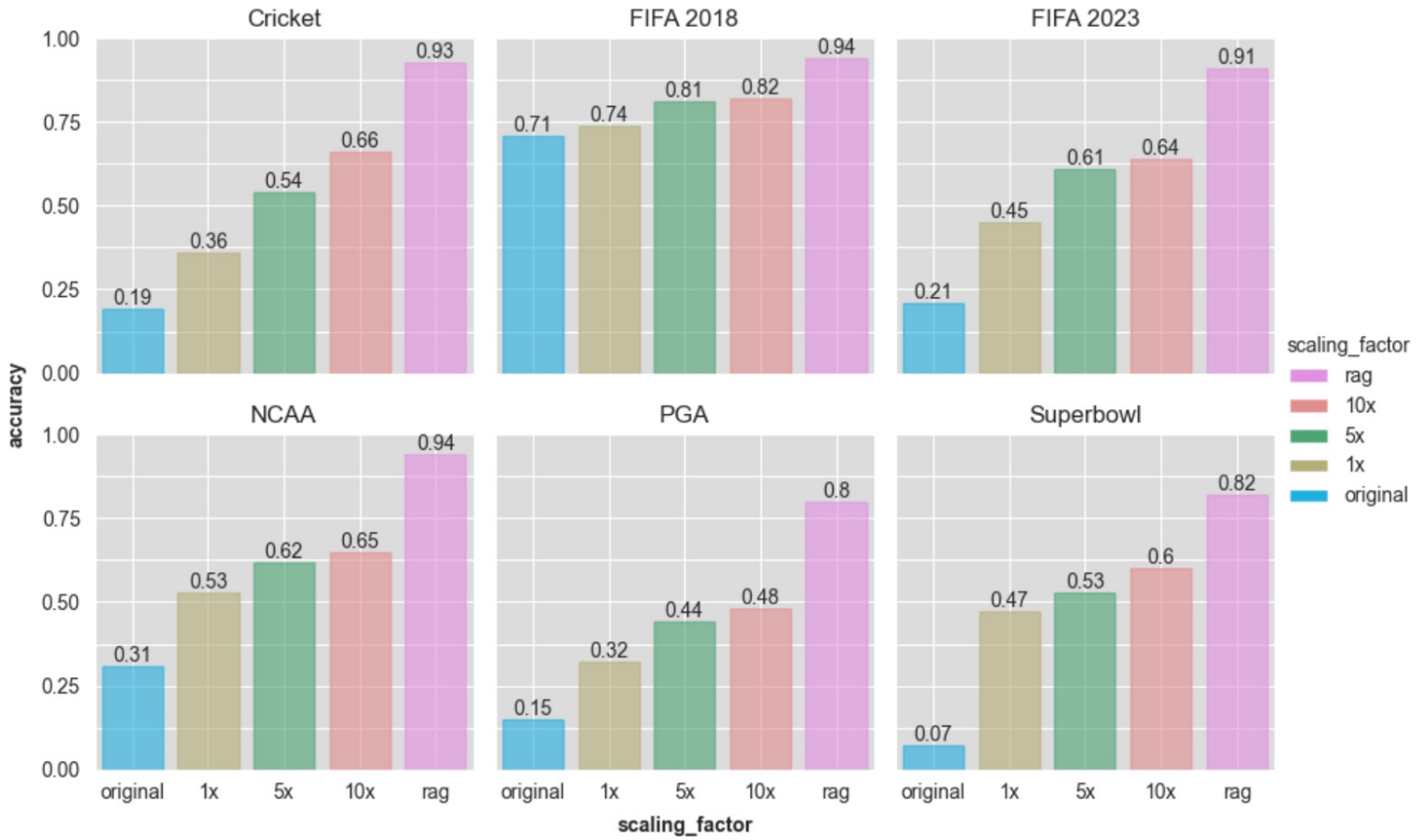


Figure 4: **Fact-based evaluation set** accuracy for our six documents across 1x, 5x, and 10x scaling with models trained on **token-scaled datasets**.

5. Conclusion

- SFT를 통해 보다 효과적으로 knowledge injection할 수 있는 data generation 방법론 제안
- Out-of-domain knowledge 관련 Q&A 태스크에서 성능 향상에 도움
- LLM의 domain (out-of-domain, new information) adaptation을 이해하고자 했음
- 해당 도메인의 factuality 향상 가능성을 보임

SEEKING NEURAL NUGGETS: KNOWLEDGE TRANSFER IN LARGE LANGUAGE MODELS FROM A PARAMETRIC PERSPECTIVE

Ming Zhong[1], **Chenxin An**[2], **Weizhu Chen**[3], **Jiawei Han**[1], **Pengcheng He**[3]
[1]University of Illinois Urbana-Champaign, [2]The University of Hong Kong, [3]Microsoft Azure AI
{mingz5, hanj}@illinois.edu, cxan23@connect.hku.hk
wzchen@microsoft.com, Herbert.he@gmail.com

ICLR 2024

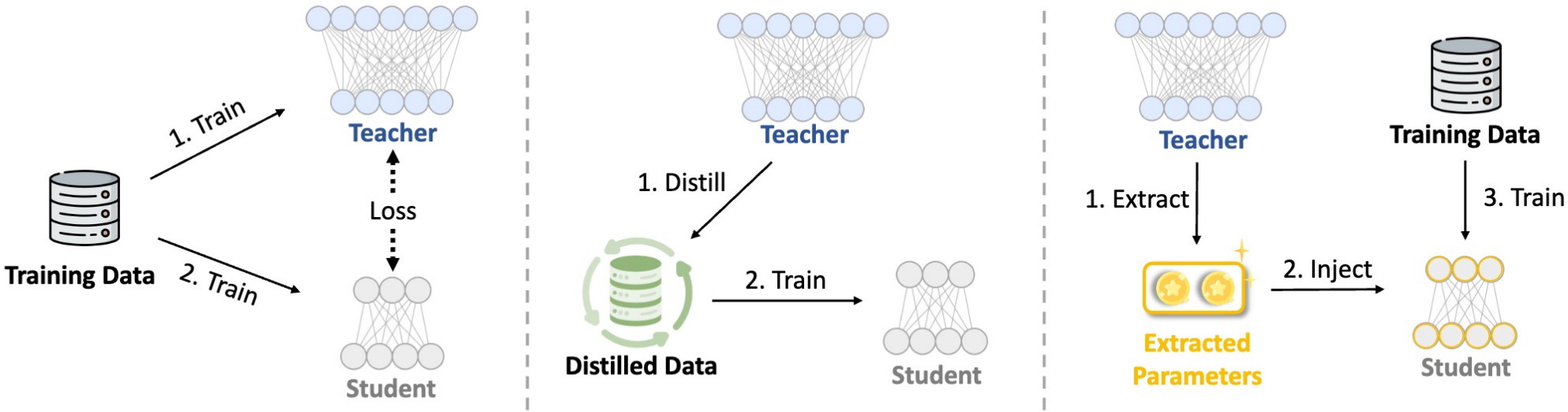
Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

1. Introduction

- LLM은 막대한 양의 corpora에 PT되면서 파라미터 안에 풍부한 지식을 내재적으로 가짐
- 이전 연구에서는 이러한 **파라미터에** 대해 **직접 작용**하여 **내부 지식을 조작**하고자 했음
  - Detection, editing, merge 등
- 다른 크기의 모델 간의 전이 가능성에 관련해서는 명확한 이해를 한 연구는 없었음
- 본 연구에서는 larger model에서 smaller model로의 parameter knowledge transfer를 실험적으로 연구함
- 특히 teacher의 task-specific parameter를 추출하여 이를 student에 transfer 가능한지 연구 – downstream task에 도움이 되는지 확인

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

2. Knowledge Transfer



(a) Online Distillation (b) Offline Distillation (c) Parametric Knowledge Transfer

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

3. Parametric Knowledge Transfer

- Task-specific parametric knowledge를 teacher로부터 선별적으로 student 모델에 transfer하는 것이 목표
- 태스크 $\mathcal{T}$가 주어졌을 때, teacher model M_T (parameter Θ_T) 로부터 student model M_S (parameter Θ_S) 에게 transfer 과정
- 1단계: extraction $\mathbf{Extract}(\mathbf{\Theta}_T; \mathbf{\Theta}_S; \mathcal{T}) = \mathbf{\Theta}_{T_{\text{extract}}},$
- 2단계: injection $\mathbf{Inject}(\mathbf{\Theta}_S; \mathbf{\Theta}_{T_{\text{extract}}}) = \mathbf{\Theta}'_S,$

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

3. Parametric Knowledge Transfer – Knowledge Extraction

- Parameter sensitivity: 특정 parameter를 0으로 세팅하고 loss의 변화 정도를 보고 sensitivity 측정
  - Parameter removal이 loss에 큰 영향을 줬다면 이 파라미터는 high sensitivity
- i번째 파라미터, j 번째 샘플 (in task $\mathcal{T}$) 에 대한 sensitivity:

$$S_{i,j} = \left| \mathbf{\Theta}_{T_i}^{\top} \nabla_{\mathbf{\Theta}_T} \mathcal{L}(x_j) \right| .$$

- S_i는 loss의 변화 정도를 제공:

$$\mathbf{\Theta}_{T_i}^{\top} \nabla_{\mathbf{\Theta}_T} \mathcal{L}(x_j) \approx \mathcal{L}(\mathbf{\Theta}_T) - \mathcal{L}(\mathbf{\Theta}_T - \mathbf{\Theta}_{T_i}).$$

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

3. Parametric Knowledge Transfer – Knowledge Extraction

- 모델간 레이어 수, 차원 크기가 다르기 때문에 layer selection과 dimensionality reduction을 sensitivity score에 기반하여 진행
- Teacher 모델의 layer 별 sensitivity 점수를 계산하고, 점수가 가장 높은 Top L_S (student 모델 레이어 수에 해당) 레이어를 고름
- 원래 sequential order를 유지하면서 student에 매핑
- 일반적으로 teacher가 student보다 큰 dimension을 가지기에 모든 2D matrix에 대해 parameter dimension 진행: (1) pass-by-neuron, (2) selecting rows and columns, **(3) direct extraction of submatrices**

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

3. Parametric Knowledge Transfer – Knowledge Injection

- 추출한 teacher parameters를 factorize한 후 student의 LoRA 모듈로 합침

$$\boldsymbol{W}_i^* = \boldsymbol{W}_i + \boldsymbol{B}_i \boldsymbol{A}_i,$$

frozen

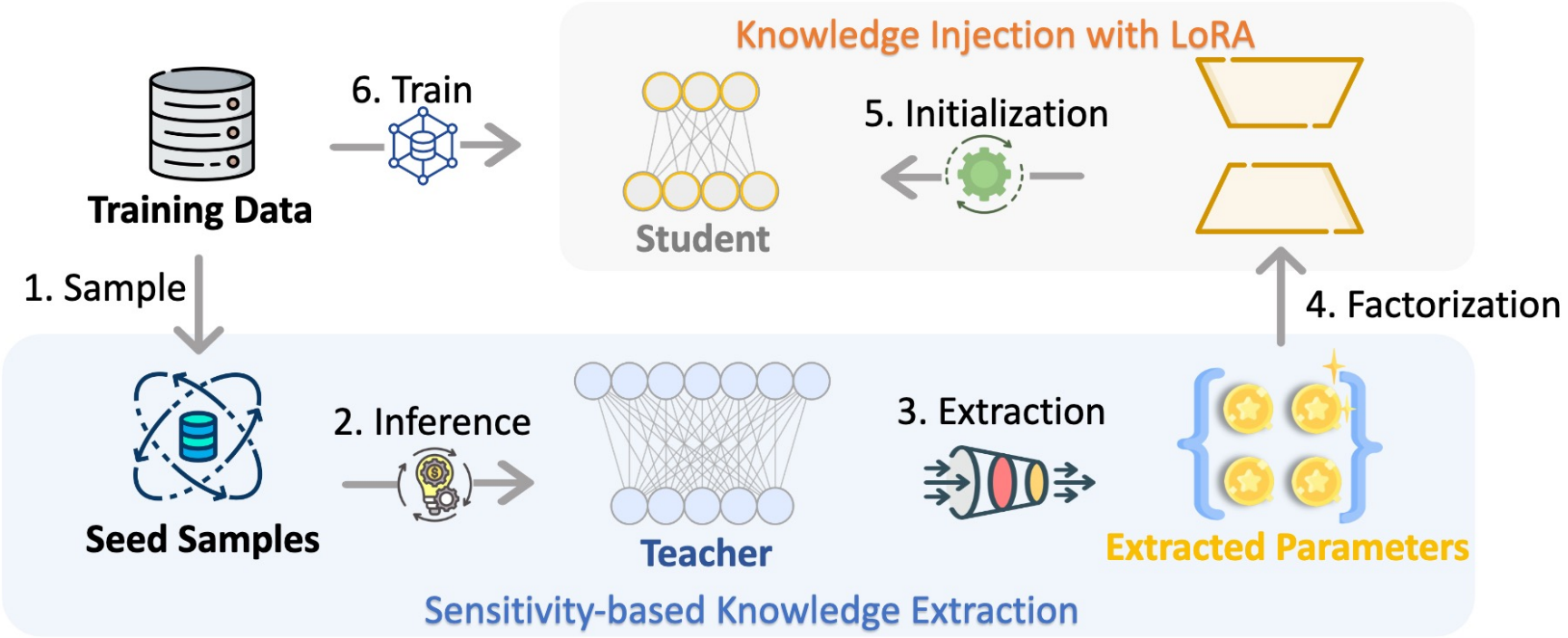
$$\boldsymbol{W}_i^* = \boldsymbol{W}_i - \boldsymbol{W}_{T_i,\text{extract},r} + \boldsymbol{B}_i \boldsymbol{A}_i,$$

Initially equivalent

- Student의 PT weight로부터 학습 시작하도록 보장
- LoRA 모듈은 extracted knowledge에서 효율적으로 가장 영향력 있는 feature들을 활용

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

3. Parametric Knowledge Transfer



Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

4. Experiments

- 네 가지의 벤치마크에 대해 유효성 확인: (1) reasoning (GSM), (2) professional knowledge (MMLU), (3) instruction-driven NLP tasks (Super NI), (4) open-ended conversation (AlpacaFarm)
- 큰 사이즈의 Llama 모델이 teacher 작은 모델이 student
- Student 학습을 위해 각 벤치마크마다 1,000 instances 랜덤샘플하여 사용
- LoRA는 rank 16, embedding layer, FFN, self-attention layer에 추가

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

4. Experiments

Table 1: Results for parametric knowledge transfer. "7B-LoRA + 13B Param." represents that we extract parameters from the 13B teacher model and transfer them to the 7B student model.

Models	GSM		MMLU		Super NI		AlpacaFarm	Average
	0-shot	8-shot	0-shot	5-shot	EM	R-L	Win Rate%	-
LLaMA-1								
Vanilla 7B	4.70	10.77	32.10	35.30	0.67	5.55	-	-
7B-LoRA	17.26	16.93	43.43	38.90	22.91	40.49	9.07	27.00
+ 13B Param.	**18.73**	**18.85**	44.03	39.77	24.51	42.37	9.28	28.22
+ 30B Param.	18.63	18.52	**45.20**	**40.60**	**25.01**	**43.08**	**9.40**	**28.63**
Vanilla 13B	4.93	17.44	43.50	46.80	2.18	7.78	-	-
13B-LoRA	26.18	23.78	50.43	50.03	27.34	45.53	13.91	33.89
+ 30B Param.	**27.85**	**27.70**	**51.30**	**51.03**	**27.51**	**46.09**	**17.27**	**35.54**
LLaMA-2								
Vanilla 7B	3.34	15.54	41.70	45.80	0.00	4.68	-	-
Vanilla 13B	6.52	27.82	52.10	55.20	0.00	4.84	-	-
7B-LoRA	23.38	21.05	47.77	**47.07**	24.93	41.25	20.50	32.28
+ 13B Param.	**25.30**	**26.31**	**49.37**	46.53	**26.16**	**42.98**	**24.64**	**34.47**

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

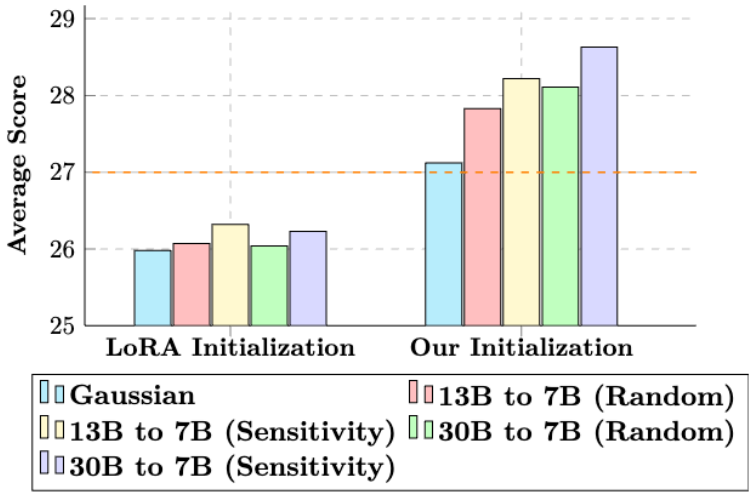
4. Experiments

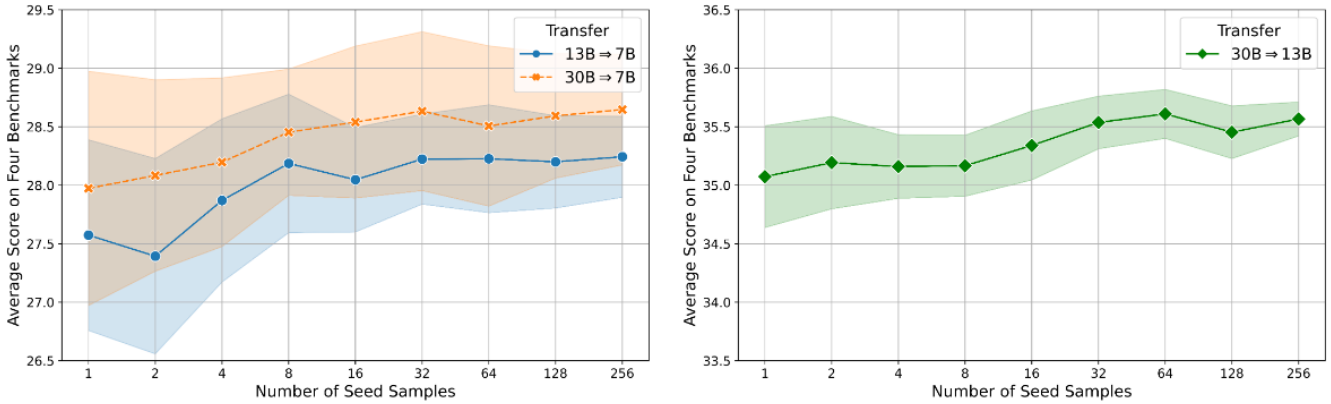
Table 2: Transfer experiments with different task-specific extracted parameters. The leftmost column indicates the dataset on which the knowledge extraction is based. The teacher model and student model are LLaMA-2 13B and 7B, respectively.

Models	GSM		MMLU		Super NI		AlpacaFarm	Average
	0-shot	8-shot	0-shot	5-shot	EM	R-L	Win Rate%	-
Vanilla 7B	3.34	15.54	41.70	45.80	0.00	4.68	-	-
7B-LoRA	23.38	21.05	47.77	47.07	24.93	41.25	20.50	32.28
GSM	**25.30**	**26.31**	48.40	45.97	24.45	42.11	23.68	33.75
MMLU	24.11	25.47	**49.37**	46.53	25.55	42.55	24.01	33.94
Super NI	23.78	24.11	48.60	46.70	**26.16**	**42.98**	24.31	33.81
LIMA	24.08	25.60	49.03	**47.23**	25.63	42.83	**24.64**	**34.15**

Seeking Neural Nuggets: Knowledge Transfer in Large Language Models from a Parametric Perspective

4. Experiments





5. Conclusion

- 다른 scale의 LLM간의 parametric knowledge transfer 연구
- Knowledge extraction & injection 방식 제안
- Sensitivity 기반 selection 방식과 dimension reduction 방식에 대한 다양한 실험을 통해 가장 효과적인 세팅 제안

Q&A

감사합니다.