

### Q. 사용하는 벡터 DB 최대 차원수는 어디까지 지원하는지?

Elasticsearch의 버전에 따라 지원하는 최대 벡터 차원 수는 다음과 같습니다:

- Elasticsearch 8.9.2 이하: 최대 1024차원까지 지원
- Elasticsearch 8.10.0 이상: 최대 2048차원까지 지원
- Elasticsearch 8.11.0 이상: 최대 4096차원까지 지원

또한, 8.11.0부터는 vector dimension 파라미터를 명시하지 않아도, 최초 인덱싱된 벡터의 차원을 기준으로 자동 설정됩니다. 따라서 해당 버전 이후에는 보다 유연하게 벡터를 다룰 수 있습니다.

※ 현재 사용하는 Elasticsearch 버전에 따라 제한이 다르므로, 실제 프로젝트에서는 해당 버전 확인이 필요합니다.

### Q. Vector Model은 주로 어떤 걸 사용하는지?

A. bge-m3, E5 모델 기반의 다국어 처리 임베딩 모델 사용. 삼성 및 LG 등의 여러 사업을 진행하면서, 다양한 Vector 모델에 대한 테스트를 진행 하였습니다. 현재 주로 많이 사용하는 모델은 한국어 인식이 좋다고 판단되는 E5 모델이나 M3 모델을 활용 하고 있습니다. 이 부분은 삼성전자 고객 요청으로 여러 모델을 가지고 벤치마킹 테스트도 진행해서 성능을 평가했었습니다.

### Q. 벡터 임베딩 모델 설정 기준은?

A. 벡터 임베딩 모델을 설정할 때는 다음 기준을 고려:

- 업무·도메인 특성: 처리하려는 데이터 유형(예: 뉴스, 대화, 과학문서 등)과 작업 유형(검색, 분류, 추천 등)에 적합한 모델을 선택합니다.
- 언어 및 데이터 규모: 주로 다룰 언어와 사용 가능한 학습/추론 데이터 크기에 따라 대형 모델 또는 소형 경량 모델을 결정합니다.