

## GPT OSS – 로컬 pc(LM STUDIO)

GGUF Model size: 20.9B params Architecture: gpt-oss Chat template

Hardware compatibility

Log In to view the estimation

2-bit: IQ2\_XXS | 11.5 GB, IQ2\_XS | 11.5 GB, IQ2\_S | 11.5 GB, Q2\_K | 11.6 GB, IQ2\_M | 11.5 GB, Q2\_K\_L | 11.8 GB

3-bit: IQ3\_XXS | 11.6 GB, IQ3\_XS | 11.6 GB, Q3\_K\_S | 11.6 GB, IQ3\_M | 11.6 GB, Q3\_K\_M | 11.6 GB, Q3\_K\_L | 11.5 GB, Q3\_K\_XL | 11.8 GB

4-bit: IQ4\_XS | 11.6 GB, Q4\_K\_S | 11.7 GB, IQ4\_NL | 11.6 GB, MXFP4\_MOE | 12.1 GB, Q4\_0 | 11.5 GB, Q4\_1 | 11.6 GB, Q4\_K\_M | 11.7 GB

5-bit: Q5\_K\_S | 11.7 GB, Q5\_K\_M | 11.7 GB

6-bit: Q6\_K | 12 GB, Q6\_K\_L | 12 GB

8-bit: Q8\_0 | 22.3 GB

16-bit: BF16 | 11.9 GB, BF16 | 11.9 GB, BF16 | 41.9 GB, **BF16 | 28.1 MB**

< [https://huggingface.co/bartowski/huihui-ai\\_Huihui-gpt-oss-20b-BF16-abliterated-GGUF](https://huggingface.co/bartowski/huihui-ai_Huihui-gpt-oss-20b-BF16-abliterated-GGUF) >

### Imatrix Quantization

## IQ1, IQ2 (I-quants): imatrix 기반의 고급 양자화

- **imatrix 없이:** 모델 전체에 비교적 균일한 압축을 적용합니다.
- **imatrix 사용 시:** 모델의 핵심적인 부분을 덜 압축하고, 덜 중요한 부분은 더 많이 압축해서 전반적인 성능을 보존합니다.

...-imatrix.gguf (크기: **28.1 MB**)

- 이것은 양자화할 때 사용한 '**재료(importance matrix)**'일 뿐, 완성된 모델이 아닙니다.

imatrix.gguf는 작성자가 만든 양자화 모델에서 바로 이 imatrix를 사용해서 이런 과정을 거쳐 만들어졌습니다"라고 자신의 작업 과정을 투명하게 공개하는 것과 같습니다.

요약하자면, imatrix.gguf 파일은 다음과 같은 사람들을 위해 존재합니다.

- 모델을 실행만 하려는 일반 사용자 (User) → 필요 없음
- 모델이 어떻게 만들어졌는지 검증하려는 연구자 (Researcher) → 필요함
- 이 모델을 기반으로 다른 버전을 만들려는 개발자 (Developer) → 필요함

LM Studio - 0.3.23

My Models

Models Directory: C:\Users\SH\lmstudio\models

Filter models... (Ctrl + F)

Filter by: LLMs Text Embedding

Arch	Params	Publisher	Model	Quant	Size	Date Modified	Actions
gpt-oss	20B	openai	openai/gpt-oss-20b	MXFP4	12.11 GB	today	⋮ ⋯

LM Studio - 0.3.23

ChatGPT Introduction

openai/gpt-oss-20b

Thought for 1.80 seconds

### ROS 2 (Humble) 초보자를 위한 "첫걸음" 학습 로드맵

목표:

1 ROS 2 기본 개념을 이해하고, 2 간단한 노드를 만들고 실행해 보며,  
3 실제 프로젝트에 바로 적용할 수 있는 실전 스클립을 갖추는 것.

아래 단계별 가이드를 따라가면 "초보 → 중급"으로 자연스럽게 넘어갈 수 있어요. 각 단계마다 핵심 리소스(문서·튜토리얼·코드)와 실습 예시를 함께 제공하니, 바로 실행해 보세요!

#### 1 기본 개념 정리 (약 2~3시간)

Send a message to the model...

Reasoning Effort

Input token count: 0 Context is 317.6% full

LM Studio 0.3.23 (Build 3) User Power User Developer RAM: 3.31 GB | CPU: 0.00 %

LM Studio - 0.3.23

ChatGPT Introduction

다른 로컬 PC LLM 모델과 gpt oss의 차이점과 장단점

#### 1 개념 정리

구분	정의
로컬 PC LLM	개인 혹은 조직 내부에서만 실행되는 대형 언어 모델(예: GPT-Neo, GPT-J, LLaMA). 하드웨어(특히 GPU)와 전력 비용이 필요하며, 사용자는 전체 파이프라인을 직접 구축한다.
GPT-OS S	"오픈소스 GPT"라는 개념으로, 모델 아키텍처 + 툴킷 + 사전 학습 가중치가 공개된 프로젝트를 의미한다. Hugging Face Transformers, EleutherAI, Meta LLaMA 등이 대표적이다.

두 용어는 겹칠 수 있다: 로컬 PC에서 실행되는 GPT-OS 모델도 "로컬 PC LLM"이라 부른다.

#### 2 차이점

Send a message to the model...

Reasoning Effort

Input token count: 0 Context is 434.0% full

Context Model Program

Preset Discard Unused

Enter a name for the preset... Save (x)

System Prompt

you are ros2 expert.  
Summarize it briefly. do not exceed token and only answer korean.

Token count: 21

LM Studio 0.3.23 (Build 3) User Power User Developer RAM: 2.41 GB | CPU: 0.00 %

