**Dynamic routing test strategy**

**ollama version is 0.3.9**

23 September, 2024



Figure 1. Test strategy for dynamic routing

Perform the test as shown in Figure 1. By changing the embedding models such as

1. **all-minilm:33m,**
2. **nomic-embed-text,**
3. **snowflake-arctic-embed:110m and**
4. **mxbai-embed-large.**

Change the **THRESHOLD value** for each of the embed model.

|  |  |
| --- | --- |
| Embed Model | Threshold Value |
| **all-minilm:33m** | **0.4950** |
| **nomic-embed-text** | **0.63** |
| **snowflake-arctic-embed:110m** | **0.66** |
| **mxbai-embed-large** | **0.62** |

We keep fix the LLM in this whole study as “**qwen2:0.5b-instruct**”.

**For zero-shot:** create\_multi\_shot\_prompt() function has only below text:

|  |
| --- |
| **examples = """**  Now, respond to this query:  User: """ + prompt + """  Assistant:  **"""** |

**For one-shot:** create\_multi\_shot\_prompt() function has only below text that consists of Example 1 from each dynamic router.

|  |
| --- |
| **examples = """**  Example 1:  ….  Example 1:  ….  Example 1:  …  Now, respond to this query:  User: """ + prompt + """  Assistant:  **"""** |

**For few-shot:** create\_multi\_shot\_prompt() function has only below text that consists of Example 1, Example 2 and Example 3 from each dynamic router.

|  |
| --- |
| **examples = """**  Example 1:  ….  Example 2:  ….  Example 3:  …  Example 1:  ….  Example 2:  ….  Example 3:  …  Example 1:  ….  Example 2:  ….  Example 3:  …  Example 1:  ….  Example 2:  ….  Example 3:  …  Now, respond to this query:  User: """ + prompt + """  Assistant:  **"""** |