### Ollama Model Deployment Documentation

**Overview:** This document details the deployment and usage of Ollama models: bge-large:latest for generating embeddings and deepseek-r1:latest for generating responses.

#### Models Deployed:

**bge-large:latest**

* 1. **Description:** A BERT-based model used for generating embeddings.
  2. **Model Format:** GGUF
  3. **Parameter Size:** 334.09M
  4. **Quantization Level:** F16
  5. **Size:** 670,532,029 bytes
  6. **Digest:** b3d71c92805938e2c9d78e6f35e82d7cc1bd77e05b47dd2ab3016240f4021c15
  7. **Modified At:** 2025-02-14T13:53:52.934820797+05:30

**deepseek-r1:latest**

* 1. **Description:** A model designed for generating responses.
  2. **Model Format:** GGUF
  3. **Parameter Size:** 7.6B
  4. **Quantization Level:** Q4\_K\_M
  5. **Size:** 4,683,075,271 bytes
  6. **Digest:** 0a8c266910232fd3291e71e5ba1e058cc5af9d411192cf88b6d30e92b6e73163
  7. **Modified At:** 2025-02-14T13:33:56.652914656+05:30

#### Usage:

**Embedding Generation:**  
The bge-large:latest model is used for generating embeddings from textual data. These embeddings can be used to store and search vectors in a vector database.

**Response Generation:**  
The deepseek-r1:latest model is utilized for generating natural language responses based on provided input prompts.