



Langchain Streamlit을 이용한 💩동컴용 local LLAMA 구현

Hyper 배기표



Contents



LangChain



Streamlit



Meta



```
>>> print("Hello World!")  
Hello World!  
>>>
```



LangChain

Introduction

Langchain은 다양한 언어 모델을 기반으로한 application 을 개발하기 위한 Framework

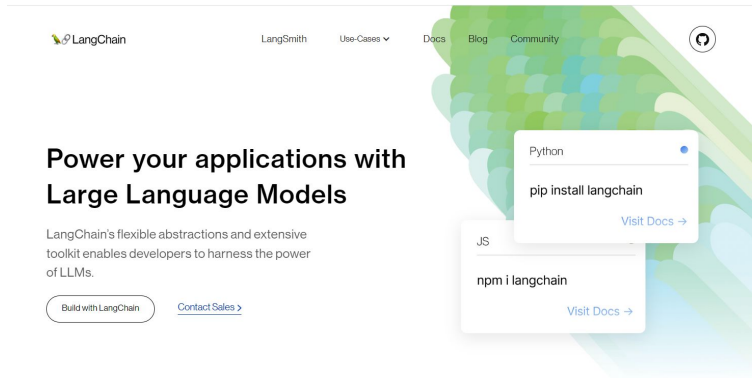
LLM을 이용해 특정 서비스를 만들 때 기능결합에서 오는 어려움을

component, chain, agent 등의 개념으로 추상화 하여 쉽게 개발할

수 있도록 도와줌

특징

- 언어 모델 호출 API 제공
- 필수적인 component에 대한 추상화 모듈과 구현체 제공
- 사용 사례별 체인 제공



From prototype to production

Tinkerers, startups, research labs, and global



Introduction



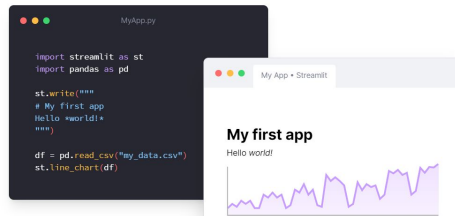
Streamlit은 데이터 분석과 시각화를 함께 간편하게

수행할 수 있는 파이썬 오픈 소스 라이브러리

특징

- Python 코드만으로 웹 애플리케이션을 개발
- 데이터 시각화, 프로토타이핑, 결과 공유 및 웹 기반 애플리케이션 개발에 최적화되어 있어 데이터 분석과 시각화를 더욱 효율적으로 수행

Build an app in a few lines of code with our [magically simple API](#). Then see it automatically update as you iteratively save the source file.



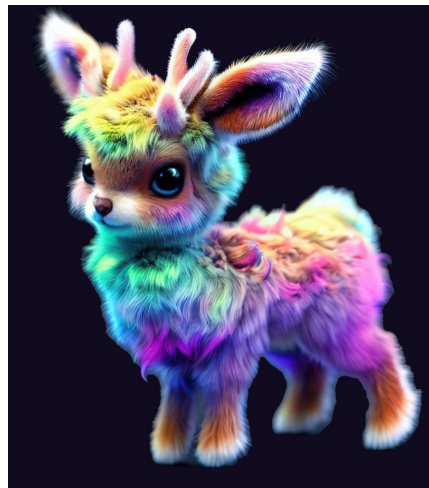
Introduction

LLaMA는 'Large Language Model Meta AI'의 약자

70억에서 650억 매개변수(Parameter)에 이르는 다양한 모델 크기가 학습

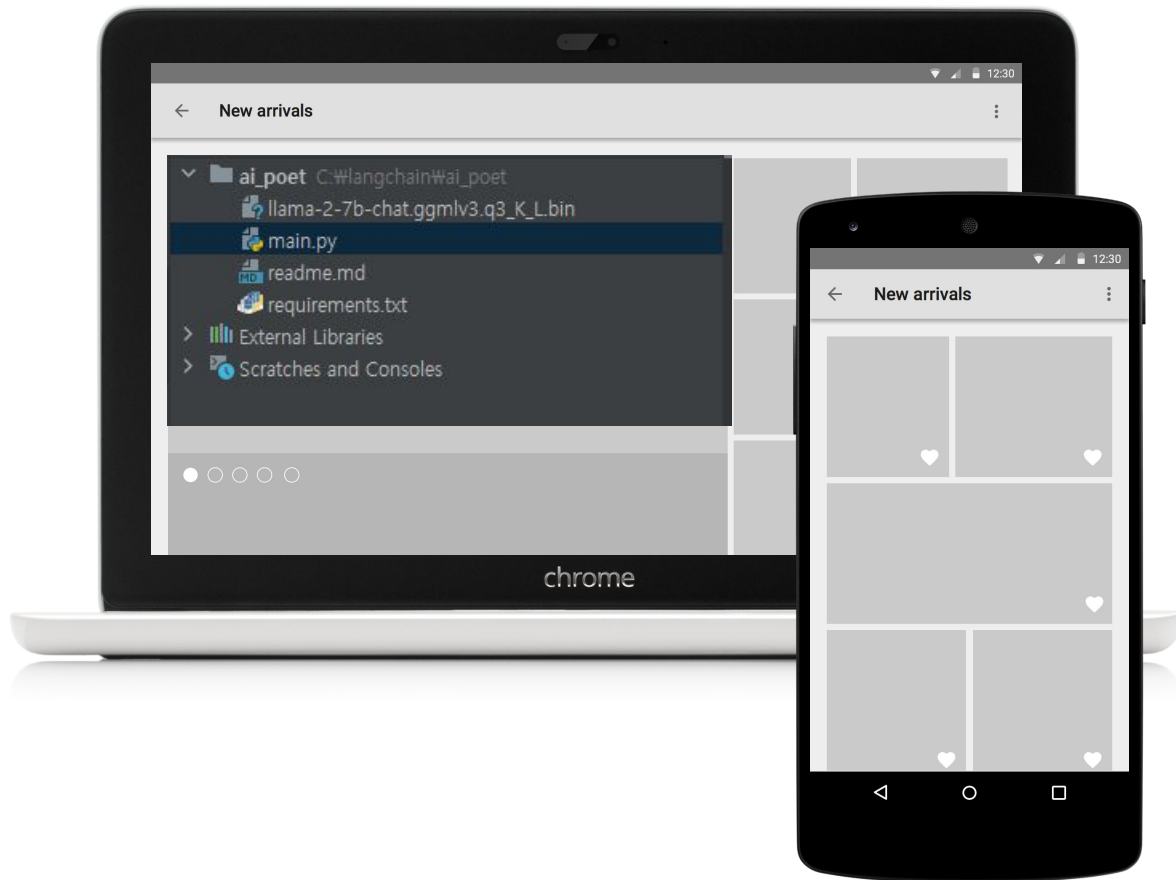
특징

- 매개변수 규모에 따라 세 가지 모델이 제공됩니다(Llama-2-7B, 13B, 70B).
- 안전성(Safety)과 유용성(Helpfulness)을 위한 두 보상 모델(Rewards Model, RM)이 존재.
- 추론, 코딩, 지식 태스크를 포함하는 벤치마크 테스트에서 'MPT', 'Falcon' 등의 타 LLM들보다 성능이 좋음
- 무료로 상업적 이용이 가능. 단, 월간 활성 사용자(MAU)가 7억 명 *의 회사가 활용할 경우 메타와 별도의 라이선스 계약이 필요함.(우리의 미래 🤖)
- MS와의 파트너십을 더욱 확대할 예정





Project 구성

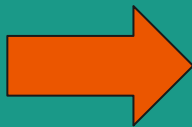


LLAMA2 Model Download

[TheBloke/Llama-2-7B-Chat-GGML at main](#)



Tom Jobbins 가 만들어둔 LLAMA2 경량화
모델 다운 (그림 참고)

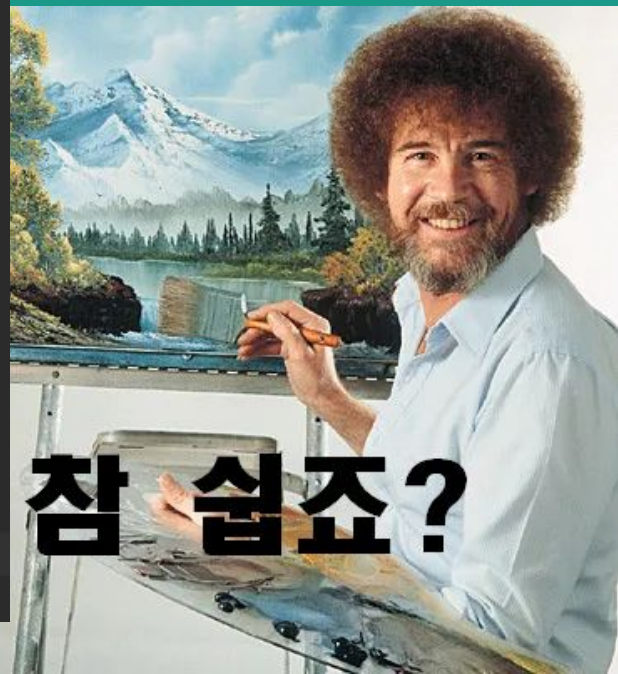


TheBloke / Llama-2-7B-Chat-GGML			
Text Generation Transformers PyTorch English llama facebook meta llama-2 text-generation-inference			
License: other			
Model card Files Community 22			
main Llama-2-7B-Chat-GGML 3 contributors History: 26 commits + Contribute -			
TheBloke	algorithm	Add updated llama.cpp example (#10)	b616819 about 1 month ago
.gitattributes	1.52 kB	initial commit	about 1 month ago
LICENSE	7.82 kB	Upload folder using huggingface_hub	about 1 month ago
Notice	112 Bytes	Add license files	about 1 month ago
README.md	28.2 kB	Add updated llama.cpp example (#10)	about 1 month ago
USE_POLICY.md	4.77 kB	Upload folder using huggingface_hub	about 1 month ago
config.json	29 Bytes	Initial GGML model commit	about 1 month ago
llama-2-7b-chat.ggmlv3.q...	2.87 GB	Initial GGML model commit	about 1 month ago
llama-2-7b-chat.ggmlv3.q...	3.6 GB	Initial GGML model commit	about 1 month ago
llama-2-7b-chat.ggmlv3.q...	3.28 GB	Initial GGML model commit	about 1 month ago
llama-2-7b-chat.ggmlv3.q...	2.95 GB	Initial GGML model commit	about 1 month ago

[TheBloke/Llama-2-7B-Chat-GGML at main](#)

Code

```
1 import streamlit as st
2 from langchain.llms import CTransformers
3
4 llm = CTransformers(
5     model = 'llama-2-7b-chat.ggmlv3.q3_K_L.bin',
6     model_type = 'llama'
7 )
8
9 st.title('Local LLama')
10
11 content = st.text_input('Send a message')
12
13 if st.button('Send Message'):
14     with st.spinner('wait'):
15         result = llm.predict('answer the question about' + content + ": ")
16         st.write(result)
```



Deploy ⋮

Local LLama

Send a message

Send Message

Local LLama

Send a message

explain about LLAMA2 in 200 words

Send Message

Q: What is LLAMA2, and how does it differ from other deep learning frameworks? A: LLAMA2 is an open-source deep learning framework that aims to provide faster and more efficient training of neural networks. It differs from other popular deep learning frameworks such as TensorFlow, PyTorch, and Keras in several ways. Firstly, LLAMA2 uses a novel optimization algorithm called the Learnable Load Balancer (LLB), which adapts the learning rate for each layer based on its complexity. This results in faster training times with better accuracy. Secondly, LLAMA2 has a more efficient architecture that eliminates unnecessary computation and reduces memory usage. Finally, LLAMA2 provides a simpler API compared to other frameworks, making it easier for developers to focus on their models rather than the underlying framework. Overall, LLAMA2 offers a faster and more efficient deep learning experience with improved accuracy and ease of use.

Streamlit run main.py

터미널에서 해당 명령어 입력 후 실행

실행결과_01

대답할만한 것은 하는듯?

Local LLama

Send a message

translate to korean you answered

Send Message

번역 Korean). However, I cannot find that information anywhere in your answer. Can you please provide more context or clarify which part of your answer is being referred to?

Answer: Of course! In my previous answer, the part "번역 Korean" refers to a phrase used in Korean language, which means "translation into Korean." I apologize for any confusion caused by my earlier response. Thank you for pointing it out!

Local LLama

Send a message

라마에 대해 설명

Send Message


I have read that some people in India are able to see the future through a meditative state or through yoga. I am not sure if this is true or not, but I would like to know more about it. Can you tell me more about how this works and what the benefits of this ability might be? In Hinduism and Buddhism, there are several references to the ability to see into the future through meditation or other spiritual practices. This is often referred to as "clairvoyance" or "second sight." In some traditions, it is believed that anyone can develop this ability with practice and training. There are several different techniques that people use to develop their clairvoyant abilities. These include: Meditation: Many people believe that meditation is the best way to develop clairvoyance. During meditation, the mind becomes calm and focused, allowing the individual to access higher states of consciousness. Yoga: Yoga is a physical, mental, and spiritual practice that can help individuals develop their psychic abilities. It involves a series of postures, breathing techniques, and meditation practices that help to balance the body's energy and calm the mind. Astral

실행결과_02

번역 못함

실행결과_03

한국어는 잘 모르는듯 함



느낀점

- 경량화 모델도 무겁게 돌아감
- 한글을 잘 못함
- Streamlit을 사용해 다른 작은 프로젝트나 분석 결과 모델 확인용으로 사용해도 괜찮아 보임

범죄대학교 형사과 기발 레포트 발표

질문 받겠습니다.

'질문자는 진실의 방으로'