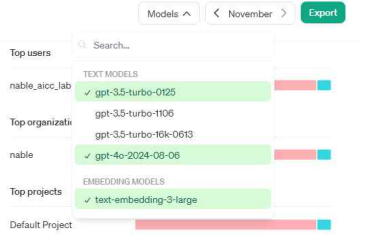
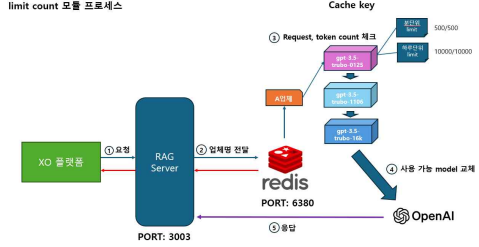
gpt-3.5-turbo-0125	200,000 TPM	500 RPM 10,000 RPD	2,000,000 TPD
gpt-3.5-turbo-1106	200,000 TPM	500 RPM 10,000 RPD	2,000,000 TPD
gpt-3.5-turbo-16k	200,000 TPM	500 RPM 10,000 RPD	2,000,000 TPD

gpt-3.5-turbo-0125 model만 진행할 때는 분단위 요청 limit 500, 일단위 요청 limit 10000까지인데
gpt-3.5-turbo-0125, gpt-3.5-turbo-1106, gpt-3.5-turbo-16k 각 성능 테스트를 진행했을때
정확도는 차이가 별로 나지 않는 것을 확인했습니다.



playground에서 각 요청했을때 gpt-3.5-turbo 시리즈가 구분되어서 출력되는 것을 확인했고
각 모델로 인식하기 때문에 교체 모듈 활용하면
분단위 요청 limit: 500 -> 1500, 일단위 요청 limit: 10000 -> 30000 확장해서 사용할 수 있습니다.
하루 24시간 기준으로 분할해서 진행할경우 평균 분단위 요청: 7->21회 가능합니다.

Tier limit에 따른 모델 교체 모듈 프로세스



gpt-3.5-turbo-0125, gpt-3.5-turbo-1106, gpt-3.5-turbo-16k, gpt-4o 요청수 기준
text-embedding-3-large는 token 사용 기준으로 구현했습니다.

embedding model token 사용 기준으로 별도 구현한 이유

text-embedding-3-large	1,000,000 TPM	3,000 RPM	3,000,000 TPD

embedding model은 RPD가 없고 TPD 기준으로 되어있기 때문에 token 사용 기준으로 모듈 교체될 수 있게 구현했습니다.