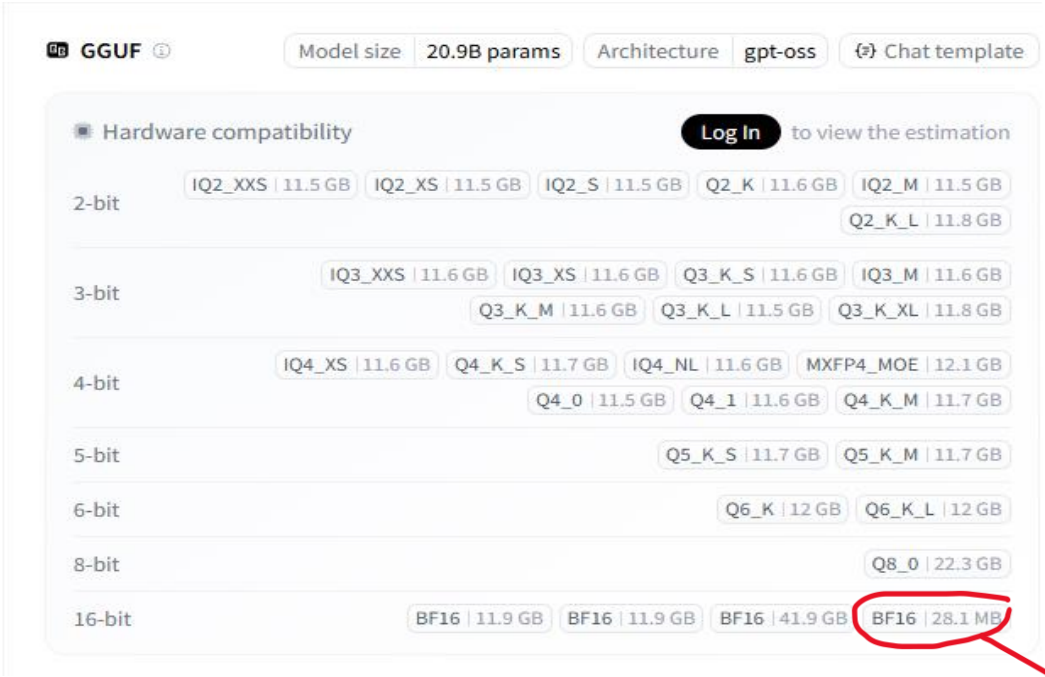


Imatrix Quantization

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.

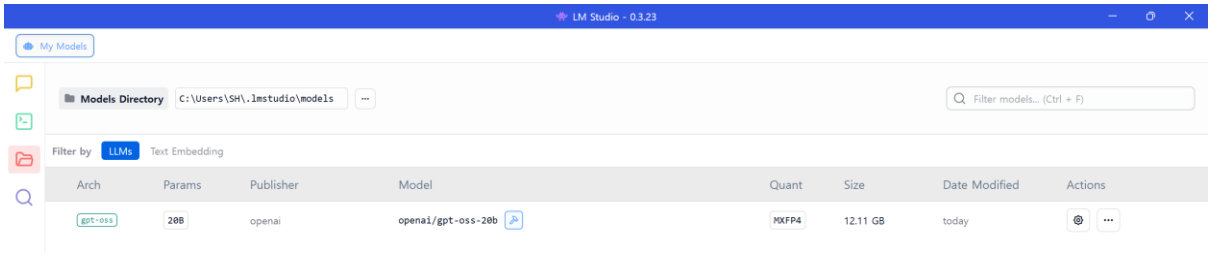
...-imatrix.gguf (크기: **28.1 MB**)

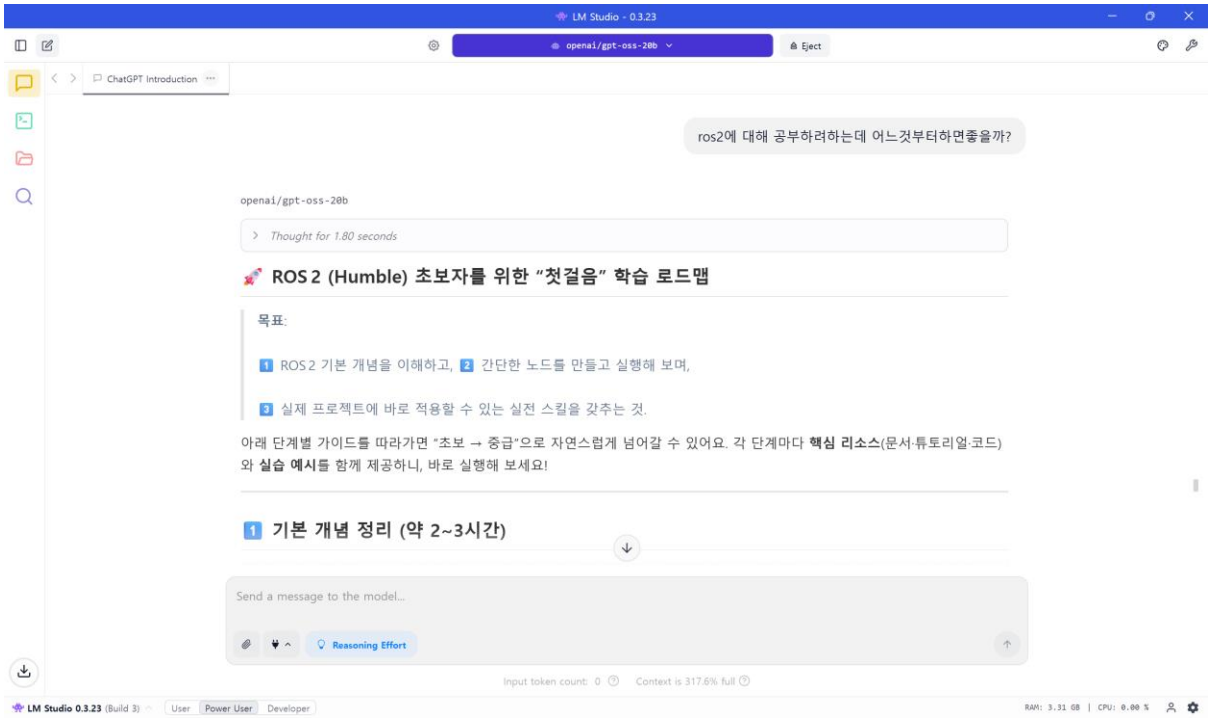
- 이것은 양자화할 때 사용한 **'재료(importance matrix)'**일 뿐, 완성된 모델이 아닙니다.

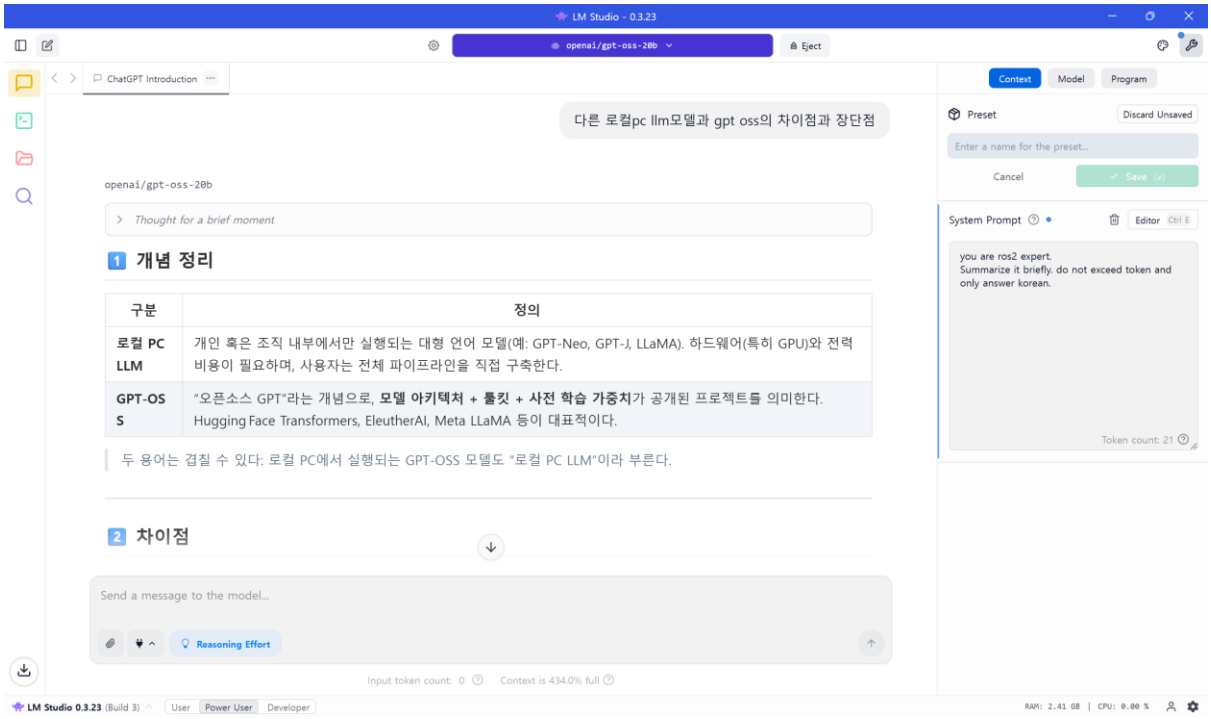
imatrix.gguf 는 작성자가 만든 양자화 모델에서 바로 이 imatrix 를 사용해서 이런 과정을 거쳐 만들어졌습니다"라고 **자신의 작업 과정을 투명하게 공개**하는 것과 같습니다.

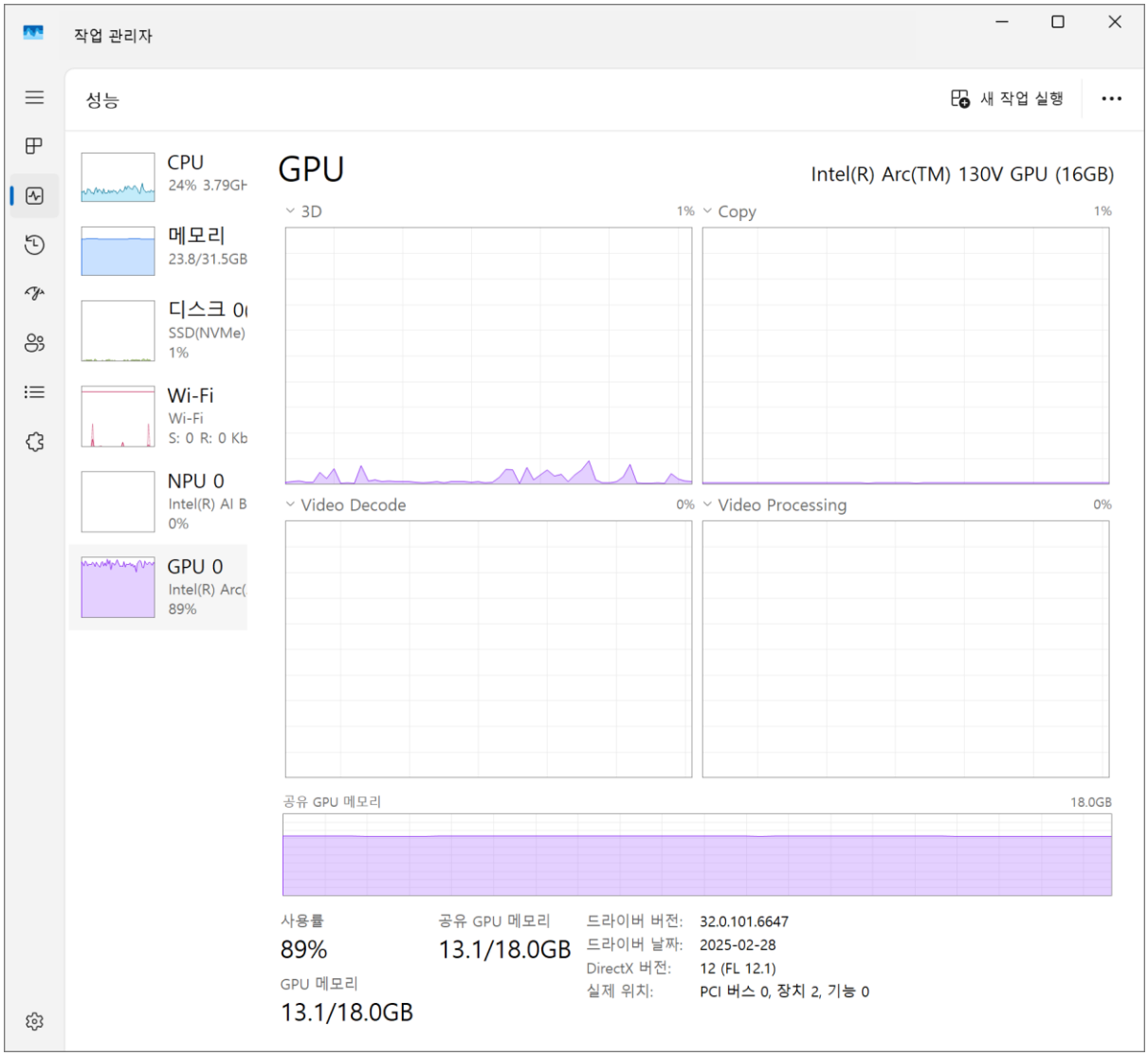
요약하자면, imatrix.gguf 파일은 다음과 같은 사람들을 위해 존재합니다.

- **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
openai/gpt-oss-20b
Eject
ChatGPT Introduction
ros2에 대해 공부하려하는데 어느것부터하면좋을까?
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 스킬을 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
openai/gpt-oss-20b
Eject
ChatGPT Introduction
다른 로컬pc llm모델과 gpt oss의 차이점과 장단점
openai/gpt-oss-20b
Thought for a brief moment
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-OSS
"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-OSS 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 Unsaved
Enter a name for the preset...
Cancel
Save
System Prompt
Editor
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 %


작업 관리자
성능
새 작업 실행
CPU
24% 3.79GH
메모리
23.8/31.5GB
디스크 0(
SSD(NVMe)
1%
Wi-Fi
Wi-Fi
S: 0 R: 0 Kb
NPU 0
Intel(R) AI B
0%
GPU 0
Intel(R) Arc(
89%
GPU
Intel(R) Arc(TM) 130V GPU (16GB)
3D
1%
Copy
1%
Video Decode
0%
Video Processing
0%
공유 GPU 메모리
18.0GB
사용률
89%
공유 GPU 메모리
13.1/18.0GB
GPU 메모리
13.1/18.0GB
드라이버 버전:
32.0.101.6647
드라이버 날짜:
2025-02-28
DirectX 버전:
12 (FL 12.1)
실제 위치:
PCI 버스 0, 장치 2, 기능 0