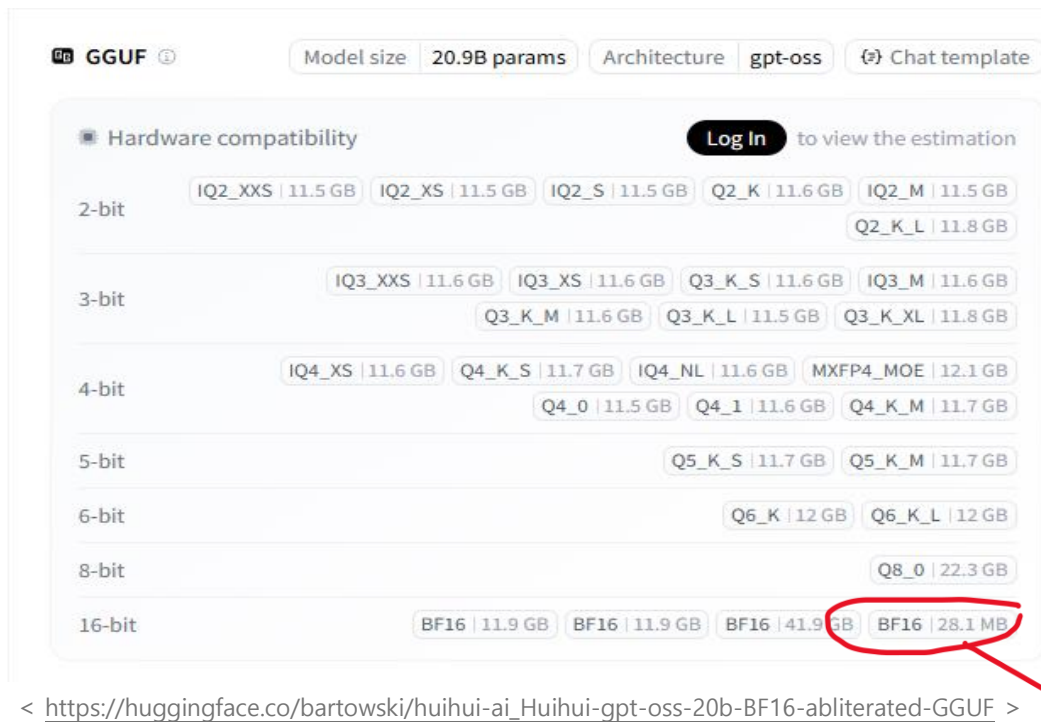


GPT OSS – 로컬 pc(LM STUDIO)



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

