

7.2 챗봇 기능 확장

#출처제공 #실시간검색 #외부API연동

Q. 검색 레퍼런스 확인 기능 제공하는지?

A. 하이브리드 검색 결과 기반으로 출처 포함 응답 가능합니다.

Q. 실시간 데이터에 대한 답변은 어떻게 처리하는지?

A. RAG 구성에서 데이터 수집 주기 및 임베딩 처리에 대한 구성이 중요하며 알맞는 아키텍처 구성이 된다면 학습과 다르게 실시간으로 수집되는 데이터가 임베딩되서 검색 및 답변 추론에 포함되서 구성 될 수 있습니다.

8. 하드웨어 인프라 관련

8.1 Local LLM GPU 서버 관련

Q. A100 GPU 서버와 T4 GPU 서버의 사양은 어떻게 차이 나는가?

A. NVIDIA A100과 T4는 성능/메모리 면에서 큰 차이가 있습니다.

구분	NVIDIA A100 (40GB/80GB HBM2e)	NVIDIA T4 (16GB GDDR6)
아키텍처	Ampere (7nm, 6912 CUDA 코어, 432 Tensor 코어)	Turing (12nm, 2560 CUDA 코어, 320 Tensor 코어)
메모리 유형/용량	HBM2e, 40GB 또는 80GB (메모리 대역폭 최대 1555GB/s)	GDDR6, 16GB (메모리 대역폭 약 320GB/s)
연산 성능(TFLOPS)	FP32 19.5 TFLOPS, TF32 156 TFLOPS, FP16 312 TFLOPS (Tensor Core)	FP32 8.1 TFLOPS, FP16 65 TFLOPS
가격(2025기준)	A100 40GB: \$12,000~13,000 A100 80GB: \$15,000~16,000	T4: \$1,200~1,400