사용 가능한 텍스트 임베딩 모델

Vertex AI에서 지원하는 주요 텍스트 임베딩 모델과 특성은 다음 표와 같습니다. 각 모델은 생성하는 벡터의 길이가 다르며, 주로 768차원 또는 3072차원을 사용합니다.

모델 이름	설명	출력 임베딩 차원
textembedding-gecko@001	영어 텍스트 전용 Gecko 임베딩 모델 (기본 버전)cloud.google.com	768cloud.google.com
textembedding-gecko@002	영어 텍스트 전용 Gecko 임베딩 모델 (버전 2)cloud.google.com	768cloud.google.com
textembedding-gecko@003	영어 텍스트 전용 Gecko 임베딩 모델 (버전 3)cloud.google.com	768cloud.google.com
text-embedding-004	영어 텍스트용 임베딩 모델 (구버전)cloud.google.com	768cloud.google.com
text-embedding-005	영어 및 코드 특화 임베딩 모델 cloud.google.com	768cloud.google.com
textembedding-gecko-multilingual@001	다국어 텍스트용 Gecko 임베딩 모델 cloud.google.com	768cloud.google.com
text-multilingual-embedding-002	다국어 텍스트 특화 임베딩 모델 cloud.google.com	768cloud.google.com
gemini-embedding-001	최첨단 대규모 임베딩 모델 (영어·다국어·코드 전반 지원)cloud.google.com	3072cloud.google.com

- **Gecko 모델**(textembedding-gecko@001~003, textembedding-gecko-multilingual@001)은 Google에서 제공하는 범용 임베딩 시리즈로, 영어 전용(Gecko)과 다국어(Gecko Multilingual) 버전이 있습니다. 이 모델들은 768차원 벡터를 생성하며, 대규모 텍스트 데이터로 학습되어 일반 검색·분류 등에 사용합니다cloud.google.com.
- **Text-Embedding 004/005**는 주로 영어 문서와 코드 작업에 특화된 모델로, text-embedding-005가 최신(기본) 모델입니다cloud.google.com.
- **Text-Multilingual 002**는 다양한 언어(예: 한국어 포함)의 텍스트에 대해 학습된 다국어 임베딩 모델입니다cloud.google.com.

- **Gemini Embedding 001**은 가장 최신의 대규모 임베딩 모델로, 이전의 005/002 모델을 통합해 영어·다국어·코드 작업 모두에서 최상의 성능을 제공합니다 cloud.google.com. 기본 출력 벡터는 3072차원입니다cloud.google.com.

각 모델은 사용 목적과 지원 언어가 다르므로, 한국어 등 특정 언어가 필요한 경우 지원 언어 목록을 확인하세요cloud.google.comcloud.google.com. 또한, 출력 임베딩 차원(output_dimensionality 매개변수)을 조절하여 벡터 크기를 선택할 수 있습니다(기본값 768 또는 3072)cloud.google.com.

임베딩 모델 사용 예시

Vertex AI 텍스트 임베딩은 REST API 또는 Google GenAI SDK(또는 Vertex AI SDK)를 통해 사용할 수 있습니다. 예를 들어 **Python SDK (google-genai)**를 사용하여 텍스트 임베딩을 요청하면 다음과 같습니다cloud.google.com:

python

CopyEdit

```
from google import genai
from google.genai.types import EmbedContentConfig

client = genai.Client()
response = client.models.embed_content(
    model="gemini-embedding-001",
    contents="텍스트 임베딩 예시 문장입니다.",
    config=EmbedContentConfig(
        task_type="RETRIEVAL_DOCUMENT",  # 문서 검색 용도로 분류
        output_dimensionality=3072,      # 출력 벡터 차원 (3072)
        title="예시 제목",               # Optional
    ),
)
print(response.embeddings.values)  # 임베딩 벡터 출력
```

또는 **REST API**를 직접 호출할 수도 있습니다cloud.google.com. 예를 들어, gemini-embedding-001 모델을 사용하는 경우 :predict 엔드포인트에 POST 요청을 보냅니다:

bash

CopyEdit

```
curl -X POST ₩
  -H "Authorization: Bearer $(gcloud auth print-access-token)" ₩
  -H "Content-Type: application/json; charset=utf-8" ₩
  -d '{
        "instances": [
          { "content": "임베딩 생성 대상 문장" }
        ],
        "parameters": { "autoTruncate": true }
      }' ₩
  "https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID/locations/us-central1/publishers/google/models/gemini-embedding-001:predict"
```

위 예시에서 "instances" 필드에 텍스트를 넣고, 필요에 따라 "parameters"에서 autoTruncate(길이 초과 시 잘림 여부) 등을 설정할 수 있습니다cloud.google.com. Python SDK 사용 시에도 유사한 파라미터(EmbedContentConfig)를 지정할 수 있습니다 cloud.google.com.

주요 제한사항 및 권장사항

- **입력 텍스트 수 제한**: 비-Gemini 모델의 경우 한 요청당 최대 250개 텍스트(각각 임베딩 하나 생성)가 허용되며, Gemini Embedding 모델은 한 번에 하나의 텍스트만 입력할 수 있습니다cloud.google.com.
- **토큰 길이 제한**: 요청 전체의 최대 토큰 수는 20,000 토큰이며, 이 한도를 초과하면 오류가 발생합니다cloud.google.com. 개별 텍스트는 모델당 최대 2048토큰까지 처리되며(이를 초과하면 잘림), 실험 모델(text-embedding-large-exp-03-07 등)는 최대 8192토큰까지 허용합니다cloud.google.comcloud.google.com. 필요 시 autoTruncate=false로 설정해 초과 시 오류를 발생시킬 수 있습니다.

- **리전별 제한**: us-central1 리전에서는 요청당 최대 250개 텍스트(비-Gemini 모델 기준)를 지원하나, 다른 리전에서는 최대 5개로 제한됩니다. 실험 모델은 us-central1에서만 지원되며, 요청당 입력은 1개로 제한됩니다cloud.google.com.
- **모델 지정 규칙**: 모델 이름을 지정할 때는 반드시 @버전 또는 @latest 접미사를 포함해야 합니다. 예를 들어 "text-embedding-005@latest" 또는 "gemini-embedding-001"과 같이 정확히 지정해야 하며, 접미사 없이 모델 이름만 쓰면 유효하지 않습니다cloud.google.com.
- **출력 차원 조정**: 기본적으로 모든 모델은 풀사이즈 벡터(768 또는 3072차원)를 반환합니다. 그러나 output_dimensionality 매개변수를 사용하여 벡터 크기를 줄일 수 있습니다. 낮은 차원을 선택하면 저장 공간 및 계산 비용을 절감하면서 품질 저하를 최소화할 수 있습니다cloud.google.com.
- **문서 제목 및 태스크 유형**: RETRIEVAL_DOCUMENT 같은 task_type 파라미터와 title 필드를 사용하면 검색-문서 형태 등의 문맥 정보를 모델에 전달할 수 있습니다. (기본 task_type은 RETRIEVAL_QUERY입니다cloud.google.com.)

기타 유용한 기능 및 관련 리소스

- **Vertex AI 벡터 검색(Vector Search)**: 생성된 임베딩을 벡터 데이터베이스에 저장하여 빠른 유사도 검색을 수행할 수 있습니다. Vertex AI에서는 벡터 검색 기능을 지원하며, 생성된 임베딩을 벡터로 저장해 검색 엔진으로 활용할 수 있습니다 cloud.google.com.
- **배치 임베딩(Batch Predict)**: 대량의 텍스트에 대해 임베딩을 얻어야 할 경우, 배치 예측 기능을 사용할 수 있습니다. Vertex AI의 배치 예측 API로 한 번에 많은 텍스트에 대한 임베딩을 생성하고 결과를 저장할 수 있습니다cloud.google.com.
- **멀티모달 임베딩**: 텍스트뿐 아니라 이미지·오디오 등의 임베딩을 얻는 멀티모달 임베딩 기능도 제공합니다. 자세한 내용은 Vertex AI 멀티모달 임베딩 문서를 참고하세요cloud.google.com.
- **임베딩 튜닝**: 특정 도메인에 맞게 임베딩을 미세조정할 수 있는 튜닝 기능이 제공됩니다. 필요 시 튜닝 과정을 통해 도메인 특화 임베딩을 만들 수 있습니다 cloud.google.com.
- **제한사항 문서 및 예제**: API 레이트 제한 등 자세한 쿼터 정보는 Vertex AI 한도 문서에서 확인할 수 있습니다cloud.google.com. 또한 Google이 공개한 임베딩 연구 자료(예: Gecko 모델 관련 논문)도 참고하면 모델 특성 이해에 도움이 됩니다 cloud.google.com.

위 리소스를 활용하면 Vertex AI 텍스트 임베딩 기능을 효과적으로 사용하고, 생성된 임베딩을 검색·분류 등의 애플리케이션에 활용할 수 있습니다 cloud.google.comcloud.google.com.

출처: Google Cloud Vertex AI 공식 문서 및 안내 cloud.google.comcloud.google.comcloud.google.comcloud.google.comcloud.google.comcloud.google.com (2025년 기준).