Embedding API 선정

* **개요**: 금융 특화 번역 LLM RAG의 embedding model 선정 필요
* **모델 후보**:  
  1. text-embedding-3-small (OpenAI)  
  2. text-embedding-3-large (OpenAI)
* **선정 방법**: 가격, 성능을 고려하여 선정

**모델 preview**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **model** | **Output dimension** | **Max input** | **pricing** | **Page per dollar**  **(assuming ~800 tokens per page)** | **Performance on MTEB eval** |
| text-embedding-3-large | 3,072 | 8191 | $0.130/1M tokens | 9,615 | 64,6% |
| text-embedding-3-small | 1,536 | 8191 | $0.020/1M tokens | 62,500 | 62.3% |
| text-embedding-ada-002 | 1,536 | 8191 | $0.100/1M tokens | 12,500 | 61.0% |

**모델 성능 비교 (유사도: cosine similarity)**

|  |
| --- |
| 두 모델 사이의 큰 성능차이는 없음 (별첨 result\_output.xlsx 참고) |

**모델 특장점 비교**

|  |  |  |
| --- | --- | --- |
| **비교 항목** | **text-embedding-3-small** | **text-embedding-3-large** |
| 학습 데이터 | 짧고 명확한 용어 중점 | 복잡하고 논리적인 문장 중점 |
| 분석 유형 | 단어 중점 | 의미 중점 |
| 효율성 | 추론 속도 빠름 | 추론 속도 느림 |

**결론**

|  |
| --- |
| text-embedding-3-small 사용 |