📄 Research and Selection of Embedding Models for Multilingual RAG-based Chatbot

Selected Models Overview Table

Model Name	Developer	Training Dataset	Architecture	Multilingual Support	Domains / Use Cases	Open Access	URL
E5 small	Intel Labs & Hugging Face	C4, Wikipedia, multilingual QA pairs (task-specific)	Encoder-only transformer (E5 series, optimized BERT)	~100 (including English, Urdu, Arabic, Hindi, etc.)	Multilingual RAG, semantic search, QA	Yes	Link
BAAI/bge-m3	Beijing Academy of AI	Multilingual web corpus, domain-specific corpora	Transformer-based (BERT variant)	100+ languages	Retrieval, multilingual RAG, enterprise search	Yes	Link
distiluse-base-multilingual-cased-v2	Sentence-Transformers (UKP Lab)	Multilingual sentence pairs and parallel corpora	DistilBERT (Siamese network)	50+ languages	Semantic similarity, chatbots, sentence matching	Yes	Link

◇ 1. Multilingual-e5-small

- **Model Name and Developer**: Intel Labs & Hugging Face
- **Description**:

The E5 model family is specifically optimized for **embedding-based retrieval**, such as search, QA, and retrieval-augmented generation (RAG). The **multilingual E5-small** version supports over 100 languages, making it a powerful and efficient alternative to LaBSE. It generates high-quality semantic embeddings that align similar meanings across languages.

- **Training Dataset(s)**:

Trained on large-scale multilingual data including:

  - **CCMatrix**: Parallel web-scraped bilingual data
  - **Wikipedia**
  - **Multilingual NLI (XNLI)**
  - **Web-based QA and retrieval data**
- **Model Architecture**:

Single-tower Transformer encoder (BERT-like) optimized for retrieval using **contrastive learning** and **instruction tuning** (with prompts like "query: ..." and "passage: ...").

- **Multilingual Capabilities**:

Supports **100+ languages**, including **English**, **Urdu**, **Arabic**, **Chinese**, **French**, **German**, etc.

- **Best Use Cases**:
  - Cross-lingual dense retrieval
  - Semantic search (multi-language)
  - Retrieval-Augmented Generation (RAG)
  - Open-domain QA and document matching
  - Multi-language chatbot grounding
- **Justification for Chatbot**:

`multilingual-e5-small` is highly suitable for building **retrieval-based multilingual chatbots**. It produces consistent embeddings for both queries and passages across languages, making it effective for aligning user questions with semantically relevant answers, even when they're in different languages. It's lightweight, fast, and delivers **better performance than LaBSE** on several multilingual retrieval benchmarks.

◇ 2. BAAI/bge-m3

- **Model Name and Developer**: bge-m3 by Beijing Academy of Artificial Intelligence (BAAI)
- **Description**:

BGE-M3 is a multilingual embedding model optimized for RAG tasks, trained to handle multi-language queries and retrieval more effectively. It performs well in both semantic search and dense retrieval tasks across diverse languages.

- **Training Dataset(s)**:

Multilingual web datasets, QA corpora, and domain-specific text covering multiple contexts (tech, medical, general knowledge).

- **Model Architecture**: Transformer-based encoder, fine-tuned on multilingual retrieval and cross-lingual sentence matching tasks.
- **Multilingual Capabilities**:

Supports **100+ languages**, with efficient handling of both high-resource and low-resource languages.

- **Best Use Cases**:
  - RAG-based systems
  - Multilingual search & QA
  - Document retrieval
  - Chatbots & assistant systems
- **Justification for Chatbot**:

This model is specifically designed for **retrieval-enhanced generation** scenarios, making it a strong backbone for a multilingual RAG chatbot. It ensures precise, language-aware embeddings that improve both recall and relevance.

◇ 3. distiluse-base-multilingual-cased-v2

- **Model Name and Developer**: distiluse-base-multilingual-cased-v2 by Sentence-Transformers / UKP Lab
- **Description**:

This model is a multilingual version of DistilBERT trained using the sentence-transformers library. It is optimized for sentence-level semantic understanding in multiple languages, with low inference latency.

- **Training Dataset(s)**:

Trained on multiple multilingual datasets including Tatoeba, OPUS, and other translation pairs to learn cross-lingual sentence semantics.

- **Model Architecture**:

DistilBERT encoder using a **Siamese network** structure to encode sentence pairs.

- **Multilingual Capabilities**:

Supports over **50 languages**, including European, Asian, and Middle Eastern languages.

- **Best Use Cases**:
  - Chatbots
  - Semantic similarity
  - FAQ retrieval
  - Text matching
- **Justification for Chatbot**:

This model is fast and effective for understanding user queries and matching them with multilingual knowledge base entries. It is lightweight, making it ideal for real-time applications with resource constraints.