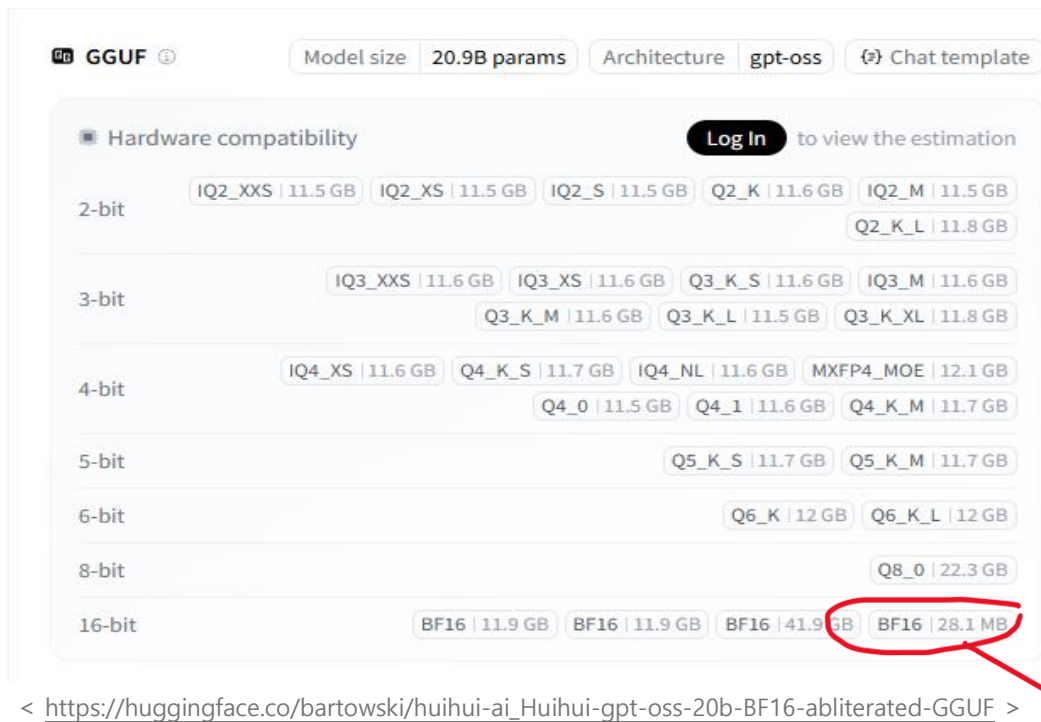


GPT OSS – 로컬 pc(LM STUDIO)



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

Hardware compatibility [Log In](#) to view the estimation

Quantization	Options
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 >

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.

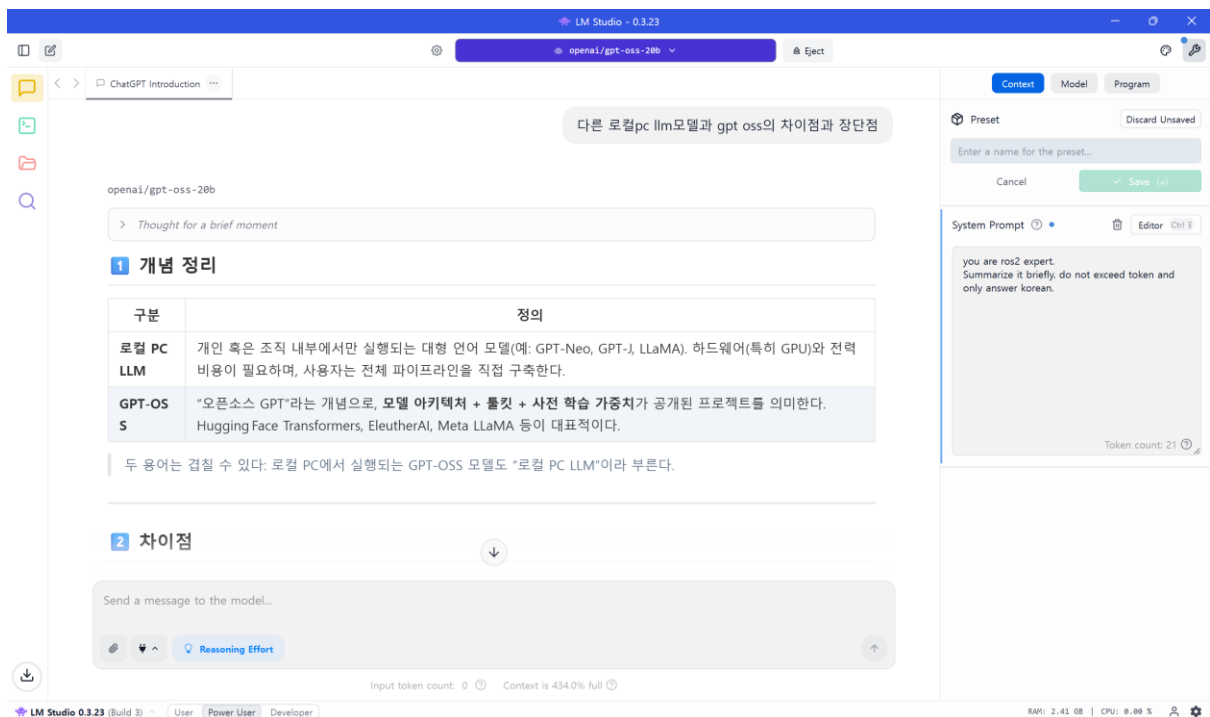
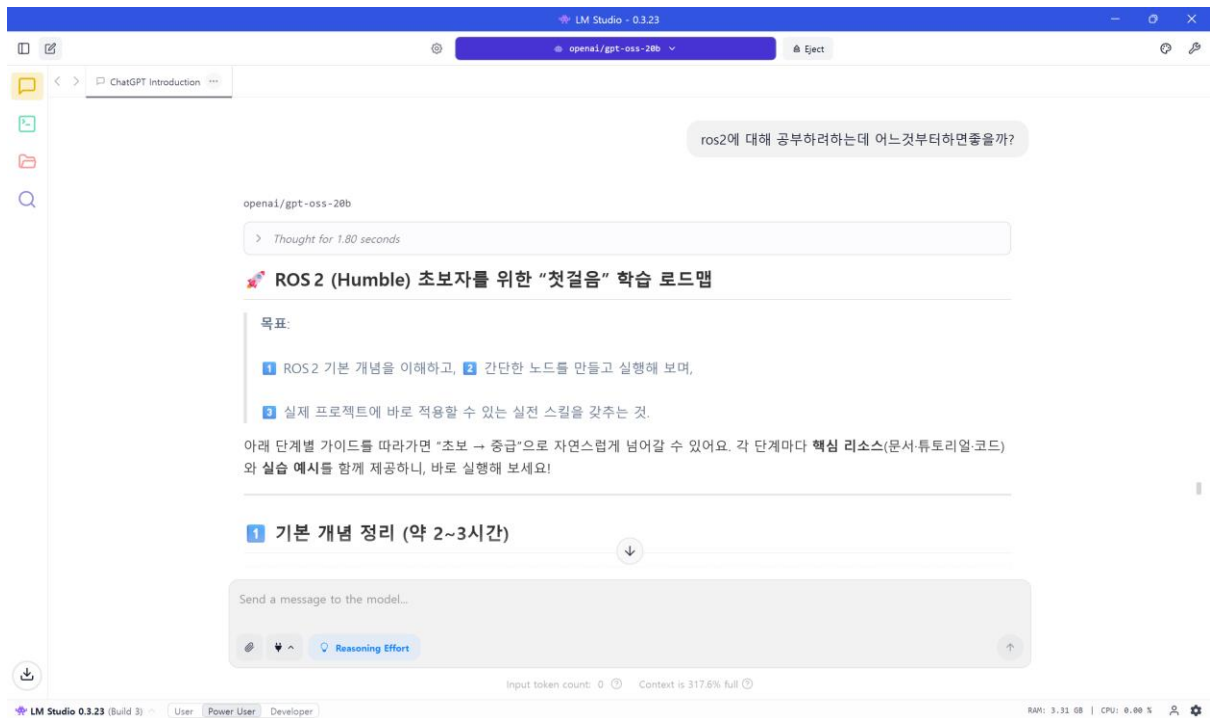
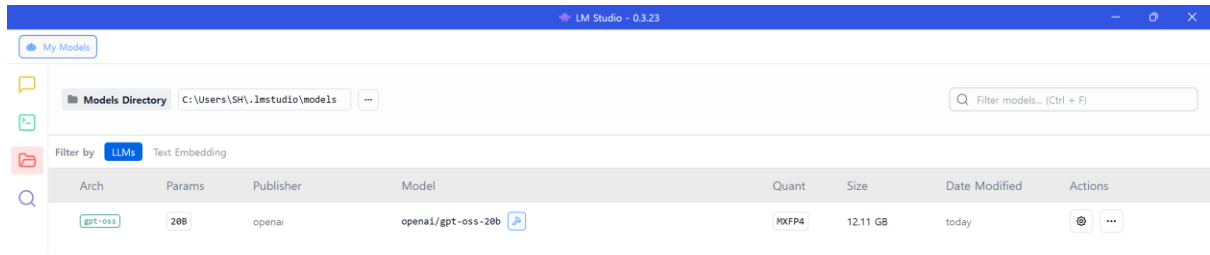
...-imatrix.gguf (크기: **28.1 MB**)

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

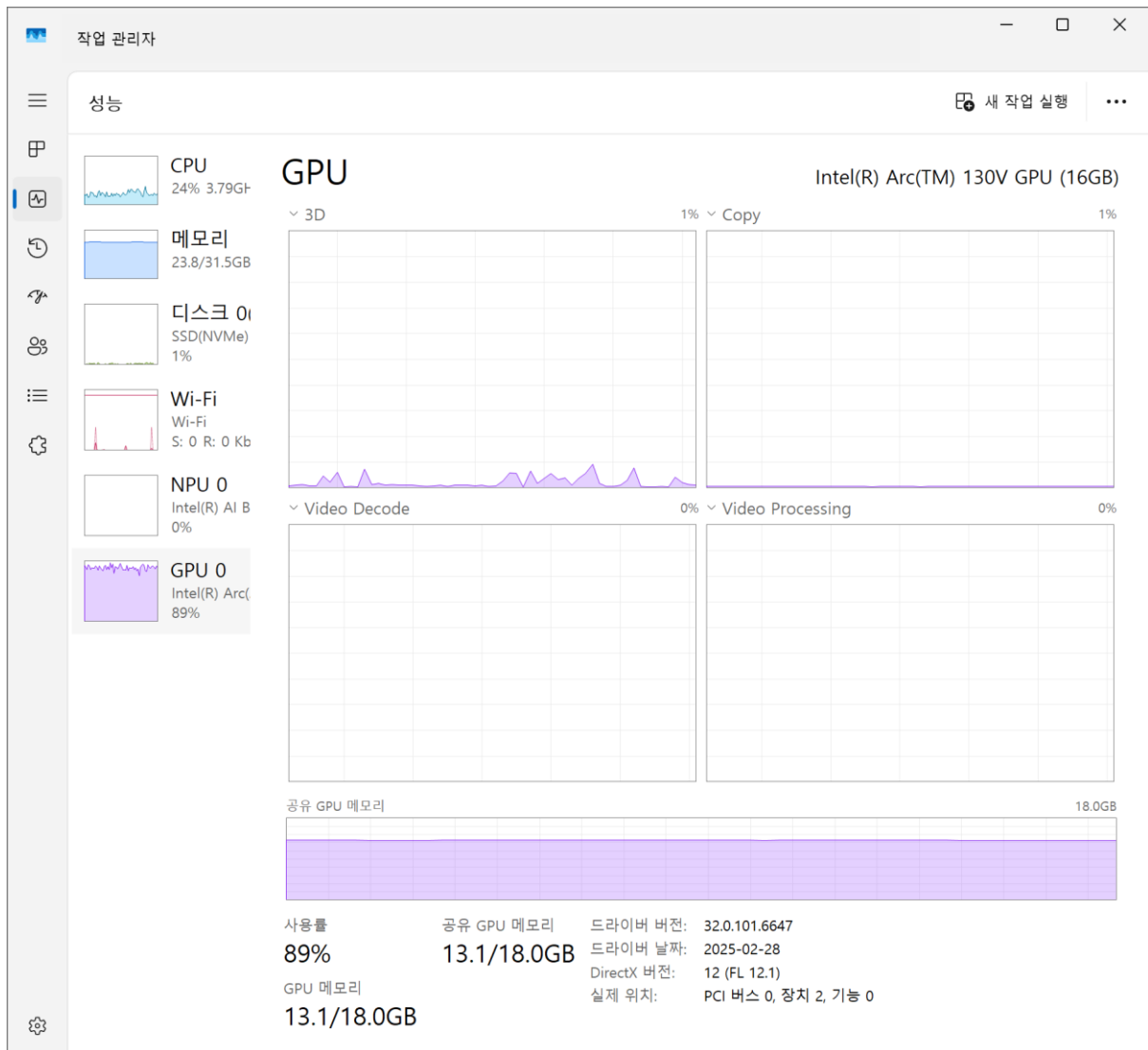
imatrix.gguf 는 작성자가 만든 양자화 모델에서 바로 이 imatrix 를 사용해서 이런 과정을 거쳐 만들어졌습니다"라고 **자신의 작업 과정을 투명하게 공개**하는 것과 같습니다.

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



- 시스템 부하



노트북 기종: 갤럭시북5 프로 NT960XHA-K52A 32GB(RAM)