

참고 문헌 기반 논문 요약 및 챗봇 구현

김경호(1조)

1. 개요

참고 문헌이 인용된 횟수를 분석 정교한 요약を提供하는 것

- 1) 논문 내 참고 문헌의 인용 횟수를 각각 분석하여 요약 반영
 - 가설: 인용이 많이 된 논문일수록 해당 논문을 이해하는 과정에 도움이 될 것
- 2) 최다 인용된 논문 5편에 대한 추가 요약 제공 및 인용 횟수 기반 점수화
- 3) 챗봇을 활용한 논문 질의응답 기능 구현 및 로컬 모델과의 비교 실험
 - **Bllossom/llama-3.2-Korean-Bllossom-3B** 모델을 챗봇에 붙여 어떤 차이가 나는지 확인

2. 주요 구성

범용적 프롬프트 구성 + RAG

- 1) One-Shot prompt 기법을 통한 범용적 프롬프트 구성
- 2) Rag를 기반으로 학습 없이 논문에 대한 챗봇 구현

"""

논문의 일부와 참고문헌 목록이 제공되었습니다.

논문 본문에서 특정 참고문헌이 인용되었다면, 해당 참고문헌의 **인용 횟수**를 `"Counter"`에 기록하고,

인용된 문장을 `"Context"`에 최대 300자 이내로 발췌하여 저장하세요.

한 번도 인용되지 않은 참고문헌은 출력에서 제외됩니다.

조건

`{condition}` => 이 곳에 인용 조건을 입력함으로써 다양한 인용 표현에 대해서도 작동하도록 설계

- 동일한 참고문헌이 여러 번 인용되면 `"Context"` 배열에 모든 인용 부분을 포함합니다.

- `"Context"`의 개별 항목은 **최대 300자 이하**를 유지해야 합니다.

출력 예시

```json

{

"10": {

"Title": "참고문헌 제목1",

"Counter": 2,

"Context": ["첫 번째 인용 문장", "두 번째 인용 문장"]

},

"3": {

"Title": "참고문헌 제목2",

"Counter": 1,

"Context": ["인용된 문장"]

}

}

```

2 Related Work

Recently, there has been a surge in interest in LLMs [9, 13, 21, 23, 49]. [23] and [27] reported that the performance of LLMs adheres to a scaling law, correlating positively with both the model size and the size of the pretraining corpus. Extensive studies have examined the knowledge encoded in the parameters of LLMs [36, 40]. [3], [15], [16], [19], [20], and [31] examined how language models learn and capture factual knowledge presented in training data. [4] demonstrated that knowledge

숫자 형태 인용일 경우

: 인용 표시: 논문 본문에서 참고문헌이
 [숫자] 형식 ([5], [27] 등) 으로
 인용된 경우만 추출하세요

et al 형태일 경우

: 인용 표시: 아래처럼 [이름 et al 연도] 로
 인용 되었음을 표시합니다.\n\n- 인용 예시: we
 evaluate pre-trained models on the
 downstream tasks\n\n GenomicsBenchmarks
 (Grešová et al. 2023)

For example, empirical studies from April 2024 observed that the frequency of certain words used in academic papers published in 2023 had changed and confirmed a strong correlation between these changes and the use of LLMs (Liang et al., 2024b; Geng and Trotta, 2024). Survey results also show that many researchers are utilizing LLMs in their work (Liao et al., 2024).

3. 서비스 시연

The screenshot displays the CitationLinkerGPT web application running on a local host (localhost:8502). The interface is divided into a left sidebar and a main content area.

Left Sidebar:

- home**
- CitationGPT** (selected)
- ImageGPT**
- arXiv 논문 ID 입력**: 2502.09606
- 사용할 모델**: gpt-4o-mini
- 최소 문장 길이**: 25 (range 10 to 100)
- 참고문헌 분석 비율**: 0.10 (range 0.10 to 1.00)
- 참고문헌 인용 조건**:
 - 인용 표시: 논문에서 (이름, et al, 연도) 형태로 인용이 되었습니다.
 - 예시: The coevolution of AI and humans has also been recognized by researchers (Pedreschi et al., 2024).

Main Content Area:

- CitationLinkerGPT** header
- Tabs: **Analyze Essay** (selected), Review based on References, CitationGPT
- 논문 분석 시작** button
- Progress Log:
 - ✓ 논문 분석이 완료되었습니다!
 - 1단계: 논문 검색 및 다운로드 중...
 - ✓ 논문 다운로드 완료! ⚡ 소요 시간: 0.87초
 - 2단계: 논문 전처리 진행 중...
 - ✓ 논문 전처리 완료! ⚡ 소요 시간: 0.44초
 - 3단계: 논문 요약 진행 중...
 - ✓ 논문 요약 완료! ⚡ 소요 시간: 9.93초
 - 4단계: 참고문헌 목록화 중...

The bottom of the image shows a Windows taskbar with various application icons and a system clock indicating 2:27 on 2025-02-15.



오늘 수업할 논문의 내용에 대해 요약해주세요



오늘 수업할 논문은 "Titans: Learning to Memorize at Test Time" 입니다. 이 논문은 Titans라는 새로운 신경망 아키텍처를 소개하고, 테스트 시에 학습하여 과거 정보를 기억하는 방법을 연구하였습니다. Titans는 단기 메모리와 장기 메모리를 결합하여, 더 나은 일반화 성능과 긴 컨텍스트를 처리할 수 있는 능력을 갖추고자 합니다. 연구에서는 Titans 아키텍처의 성능을 다양한 작업에서 평가하였고, 메모리 통합 방법의 효과를 분석하였습니다. 데이터는 FineWeb-Edu 데이터셋을 사용하였고, Titans는 메모리 업데이트 및 검색 과정을 통해 과거 정보를 효과적으로 활용합니다. 이 연구는 기존 Transformer 및 RNN 모델보다 더 효과적으로 긴 컨텍스트를 처리하며, 메모리 모델링 관점에서 중요한 기여를 한 것으로 나타났습니다.

openai 기반 챗봇 대화



논문의 주제를 요약해주세요

항해
PLUS



"Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution" 논문은 장기 시퀀스 데이터를 처리하는 데 도움이 될 수 있는 방법을 제안합니다. 이 방법은 기존의 장기 시퀀스 모델링 방법과 다르다는 점을 강조합니다. 논문은 장기 시퀀스 데이터를 처리하는 데 도움이 될 수 있는 방법을 제안합니다. 이 방법은 장기 시퀀스 데이터를 처리하는 데 도움이 될 수 있으며, 이는 기존의 장기 시퀀스 모델링 방법과 다르다는 점을 강조합니다.

로컬 모델 + 4b quantization 기반 챗봇 대화

4. 기술적 도전

다양한 인용 형태와 다양한 논문 형태를 반영하는 것

- 1) 참고 문헌 기반 요약을 구현하여 기존의 단순 요약보다 정보의 깊이를 강화
- 2) 다양한 논문 형식을 처리할 수 있도록 범용적인 설계 적용
- 3) **RAG** 기반 챗봇을 구축하여 로컬 모델과 OpenAI 모델 간의 성능 비교 실험 수행

5. 기술적 한계

긴 소요 시간 + 수식 이해 부족 + 로컬 언어 모델의 부족함

- 1) 분석 및 요약 과정에서 소요 시간이 길어(최대 10분) 실시간 처리가 어려움
- 2) AI 모델이 수식을 정확히 인식하지 못하여 논문의 깊이 있는 이해가 어려움
- 3) 로컬 모델 기반 챗봇의 성능이 **OpenAI** 모델에 비해 미흡함

6. 회고

rag 개념 + finetune 방법론 + 경량화

- 1) rag 기법에 대한 기본적 이해도 함양
- 2) 다양한 형태의 finetuning 비교(freeze 와 full finetune, lora 비교 등)
- 3) lora 방식과 quantization, 나아가 deepspeed 기법을 통한 훈련 경험
- 4) SFT 방식을 통한 언어 모델 IT 훈련