Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs

Announcement Date	@2025년 1월 9일
Conference Name	EMNLP 2024
Keywords	Fine-Tuning LLM RAG

Abstract

- LLM은 다양한 도메인에 걸쳐 다양한 질문에 답할 수 있는 능력을 통해 이를 입증하고 있음
- 그러나, 이러한 지식은 학습 데이터의 특성에 크게 의존하기 때문에 한계가 존재
- 따라서 새로운 정보를 통합하거나 이전에 학습한 정보의 능력을 개선하기 위해 외부 데이터를 사용하는 것은 challenging 한 요소가 됨

- 본 연구에서는 두 가지 접근법인 <Unsupervised Finetuning > 과 <Retrieval-Augmented Generation, RAG > 를 비교
- 다양한 주제를 포함한 knowledge-intensive task에서 두 접근법을 평가

- 연구 결과 unsupervised finetuning이 일부 개선 효과를 제공하는 반면, RAG는 학습 중에 접했던 기존 지식과 완전히 새로운 지식 모두에서 일관되게 더 우수한 성능을 보였음
- 더욱이, LLM은 unsupervised finetuning을 통해 새로운 사실 정보를 학습하는 데 어려움을 겪는 경향이 있으며, 학습 중 동일한 사실의 다양한 변형에 노출되는 것이 이러한 문제를 완화할 수 있음을 발견

Introduction

- LLM은 대규모 Pre-Training 데이터셋 덕분에 다양한 도메인에서 놀라운 수준의 지식을 보여줌
- 그러나 이러한 지식에는 두 가지 significant limitation이 존재
  - 지식은 정적이며, 시간이 지남에 따라 업데이트되지 않음
  - non-specific 하기 때문에 특정 도메인에서 전문성이 부족할 수 있음
- 이러한 두 가지의 문제의 공통 해결책 : 모델의 지식을 향상시키는 것

- 텍스트 코퍼스 형태의 지식이 주어졌을때, pretrained model에 이 지식을 가르치는 최선의 방법이 뭘까?
  - pre-trained 모델에 지식을 추가하는 한 가지 방법은 파인튜닝임
    - 파인튜닝을 통해 우리는 모델의 훈련 과정을 이어가고 가중치를 조정해가며 task-specific data로 모델을 적응시킴
  - 다른 방법은 ICL (In-Context Learning) 을 이용하는 것임
    - ICL의 주요 아이디어는 모델의 가중치를 직접 변경하지 않고 입력 쿼리를 수정하여 pre-trained LLM의 새로운 작업에 대한 성능을 개선하는 것임
    - ICL 중 한 형태는 RAG임
    - RAG는 정보 검색 기술을 사용하여 LLM이 지식 출처에서 관련 정보를 얻고 이를 생성된 텍스트에 통합할 수 있도록 함

- <시험을 치는 대학생에 비유>
  - 세 명의 대학생이 있는데, 어떤 학생이 가장 높은 점수를 받을까?
    1. 시험 중에만 교과서를 참고하는 학생 (RAG)
       - 이 학생은 시험에 필요한 정보를 미리 공부하지 않고, 시험 도중 필요한 정보를 교과서에서 즉시 찾아서 사용
    2. 시험 전에 자료를 공부한 학생 (Unsupervised Fine-Tuning)
       - 이 접근법은 LLM이 훈련 중 추가 데이터를 활용하여 모델 가중치를 업데이트(학습) 하는 방식
    3. 시험 공지 후 자료 접근 불가한 학생 (일반 LLM)
       - 시험 공지가 나간 이후로 자료에 접근할 수 없고, 스스로 알고 있는 것만으로 시험을 치러야 하는 학생

        - 사전 학습된 지식(모델의 기본 가중치)에만 의존해야 함
    - 결론
        - 첫번째 학생 (RAG)이 시험 도중에도 교과서를 참조할 수 있어 가장 높은 성과를 낼 가능성이 큼
            - RAG는 실시간으로 외부 정보를 활용할 수 있으므로 새로운 질문이나 학습되지 않은 영역에도 강함
            - 두 번째 학생(unsupervised fine-tuning)은 사전 공부를 통해 어느 정도 대비할 수 있으나, 시험 중 새로운 정보에 대응하지 못함

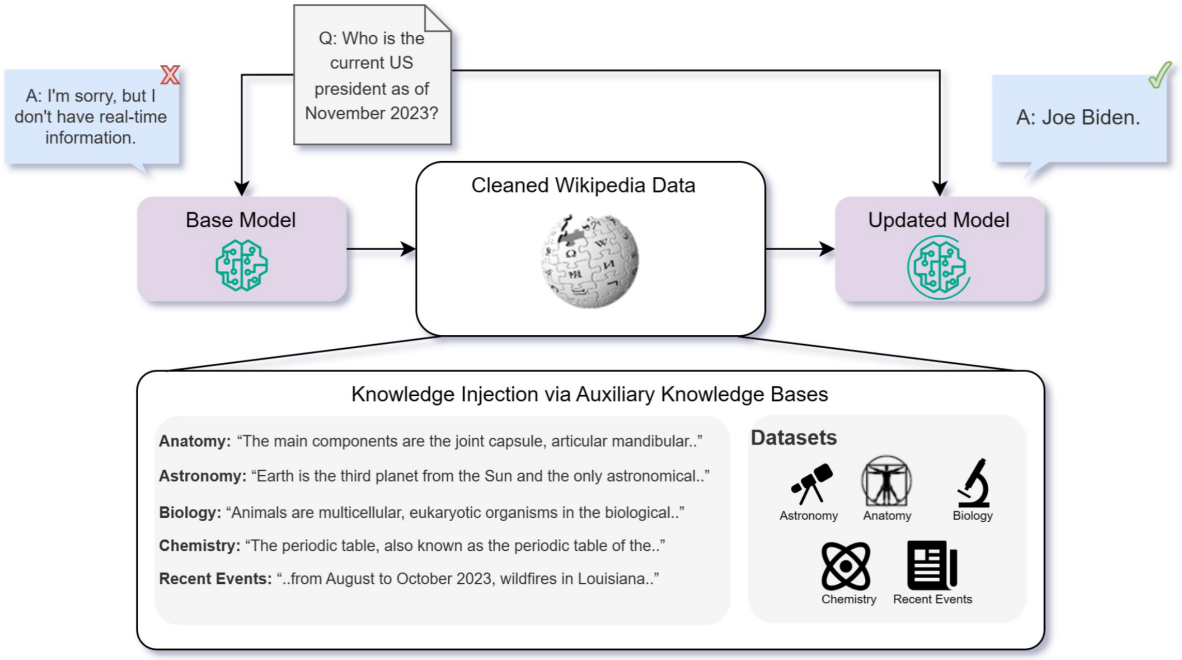

Figure 1. A visualization of the knowledge injection framework.

Knowledge Base Creation

1. Task 선택 및 이유
    - MMLU Benchmark (Massively Multilingual Language Understanding Evaluation)
        - 해부학, 천문학, 대학 생물학, 대학 화학 및 역사 주제 라는 네 가지 주제를 선택

        - 선택된 작업들은 사실 지식 (factual knowledge)에 중점을 두고 있음
    - Current Event Task
        - 최신 사건에 대한 다지선다형 질문을 제작
        - 2023년 8월~11월 사이에 발생한 미국의 최신 사건에 초점을 맞춤
        - 모델들이 해당 사실에 사전에 노출되지 않았음을 보장할 수 있음
2. 데이터 수집 및 전처리
    - Wikipedia에서 주제별로 관련 기사를 스크래핑하여 데이터셋을 수집
3. 최신 사건 작업 생성
    - GPT-4의 도움으로 새로운 다지선다형 데이터셋을 생성
    - 정답이 하나로 나오는 네 개의 구체적이고 고품질의 다지선다형 질문을 생성하도록 요청
    - 그 다음, GPT-4에게 생성된 네 가지 질문 중 가장 구체적인 두 가지를 선택하도록 요청
4. Paraphrases 생성
    - 데이터셋 생성 이후, GPT-4를 활용하여 데이터셋의 확장을 진행
    - GPT-4에게 원본 데이터를 완전히 동일한 정보를 유지하면서도 재구성된 패러프레이즈 버전을 생성하도록 요청
    - 이는 하이퍼파라미터 튜닝을 위한 validation set 로 사용하기 위해 별도로 저장-

Experiments and Results

- LLM의 knowledge-intensive task 성능을 평가하기 위해 **LM-Evaluation-Harness** 를 사용
- 이는 업계 표준이며, HuggingFace 리더보드의 기반이 되는 도구

- 추론 평가를 위해 LLAMA2-7B / Mistral-7B / Orca2-7B 세 가지 모델을 선택
- RAG 임베딩 모델로 bge-large-en 을 선택 (현 SOTA), 벡터 저장소로 FAISS 를 사용

- 기본 모델 / fine-tuned model / RAG / fine-tuned + RAG

- 5-shot 과 zero-shot 을 실험

- fine-tuning : 4개의 A100 GPU에서 5 epoch, batch 64

- 평가 방법 : 질문에 다지선다형 옵션을 추가한 뒤, 이를 모델에 전달하여 각 옵션에 대한 로그 확률 점수를 계산
- 가장 높은 점수를 가진 옵션을 모델의 선택으로 생각하고 정확도를 계산

Table 1. Results for the MMLU datasets described in Section 4.1 in terms of log-likelihood accuracy (Equation (4)).

Task	Model	Base model	Base model + RAG	Fine-tuned	Fine-tuned + RAG
Anatomy (0-shot)	Mistral 7B	0.556	**0.681**	0.570	0.659
	Llama2 7B	0.393	**0.489**	0.430	**0.489**
	Orca2 7B	0.607	**0.637**	0.600	**0.637**
Anatomy (5-shot)	Mistral 7B	0.600	**0.681**	0.622	0.674
	Llama2 7B	0.467	**0.563**	0.496	0.548
	Orca2 7B	0.570	0.659	0.593	**0.674**
Astronomy (0-shot)	Mistral 7B	0.625	0.678	0.651	**0.697**
	Llama2 7B	0.401	0.467	0.487	**0.520**
	Orca2 7B	0.645	**0.750**	0.651	**0.750**
Astronomy (5-shot)	Mistral 7B	0.658	**0.724**	0.651	0.697
	Llama2 7B	0.401	0.474	0.447	**0.520**
	Orca2 7B	0.664	**0.763**	0.664	0.743
College biology (0-shot)	Mistral 7B	0.681	0.757	0.701	**0.764**
	Llama2 7B	0.438	**0.493**	0.458	0.465
	Orca2 7B	0.583	**0.639**	0.604	0.632
College biology (5-shot)	Mistral 7B	0.722	**0.778**	0.736	0.771
	Llama2 7B	0.451	**0.521**	0.424	0.479
	Orca2 7B	0.604	**0.660**	0.625	0.653
College chemistry (0-shot)	Mistral 7B	0.470	**0.500**	0.490	**0.500**
	Llama2 7B	0.310	0.380	0.390	**0.390**
	Orca2 7B	0.370	**0.440**	0.370	0.390
College chemistry (5-shot)	Mistral 7B	0.470	**0.540**	0.500	0.500
	Llama2 7B	0.370	0.380	0.360	**0.390**
	Orca2 7B	0.430	**0.470**	0.370	0.380
Prehistory (0-shot)	Mistral 7B	0.713	**0.750**	0.719	0.731
	Llama2 7B	0.448	**0.481**	0.457	0.478
	Orca2 7B	0.642	**0.679**	0.673	0.673
Prehistory (5-shot)	Mistral 7B	0.722	**0.762**	0.725	**0.762**
	Llama2 7B	0.515	0.531	0.503	**0.537**
	Orca2 7B	0.664	**0.698**	0.667	0.694

Table 1 : MMLU의 전체 결과

Table 2. Current events results. Models that were fine-tuned on the original dataset are labeled as *FT-reg*, while those trained on the dataset with multiple paraphrases are labeled as *FT-par*.

	Base model	Base model + RAG	FT-reg	FT-par	FT-reg + RAG	FT-par + RAG
Mistral 7B	0.481	**0.875**	0.504	0.588	0.810	0.830
Llama2 7B	0.353	**0.585**	0.219	0.392	0.326	0.520
Orca2 7B	0.456	**0.876**	0.511	0.566	0.820	0.826

Table 2 : 최신 사건 작업 결과

- 결과적으로 MMLU와 최신 사건 작업 모두에서 RAG가 fine-tuning에 비해 유의미한 이점을 보임
- fine-tuning의 경우 기본 모델에 비해 결과를 개선했지만, RAG 접근법과 경쟁할 수 없었음

Conclusion

- 본 연구에서는 LLM이 새로운 지식 (전문적이거나, 완전히 새로운 지식) 에 적응할 수 있는 능력을 테스트하였음
- fine-tuning 과 RAG 를 같은 환경에서 두 접근법의 성능을 비교한 연구는 최초의 연구
- fine-tuning이 많은 사용 사례에서 유용할 수 있지만, RAG가 knowledge injection (지식 삽입)에 있어 더 신뢰할 수 있는 선택이라는 것을 발견함