

LM Studio - 허깅페이스 LLM(+ obsidian)

발표 주제

“Reachy Mini를 위한 LLM 기반 자연어 명령 인터페이스 구성 실험”

- LM Studio + Hugging Face 모델을 통해
- “로컬 PC에서 자연어 명령 처리”
- → (향후 목표) 이를 Reachy의 행동 제어로 연결할 수 있는 구조 실험

발표 구성

1. LLM을 로컬에서 쓰는 이유

- 인터넷 연결 없이, 빠르게 응답 가능한 명령 인터페이스 구축


2. LM Studio로 Hugging Face 모델 실행

- Hugging Face에서 오픈소스 LLM모델 중 GGUF파일 모델 다운

GGUF(Georgi Gerganov Unified Format)는

딥러닝 모델을 효율적으로 저장하고 배포하기 위한 새로운 파일 형식으로 CPU, GPU, TPU 등 다양한 플랫폼에서 모델을 실행할 수 있도록 지원함.

모델 다운로드

 **Hugging Face**

[Models](#) [Datasets](#) [Spaces](#) [Community](#) [Docs](#) [Enterprise](#) [Pricing](#) [Log In](#) [Sign Up](#)

Main Tasks Libraries Languages Licenses Other

Tasks

Text Generation

Any-to-Any

Image-Text-to-Text

Image-to-Text

Image-to-Image

Text-to-Image

Text-to-Video

Text-to-Speech

+ 42

Parameters

<1B

6B

12B

32B

128B

>500B

Libraries

PyTorch

TensorFlow

JAX

Transformers

Diffusers

Safetensors

ONNX

GGUF

Transformers.js

MLX

Keras

+ 41

Apps

vLLM

TGI

llama.cpp

MLX LM

LM Studio

Ollama

Jan

+ 12

Inference Providers

Novita

Cerebras

Nebius AI

Featherless AI

Together AI

Fireworks

Groq

Hyperbolic

+ 6

Models 3,878

Full-text search

Sort: Trending

beomi/Llama-3-Open-Ko-8B

Text Generation • 8B • Updated May 20, 2024 • 8.02k • 151

KoboldAI/LLaMA2-13B-Tiefighter

Text Generation • Updated Oct 20, 2023 • 1.88k • 92

KoboldAI/LLaMA2-13B-Tiefighter-GGUF

13B • Updated Jan 27 • 150k • 102

torchtorchkimtorch/Llama-3.2-Korean-GGACHI-1B-Instr...

1B • Updated May 29 • 1.28k • 8

Bllossom/llama-3.2-Korean-Bllossom-AICA-5B

Image-to-Text • 5B • Updated Mar 14 • 750k • 92

Kotokin/sophosympatheia_New-Dawn-Llama-3.1-70B-v1.1...

Text Generation • Updated Aug 16, 2024 • 3 • 2

Kororinpa/LLAMA_peft_stack

Updated Apr 11, 2023

Kororinpa/Stack-LLAMA-merged-Adapter

Text Generation • Updated Apr 15, 2023 • 3

jinooring/llama-ko-alpaca-lab-001

Updated Apr 18, 2023

Koantek/Dolly_Llama_v1

Updated Jun 19, 2023 • 2

Koantek/Dolly_Llama

Updated Jun 17, 2023

Koantek/dolly_llama-v2

Updated Jun 22, 2023 • 2

Bllossom/llama-3.2-Korean-Bllossom-8B

Text Generation • 8B • Updated Dec 18, 2024 • 245k • 350

MLP-KTLim/llama-3-Korean-Bllossom-8B

Text Generation • 8B • Updated Dec 18, 2024 • 32.6k • 184

Bllossom/llama-3.2-Korean-Bllossom-3B

Text Generation • 3B • Updated Dec 16, 2024 • 32.6k • 184

Konnect1221/The-Inception-Presets-Methception-LLama...

Updated Feb 3 • 118

beomi/KoAlpaca-llama-1-7b

Text Generation • Updated Mar 21, 2023 • 861 • 28

Kororinpa/Stack-LLAMA-Gpt-Neox-reward-model

Text Classification • Updated Apr 15, 2023 • 6

Kororinpa/Stack-LLAMA-GPT_Neox_reward_model

Text Classification • Updated Apr 16, 2023 • 6

Mission Control

Model Search

Runtime

Hardware

Hugging Face에서 모델 검색...

GGUF

Staff picks

Qwen3 30B A3B 2507

Updated version of Qwen3-30B-A3B featuring significant improvements in general capabilities includ... 2 days ago

Qwen3 Coder 480B

Qwen's most powerful code model, featuring 480B total parameters with 35B activated through Mist... 8 days ago

Qwen3 235B A22B 2507

Updated version of Qwen3-235B-A22B featuring significant improvements in general capabilities inc... 10 days ago

Llm2 1.2B

Hybrid architecture model intended for local use, by Liquid AI 15 days ago

Ernie 4.5 21B A3B

Medium-size mature-of-experts model from Baidu's new Ernie 4.5 line of foundation models 21 days ago

DeVstral Small 2507

DeVstral excels at using tools to explore codebases and editing multiple files to power software eng... 21 days ago

Gemma 3n E4B

Gemma 3n is a multimodal generative AI model optimized for use in everyday devices, such as pho... 23 days ago

Mistral Small 3.2

Update to Mistral Small 3.1 with better instruction following, fewer infinite generation issues, and an... 41 days ago

Magistral Small

MistralAI's first reasoning model, based on Mistral Small 3.1 51 days ago

Deepseek R1 0528 Qwen3 8B

Distilled version of the DeepSeek-R1-0528 model, created by continuing the post-training process o... 63 days ago

DeVstral Small 2505

DeVstral by MistralAI is based on Mistral Small 3.1. Debuts as the #1 open source model on SWE-b... 71 days ago

Phi 4 Mini Reasoning

Lightweight open model from the Phi-4 family 82 days ago

Phi 4 Reasoning Plus

google/gemma-3n-e4b

4.38k • 22

Last updated 23 days ago

Gemma 3n is a multimodal generative AI model optimized for use in everyday devices, such as phones, laptops, and tablets.

Capabilities:

Vision

Model Information

Model

google/gemma-3n-e4b

Format

GGUF

Params

4.38B

Arch

Gemma3n

Domain

11x

Download Options

GGUF

Gemma 3n E4B Instruct

4.24 GB

Applicable model file already downloaded

물론 GPU 호환성 가능

README

Gemma 3n E4B

Gemma 3n is a multimodal generative AI model optimized for use in everyday devices, such as phones, laptops, and tablets.

This model includes innovations in parameter-efficient processing, including Per-Layer Embedding (PLE) parameter caching and a MatFormer model architecture that provides the flexibility to reduce compute and memory requirements.

E = effective parameters.

Supports a context length of 32k tokens.

GGUFs are currently text-only. We are working to expand capabilities and remove this limitation.

Cancel

Use in New Chat

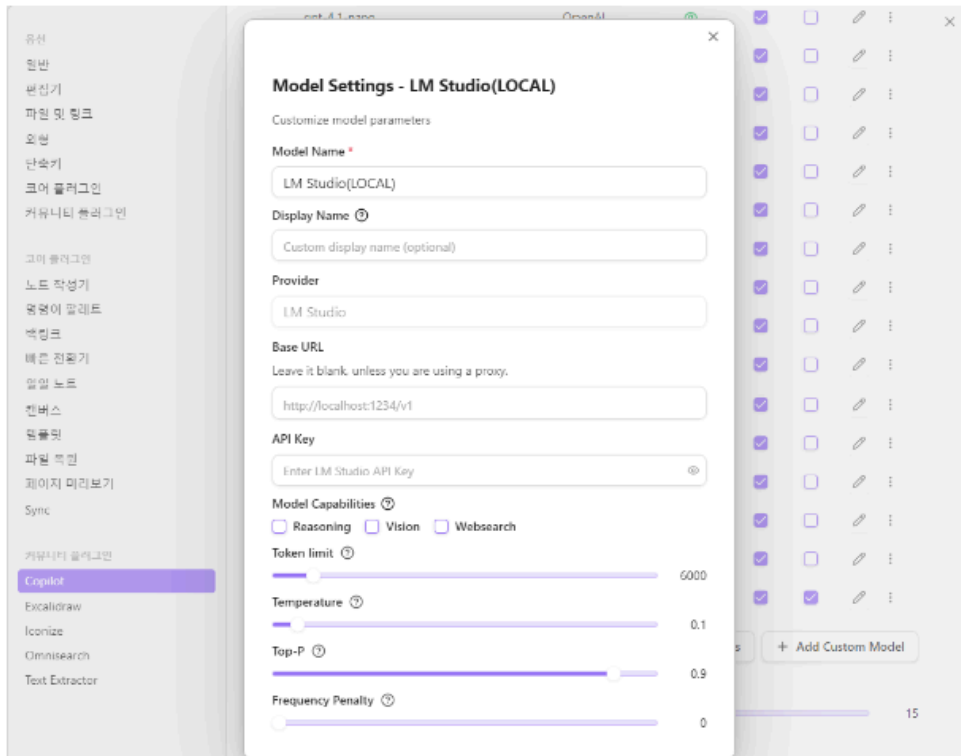
3. 시스템 프롬프트 튜닝 실험

프리셋 구성



무작위성 정도를 조절합니다. 0은 항상 같은 결과를 생성하며, 높은 값은 창의성과 다양성을 증가시킵니다

Obsidian - LM Studio(LOCAL) settings



Model Settings - LM Studio(LOCAL)

Customize model parameters

Model Name *
LM Studio(LOCAL)

Display Name ⓘ
Custom display name (optional)

Provider
LM Studio

Base URL
Leave it blank, unless you are using a proxy.
http://localhost:1234/v1

API Key
Enter LM Studio API Key ⓘ

Model Capabilities ⓘ
☐ Reasoning ☐ Vision ☐ Websearch

Token limit ⓘ
6000

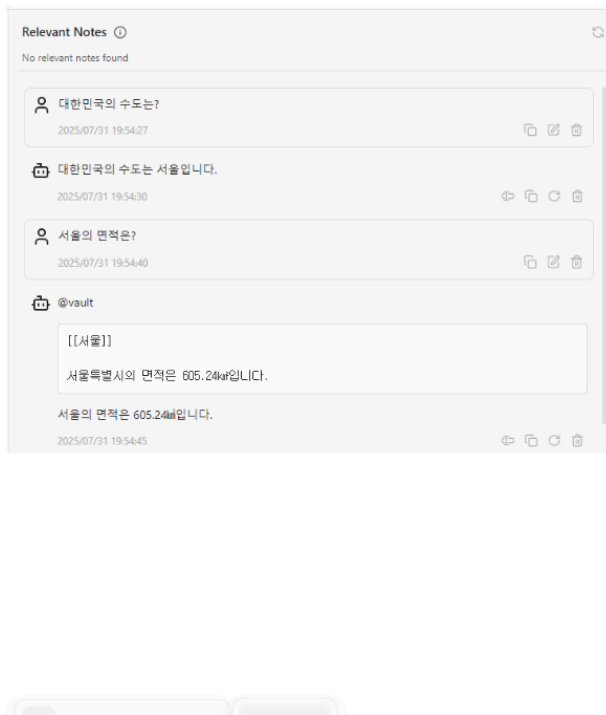
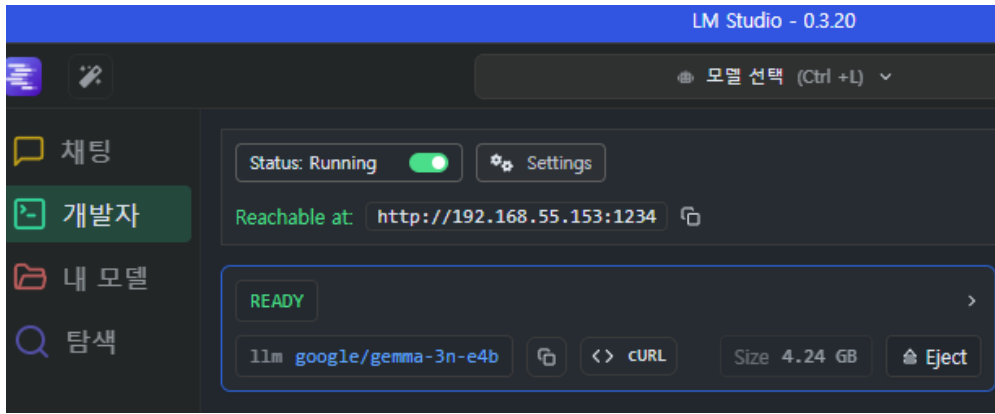
Temperature ⓘ
0.1

Top-P ⓘ
0.9

Frequency Penalty ⓘ
0

+ Add Custom Model

- System Prompt 설정 및 Obsidian 세팅 후 running.



4. Reachy Mini와의 연결 가능성

```
04-lm-studio.ipynb U X
+ Code + Markdown | ▶ Run All ⏮ Restart ⏻ Clear All Outputs | 📄 Variables 📄 Outline ...

1 from langchain_openai import ChatOpenAI
2 from langchain_core.callbacks.streaming_stdout import StreamingStdOutCallbackHandler
3 from langchain_core.prompts import PromptTemplate
4 from langchain_core.output_parsers import StrOutputParser
5
6
7 llm = ChatOpenAI(
8     base_url="http://localhost:1234/v1",
9     api_key="lm-studio",
10    model="teddylee777/EEVE-Korean-Instruct-10.8B-v1.0-gguf",
11    streaming=True,
12    callbacks=[StreamingStdOutCallbackHandler()], # 스트리밍 콜백 추가
13)
14
15 prompt = PromptTemplate.from_template(
16     """You are a helpful, smart, kind, and efficient AI assistant. You always fulfill the user's requests to the best of your ability.
17     You must generate an answer in Korean.
18
19     #Question:
20     {question}
21
22     #Answer: """
23 )
24
25 chain = prompt | llm | StrOutputParser()
```

✅ 가능성 (👍)

1. 모듈이 모두 오픈소스 + 로컬 가능

- Whisper, LM Studio, Reachy SDK 모두 무료 사용 가능

2. API 기반 연결이라 구조 자체는 단순

- STT 예시: `text = whisper.transcribe(audio)`
- LLM 예시: `requests.post(...)`
- 제어 예시: `send_command(JSON)`

✨ 결론

Reachy Mini를 포함한 로봇과 연결 가능성

이 구조는 완전한 로컬 오프라인 처리 가능: Whisper + LM Studio + Reachy SDK
STT → LLM → 제어까지 하나의 흐름으로 연결되는 로봇 자연어 인터페이스 기반입니다.

- 자연어 명령 → 행동 명령 변환 흐름도
- 통신 방법 (ROS2, socket 통신, REST API 등)

참고자료

[로컬에서 HuggingFace LLM 사용 강의](#)
[gguf 개념](#)