onlybooks / llm

Code Issues 9 Pull requests 1 Actions Projects Security Insights

LLM을 활용한 실전 AI 애플리케이션 개발

www.onlybook.co.kr/entry/llm

187 stars 152 forks 6 watching Branches Activity
Tags

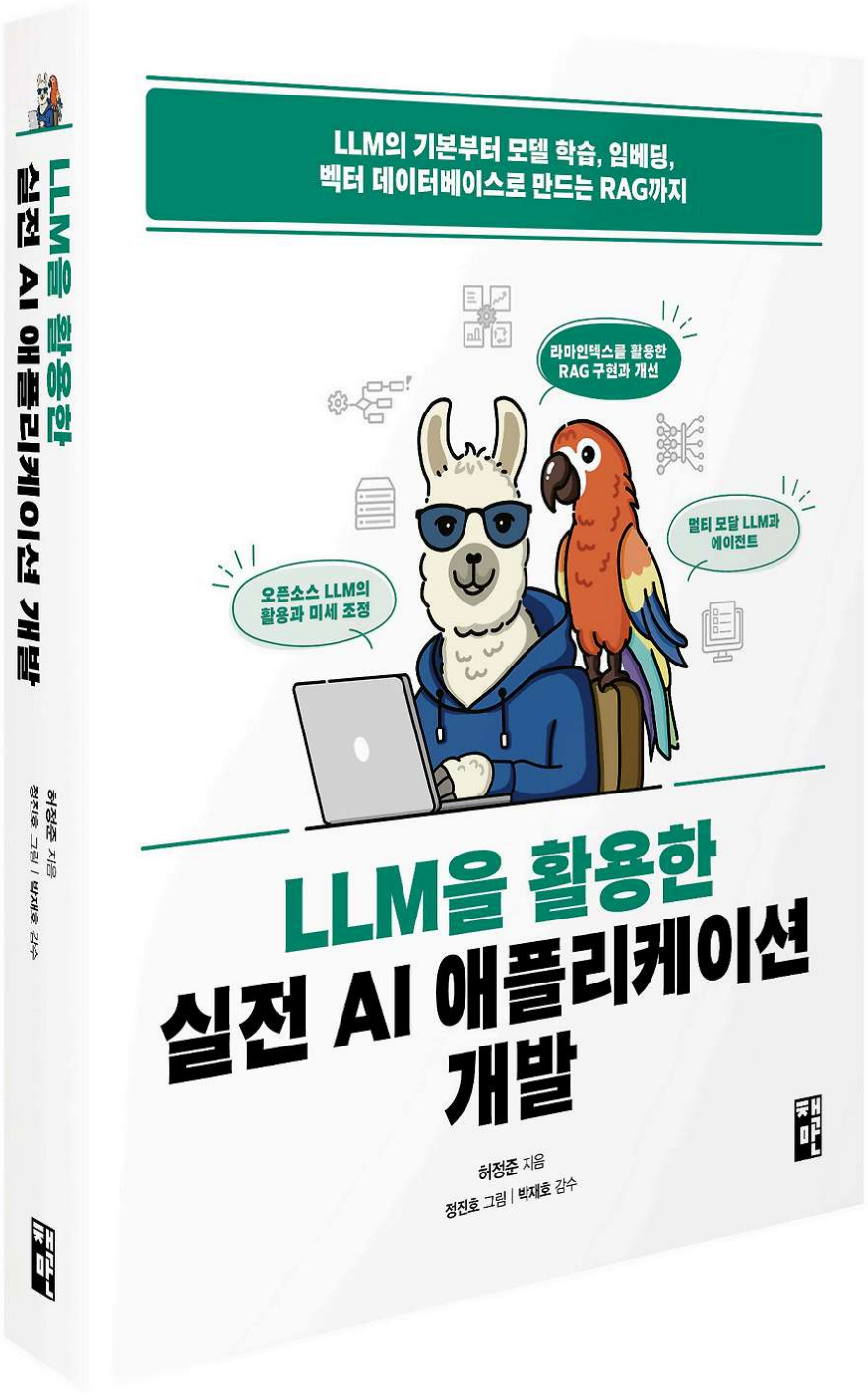
Public repository

1 Branch 0 Tags Go to file t Go to file Add file Code

onlybooks Update README.md b672610 · 7 months ago

02장	decoder input error correction	9 months ago
03장	upgrade library version	9 months ago
05장	upgrade library version	9 months ago
06장	fix/variable-typo	last year
07장	Add files via upload	last year
08장	세션 재시작 안내 추가	last year
09장	필요한 라이브러리 추가	last year
10장	Add files via upload	last year
11장	Add files via upload	last year
12장	Add files via upload	last year
14장	Add files via upload	last year
15장	Add files via upload	last year
16장	Add files via upload	last year
README.md	Update README.md	7 months ago

README



LLM을 활용한 실전 AI 애플리케이션 개발

LLM의 기본부터 모델 학습, 임베딩, 벡터 데이터베이스로 만드는 RAG까지

허정준 지음 | 정진호 그림 | 박재호 감수

556쪽 | 32,000원 | 2024년 7월 25일 출간 | 184*24*027 | ISBN 9791189909703 (93000)

트랜스포머 아키텍처부터 RAG 개발, 모델 학습, 배포, 최적화, 운영까지

라마인덱스(LlamaIndex)와 LLM을 활용한 AI 애플리케이션 개발의 모든 것

★ 정오사항 확인: https://www.onlybook.co.kr/entry/llm-errata

이 책에서는 LLM의 기본 아키텍처에서 출발해 애플리케이션의 요구사항에 맞춰 LLM을 길들이고 제한된 컴퓨팅 환경에서 동작하게 경량화해서 원활하게 서빙하게끔 기초를 다진 다음에 RAG라는 LLM의 대표적인 애플리케이션을 만드는 방법을 차근차근 설명한다. 또한, 실제 운영과정에서 부딪히는 어려움을 해소하는 방법과 멀티 모달과 더불어 에이전트와 같은 고급 주제까지 다룬다. LLM 시대를 맞이해 필수적으로 갖춰야 하는 개발 지식을 이론과 실무 양쪽 관점에서 설명하므로, 새로운 패러다임에 적응하고자 하는 개발자들에게 가뭄의 단비와도 같을 책이다.

깃허브 실습 코드 다운로드

실습 코드는 책의 깃허브 저장소(https://github.com/onlybooks/llm) 에서 확인할 수 있습니다. 깃허브의 코드는 구글 코랩에서 두 가지 방법으로 활용할 수 있다.

1. 로컬에서 업로드하기: 깃허브의 코드를 로컬 환경에 클론하거나 압축 파일 형태로 내려받은 후 진행하려는 실습 폴더의 노트북 파일(ipynb)을 구글 코랩에서 열어 실습을 진행할 수 있다.
2. 깃허브 URL로 열기: 구글 코랩에서 노트 열기(Ctrl+O)를 선택하면 다양한 노트 열기 방식 중 깃허브GitHub 탭에서 코드의 URL을 통해 실습 노트북을 열 수 있다.

| 이 책의 코드 실행 환경 |

이 책의 실습은 구글 코랩에서 실행한다. 구글 코랩은 구글에서 제공하는 노트북 실행 환경으로, 파이썬의 주피터 노트북과 유사한 UI로 브라우저에서 실행할 수 있다. 또 구글 코랩에서는 무료로 T4 GPU(16GB)를 사용할 수 있도록 제공한다. 구글 코랩의 무료 버전은 12시간의 런타임 제한이 있으며, 장시간 사용하지 않으면 연결이 끊길 수 있다.

지은이 허정준

서울대학교 기계항공공학부를 졸업하고 롯데면세점 빅데이터팀 데이터 분석가를 거쳐 현재는 프리랜서 마켓 크몽에서 AI 엔지니어로 일하고 있다. 『파이토치 라이트닝으로 시작하는 딥러닝』을 번역했으며, 최근에는 LLM을 활용한 어시스턴트(에이전트) 개발에 관심이 많다.

차례

8장 sLLM 서빙하기

8.1 효율적인 배치 전략
__8.1.1 일반 배치(정적 배치)
__8.1.2 동적 배치
__8.1.3 연속 배치
8.2 효율적인 트랜스포머 연산
__8.2.1 플래시어텐션
__8.2.2 플래시어텐션 2
__8.2.3 상대적 위치 인코딩
8.3 효율적인 추론 전략
__8.3.1 커널 퓨전
__8.3.2 페이지어텐션
__8.3.3 추측 디코딩
8.4 실습: LLM 서빙 프레임워크
__8.4.1 오프라인 서빙
__8.4.2 온라인 서빙
8.5 정리

[3부] LLM을 활용한 실전 애플리케이션 개발

9장 LLM 애플리케이션 개발하기

9.1 검색 증강 생성(RAG)
__9.1.1 데이터 저장
__9.1.2 프롬프트에 검색 결과 통합
__9.1.3 실습: 라마인덱스로 RAG 구현하기
9.2 LLM 캐시
__9.2.1 LLM 캐시 작동 원리
__9.2.2 실습: OpenAI API 캐시 구현
9.3 데이터 검증
__9.3.1 데이터 검증 방식
__9.3.2 데이터 검증 실습
9.4 데이터 로깅
__9.4.1 OpenAI API 로깅
__9.4.2 라마인덱스 로깅
9.5 정리

10장 임베딩 모델로 데이터 의미 압축하기

Releases

No releases published

Packages

No packages published

Contributors 3


onlybooks 책만


shangrilar Jungjun Hur


yunjihyung

Languages


Jupyter Notebook 98.1%
Python 1.9%