Deepseek企业级Agent项目开发实战

Part 10. Prompt Cache 工程化实现方法与使用技巧

`Prompt Cache` (提示缓存) 是 `API` 生成响应的临时存储机制。当客户端多次发送相同的提示（`Prompt`）时到大模型服务时（比如 `DeepSeek` 的 `API` 服务），可以重复使用缓存中的内容，从而用来减少响应延迟和计算开销。

通常场景下，一个模型接收到最终有效的 `Prompt` 一般是 `system` 消息 + `user` 消息。即如下形式：

```
messages: [
    {"role": "system", "content": "你是一位乐于助人的助手"},
    {"role": "user", "content": "你好，我是居居，很高兴认识你"}
]
```

经常存在的情况是：`system` 消息是固定的，而 `user` 消息是变化的。除此以外，当涉及到多轮对话的时候，为了能让大模型具备历史的上下文信息，往往需要将之前所有的对话历史都拼接起来，作为新的 `Prompt` 进行传递。即如下形式：

```
# 第一轮对话
messages: [
    {"role": "system", "content": "你是一位乐于助人的助手"},
    {"role": "user", "content": "你好，我是居居，很高兴认识你"}
]
# 第二轮对话
messages: [
    {"role": "system", "content": "你是一位乐于助人的助手"},
    {"role": "user", "content": "你好，我是居居，很高兴认识你"}
    {"role": "assistant", "content": "你好，居居。"}
    {"role": "user", "content": "请问我叫什么名字？"}
]
```

`Prompt Cache` 做的事情就是存储并利用这些 `Prompt` 中的重复内容，来降低响应延迟和推理成本。

1. DeepSeek 硬盘上下文缓存

从应用开发的角度看，大模型服务响应一个比较关键的概念就是 `首 Token 延迟`，即从客户端发送请求到模型返回第一个 `Token` 的时间间隔。比如我们正在开发的 `AssistGen` 应用全部采用流式输出接口，`首 Token 延迟` 的直观感受就是用户看到第一个问题答案的时间。所以为了提升用户体验，需要尽可能的去降低 `首 Token 延迟` 的响应时间。像模型的服务商如 `OpenAI`、`Claude`、`Gemini` 等都推出了 `Prompt Cache` 接口特性。同样，`DeepSeek` 的在线 `API` 服务也支持使用 `Prompt Cache`（提示缓存），并且无需修改代码，无需更换接口，硬盘缓存服务会自动运行，并自动按照实际命中情况计费。

DeepSeek API 创新采用硬盘缓存，价格再降一个数量级

在大模型 API 的使用场景中，用户的输入有相当比例是重复的。举例说，用户的 prompt 往往有一些重复引用的部分；再举例说，多轮对话中，每一轮都要将前几轮的内容重复输入。

为此，DeepSeek 启用上下文硬盘缓存技术，把预计未来会重复使用的内容，缓存在分布式的硬盘阵列中。如果输入存在重复，则重复的部分只需要从缓存读取，无需计算。该技术不仅降低服务的延迟，还大幅削减最终的使用成本。

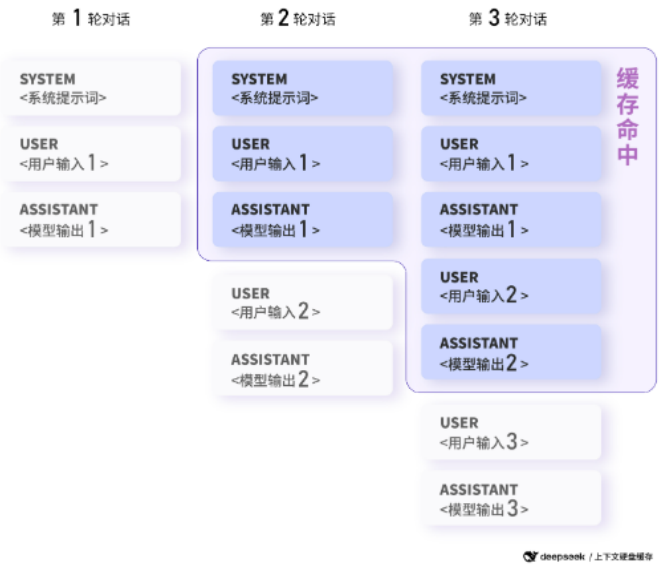
缓存命中的部分，DeepSeek 收费 0.1元 每百万 tokens。至此，大模型的价格再降低一个数量级[1]。

输入（每百万 tokens）		输出（每百万 tokens）
缓存命中	缓存不命中	2 元
0.1 元	1 元	

deepseek / 上下文硬盘缓存

`Prompt Cache` 能很好的解决 `首 Token 延迟` 的问题。 大模型在接收到输入的时候，需要先处理完整个输入才能开始生成输出，输入的序列越长，自然开始生成输出的时间就越长，所以通过 `Prompt cache`，大模型可以直接利用已缓存的处理结果，显著减少这个首次等待时间。

如下图所示，在多轮对话中，为了能让大模型具备历史的上下文信息，往往需要将之前所有的对话历史都拼接起来，作为新的 `Prompt` 进行传递。因此，对于第2轮和第3轮的对话，`Prompt Cache` 的作用是能够重复使用前面1轮的响应，而不是从头开始生成新的响应，从而加快响应时间，并最大程度地减少计算工作量。



这里我们可以通过调用 `DeepSeek` 的 `API` 接口来验证 `Prompt Cache` 的效果。这里需要注意的是：`DeepSeek` 的 `/chat/completions` 是一个“无状态” `API`，即服务端不记录请求的上下文，所以在每次请求时，需将之前所有对话历史拼接好后进行传递。其中，在返回的响应结构中，`usage` 字段中会包含 `prompt_cache_hit_tokens` 和 `prompt_cache_miss_tokens` 字段，分别表示缓存命中和未命中的 `Token` 数量。代码如下：

```
from openai import OpenAI
from dotenv import load_dotenv
import os
import time

# 加载 .env 文件
load_dotenv()
```

```python
# 实例化 DeepSeek API 客户端
client = OpenAI(
    api_key=os.getenv('DEEPSEEK_API_KEY'),
    base_url=os.getenv('DEEPSEEK_BASE_URL')
)


def chat_with_cache(messages):
    """单轮对话，返回响应并打印缓存情况"""
    start_time = time.time()

    response = client.chat.completions.create(
        model="deepseek-chat",
        messages=messages
    )

    elapsed_time = time.time() - start_time

    # 打印缓存命中情况
    cache_hit = response.usage.prompt_cache_hit_tokens
    cache_miss = response.usage.prompt_cache_miss_tokens
    print(f"\n[耗时 {elapsed_time:.2f}s]")
    print(f"缓存命中: {cache_hit} tokens")
    print(f"缓存未命中: {cache_miss} tokens")

    return response.choices[0].message

def main():
    # 初始化对话历史
    messages = [
        {"role": "system", "content": "你是一位乐于助人的助手"}
    ]

    print("开始对话 (输入 'q' 退出):")

    try:
        while True:
            # 获取用户输入
            user_input = input("\n用户: ")
            if user_input.lower() == 'q':
                break

            # 添加用户消息
            messages.append({"role": "user", "content": user_input})

            # 获取模型回复
            assistant_message = chat_with_cache(messages)

            # 添加模型消息到历史列表中
            messages.append({
                "role": assistant_message.role,
                "content": assistant_message.content
            })

            # 打印模型回复
            print(f"AI助手: {assistant_message.content}")

    except KeyboardInterrupt:
```

```
        print("\n对话已终止")
    except Exception as e:
        print(f"\n发生错误: {e}")

if __name__ == "__main__":
    main()
```

```
开始对话 (输入 'q' 退出):
```

大家也可以进入`llm_backend\app\test\deepseek_disk_cache.py`文件，运行代码，验证`Prompt Cache`的效果。如下所示：

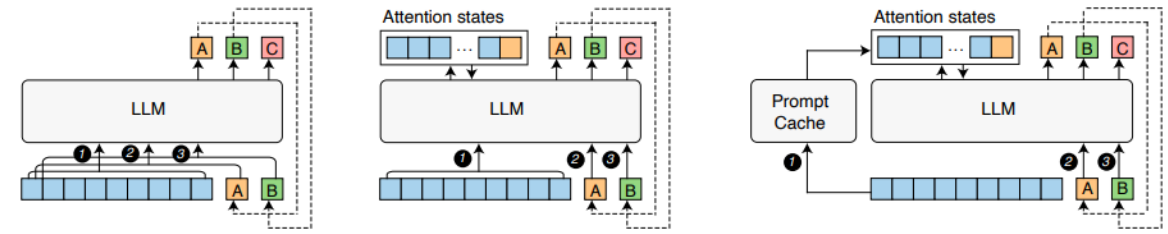
通过上图可以看到，第一轮对话中，缓存命中数为0，表示没有命中缓存，需要从头开始生成新的响应。第二轮对话中，缓存命中数为128。`DeepSeek`的缓存命中规则是：只有当两个请求的前缀内容相同时（从第 0 个 token 开始相同），才算重复。中间开始的重复不能被缓存命中。， 因此第二轮对话中，缓存命中的内容就是第一轮对话的输入 + 第一轮对话的输出。

同时，`DeepSeek`的缓存时间有效时间是几个小时到几天（官方文档说明），超出这个时间不再使用后缓存会被自动清空。

以上就是`Deepseek`官方`API`提供的硬盘缓存，大家需要重点了解。在实际使用的时候，对于一些系统指令、背景信息或者常用的工具定义等内容，放在提示词的开头位置，而将动态的内容放在结尾，这种应用方式就能极大程度的利用到缓存的优势。

2. 使用Redis实现服务请求缓存

正如上一小结内容所提，`Prompt Cache`和 大模型推理优化技术`KV-Cache`本质上都是利用“复用之前的计算结果”的方式来提高响应速度，其更偏向于模型架构、服务架构层。如果大家想更进一步的了解其实现原理，可以详细阅读`OpenAI`发布的论文：[Prompt Cache：Modular Attention Reuse for Low-Latency Inference](#)，这里不做扩展讲解。



(a) Autoregressive token generation (b) Generation with KV Cache (c) Generation with Prompt Cache

提示缓存除了在模型层进行优化外，根据平台或者特定用例也可以用不同的方式实现。在这里，我们重点介绍一个后端服务常用的缓存解决方案：**使用数据库缓存大模型对话的数据，当命中缓存后，则不向`API`服务发起调用，直接通过缓存提取某个问题的答案。**

简单来说，当用户将问题（Prompt）发送到后端服务时，首先检查缓存中是否有直接可用的响应。如果检索到，则直接返回缓存中的响应。否则，再向`API`发送新的请求，再结合`API`服务的缓存机制得到最终的响应。

缓存的概念一直是计算中一种基本优化技术。在其最基本的形式中，缓存涉及将经常访问的数据存储在与原始数据源相比更快的位置中。比如：

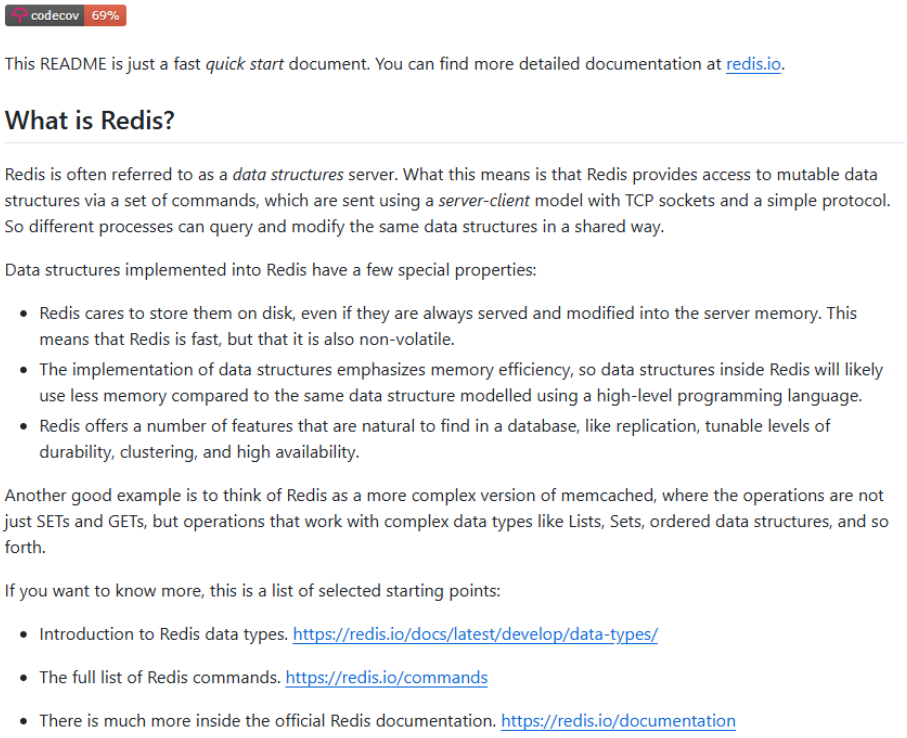
- 硬件缓存：`CPU`使用多个级别的缓存（L1，L2，L3）存储经常访问的说明和数据，从而减少了从较慢的主内存中获取信息所需的时间。
- Web缓存：`Web`浏览器和服务器使用缓存来存储静态内容，减少经常访问的网页的负载时间。

- 数据库缓存：数据库系统采用缓存机制来存储查询结果或经常访问的数据，从而改善重复查询的响应时间。

在这种场景需求下，主流且最常用的还是内存数据库`Redis`和`Memcached`，其中`Memcached`相较于`Redis`更加轻量，而`Redis`则能提供更加丰富的功能，同时具备比较高性能的读写操作，适合需要快速访问的场景。因此本节，我们就借助`Redis`来实现后端服务的`Prompt Cache`。

2.1 安装 Redis

`Redis`是一个高性能且开源的内存数据库，它提供了丰富的数据结构和功能，适合用于缓存和存储。大家可以在Redis官网了解`Redis`的更多信息，同时也能在Redis Github获取`Redis`的源码。


codecov 69%

This README is just a fast *quick start* document. You can find more detailed documentation at redis.io.

What is Redis?

Redis is often referred to as a *data structures* server. What this means is that Redis provides access to mutable data structures via a set of commands, which are sent using a *server-client* model with TCP sockets and a simple protocol. So different processes can query and modify the same data structures in a shared way.

Data structures implemented into Redis have a few special properties:

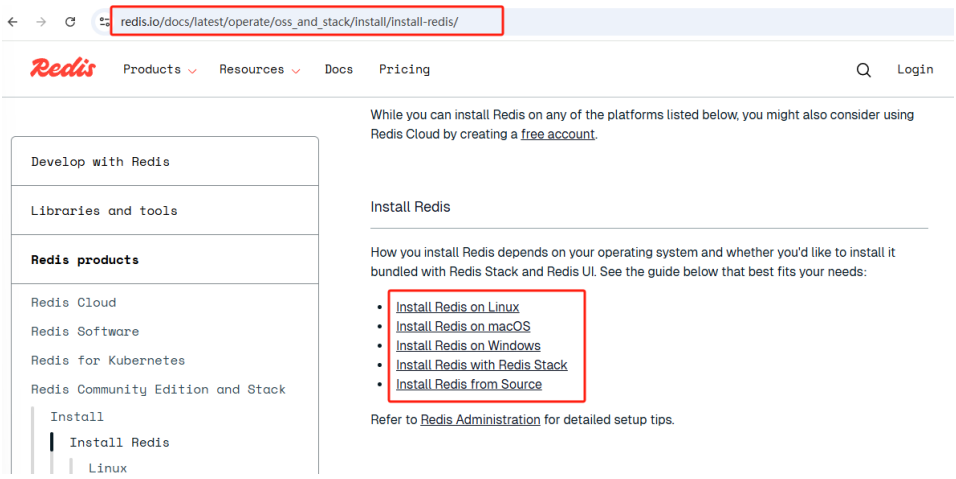
- Redis cares to store them on disk, even if they are always served and modified into the server memory. This means that Redis is fast, but that it is also non-volatile.
- The implementation of data structures emphasizes memory efficiency, so data structures inside Redis will likely use less memory compared to the same data structure modelled using a high-level programming language.
- Redis offers a number of features that are natural to find in a database, like replication, tunable levels of durability, clustering, and high availability.

Another good example is to think of Redis as a more complex version of memcached, where the operations are not just SETs and GETs, but operations that work with complex data types like Lists, Sets, ordered data structures, and so forth.

If you want to know more, this is a list of selected starting points:

- Introduction to Redis data types. https://redis.io/docs/latest/develop/data-types/
- The full list of Redis commands. https://redis.io/commands
- There is much more inside the official Redis documentation. https://redis.io/documentation

`Redis`支持多种操作系统，包括`Linux`、`macOS`、`Windows`等，但是Redis官网提供的安装教程有非常多的坑，并且描述过于简洁，导致大部分人都无法按照官方的操作文档顺利安装。



因此本节，我们就分别以`Windows`和`Linux`为例，来通过源码安装`Redis`，尽可能的降低安装难度同时也更好的适配大家的开发环境。

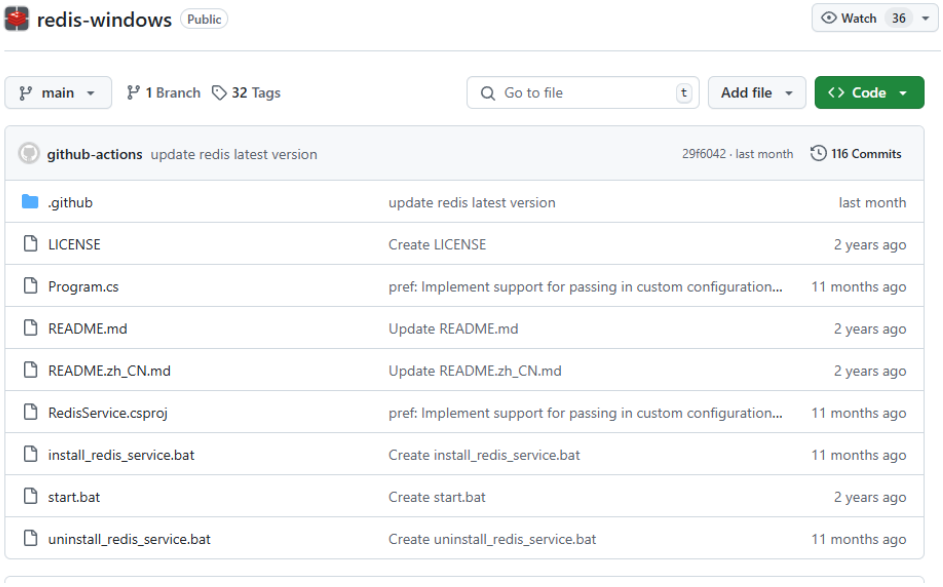
首先，我们来看`Windows`系统的安装。

2.1.1 Window源码安装Redis

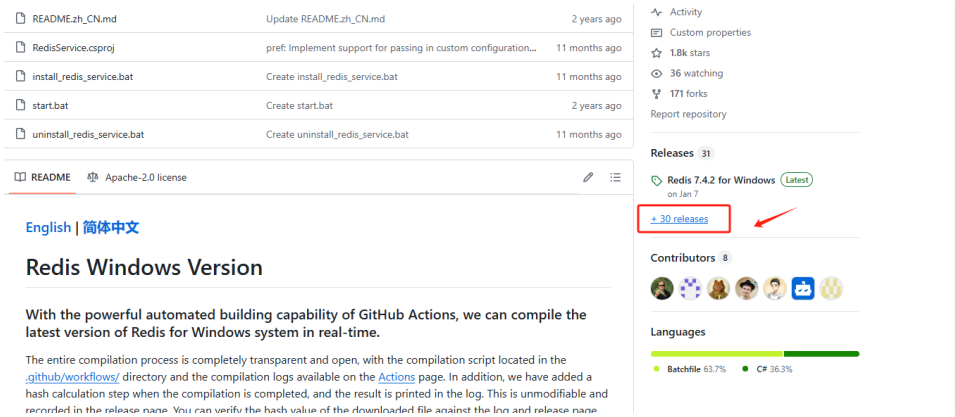
`Redis` 官方没有提供 `Windows` 版本的二进制包， 在`Windows`系统上运行官方推荐的方法是安装`WSL`后，使用`Linux`的子系统来运行`Redis`。如果大家的电脑已经安装了`WSL`，那么可以直接根据官方的安装教程进行安装。而如果不想因为一个`Redis`而安装`WSL`，毕竟一个完整的操作系统还是占用挺多系统资源的，则可以直接使用源码编译的方式来安装`Redis`。具体的操作方法如下：

- Step 1. 下载Redis的Windows的编译版本

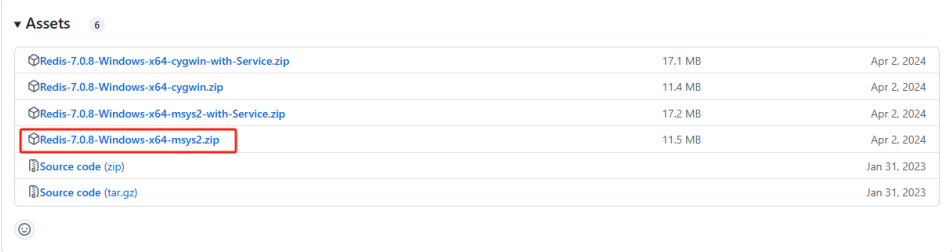
`Github`上有很多`Redis`的`Windows`编译版本的项目，这里推荐的一个项目地址为：地址为redis-windows，其最高编译的`Redis`版本为`7.4.2`，和`Redis`官方最新版本保持一致。



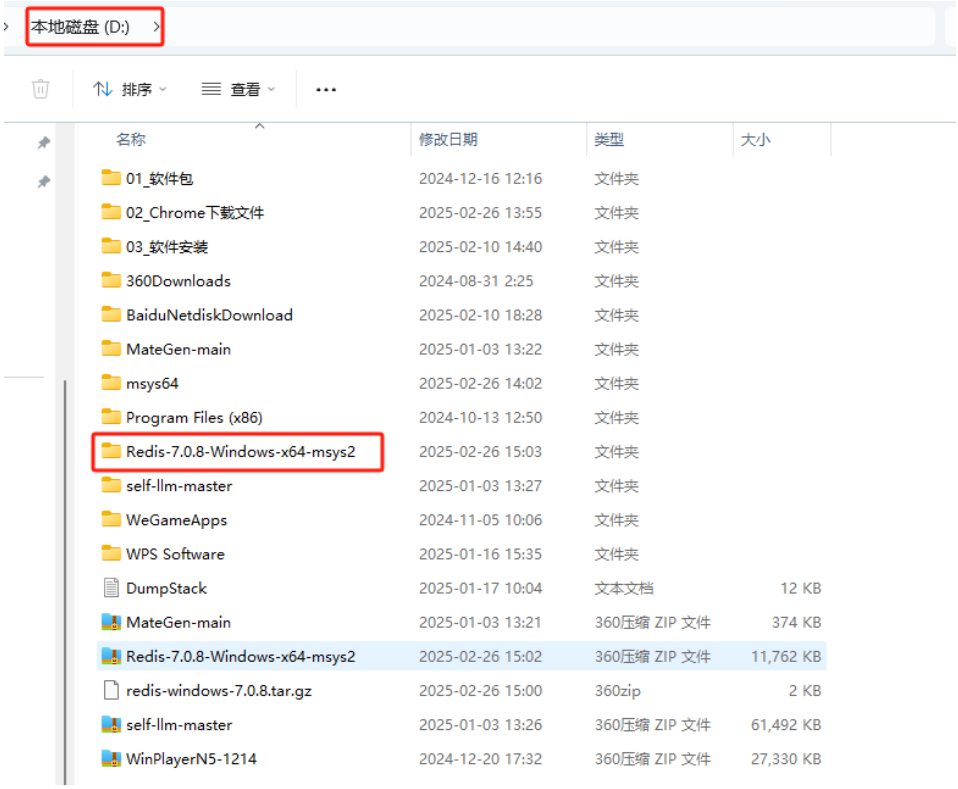
点击`Releases`按钮，即可看到`Redis`的`Windows`所有编译版本。



不建议大家选择最新的版本，因为一般最新的版本可能存在一些问题，这里我们选择`7.0.8`版本，可以点击如下链接直接进行下载：https://github.com/redis-windows/redis-windows/releases/tag/7.0.8



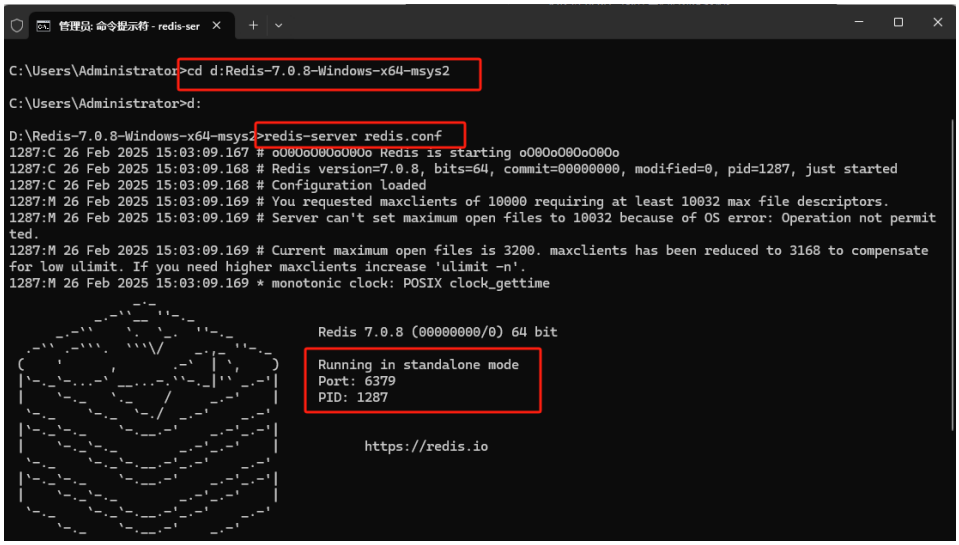
下载完成后，解压文件，得到`redis-7.0.8-win64-x64-msys2`文件夹。我这里存储到了`D:\redis-7.0.8-win64-x64-msys2`目录下。



- Step 2. 启动Redis服务

进入`Windows`系统的命令行终端（`cmd`），在命令行终端先进入`D:\redis-7.0.8-win64-x64-msys2`目录下，在执行如下命令启动`redis`服务：

```
cd d:\redis-7.0.8-win64-x64-msys2    # 这里替换为自己的解压路径
redis-server redis.conf
```



执行命令后，可以在命令行终端看到`Redis`服务启动的日志信息。

- Step 3. 连接Redis服务测试

启动`Redis`服务后，后端服务需要通过`Python`的`redis`库来连接`Redis`服务进行对其相关的操作。其调用示例在官方有较为详细的文档：https://redis.readthedocs.io/en/stable/examples.html，我们这里先进行`Redis`的连接测试。

这里需要关注的关键连接信息是：`Redis`的默认连接地址是`127.0.0.1`，端口号是`6379`，数据库是`0`。在`Redis`中，数据库 0 是默认的数据库。`Redis`默认提供了 16 个逻辑数据库，编号从 0 到 15。每个数据库都是独立的，具有自己的键空间。

```
# pip install redis  # 安装redis库

import redis

r = redis.Redis(host='localhost', port=6379, db=0)
```

建立连接后，我们就可以对`Redis`进行操作了。 相关的增删改查示例都非常简单，代码如下：

```
# 增：添加数据
def add_data(key, value):
    r.set(key, value)            # set：设置键值对
    print(f"添加数据: {key} -> {value}")

# 查：获取数据
def get_data(key):
    value = r.get(key)          # get：获取键值对
    if value is not None:
        print(f"获取数据: {key} -> {value.decode('utf-8')}")
    else:
        print(f"键 {key} 不存在")

# 改：更新数据
def update_data(key, new_value):
    if r.exists(key):            # exists：判断键是否存在
        r.set(key, new_value)    # set：设置键值对，如果键存在，则更新键值对
        print(f"更新数据: {key} -> {new_value}")
    else:
        print(f"键 {key} 不存在，无法更新")

# 删：删除数据
def delete_data(key):
    if r.delete(key):     # delete：删除键值对
        print(f"删除数据: {key}")
    else:
        print(f"键 {key} 不存在，无法删除")
```

可以调用`add_data`、`get_data`、`update_data`、`delete_data`函数来对`Redis`进行操作。

```
# 添加数据
add_data("name", "juju")
add_data("age", "30")
```

```
添加数据: name -> juju
添加数据: age -> 30
```

```
# 查数据
get_data("name")
get_data("age")
```

```
获取数据: name -> juju
获取数据: age -> 30
```

```
# 改数据
update_data("age", "31")

# 查更新后的数据
get_data("age")
```

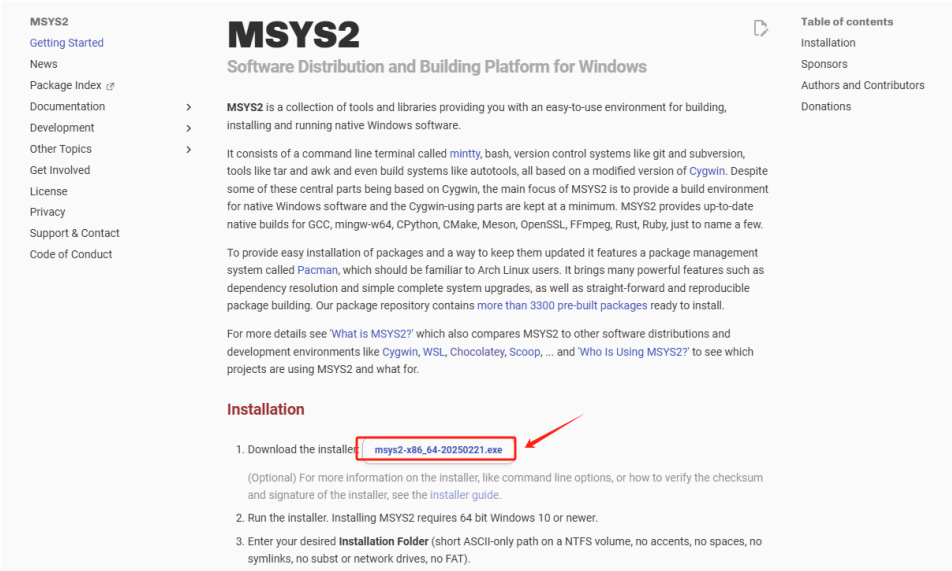
```
更新数据: age -> 31
获取数据: age -> 31
```

```
# 删数据
delete_data("name")

# 查已删除的数据
get_data("name")
```

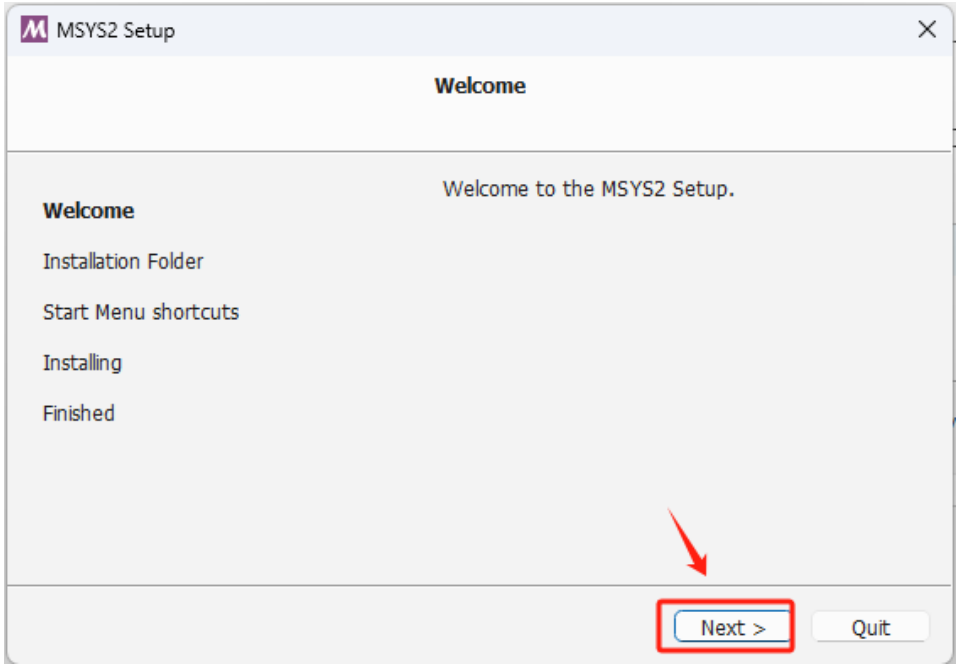
```
删除数据: name
键 name 不存在
```

注意：如果大家在启动的过程中，遇到类似 `MSYS2` 的错误，则需要大家先安装 `MSYS2`，然后再安装 `Redis`，具体的安装方法如下所示：（如果可以直接正常启动，则可以跳过该步骤）

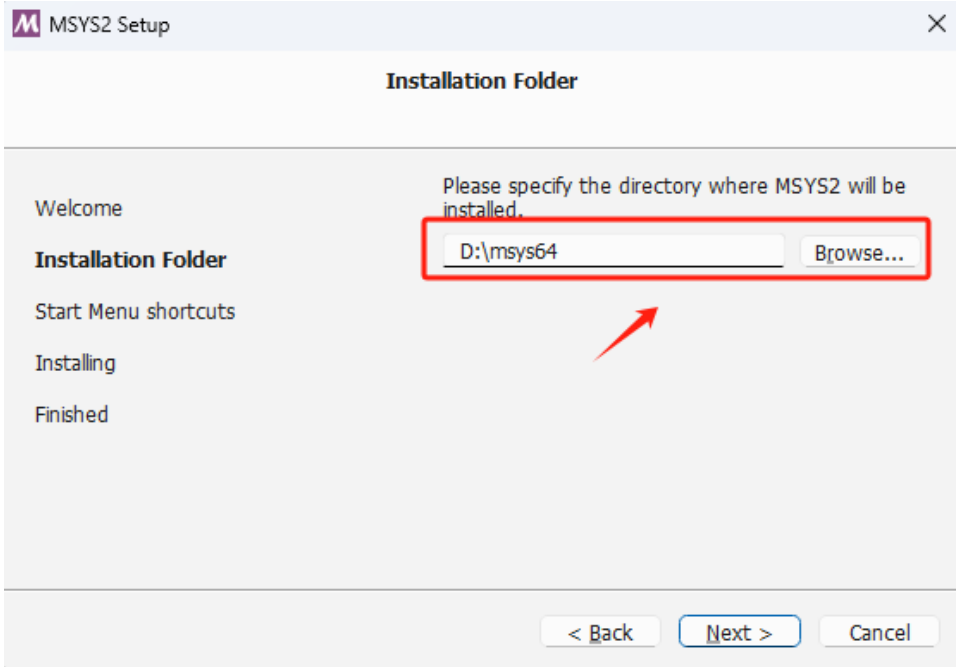
MSYS2 是一个在 Windows 上模拟 Linux 环境的工具，它提供了一个类似于 Linux 的命令行界面，并且可以安装和运行许多 Linux 软件。其官网为[MSYS2](#)，我们直接点击 `Download the installer` 官网的下载链接，下载安装包。注意：操作系统的版本要求是 64位Windows 10或更新。



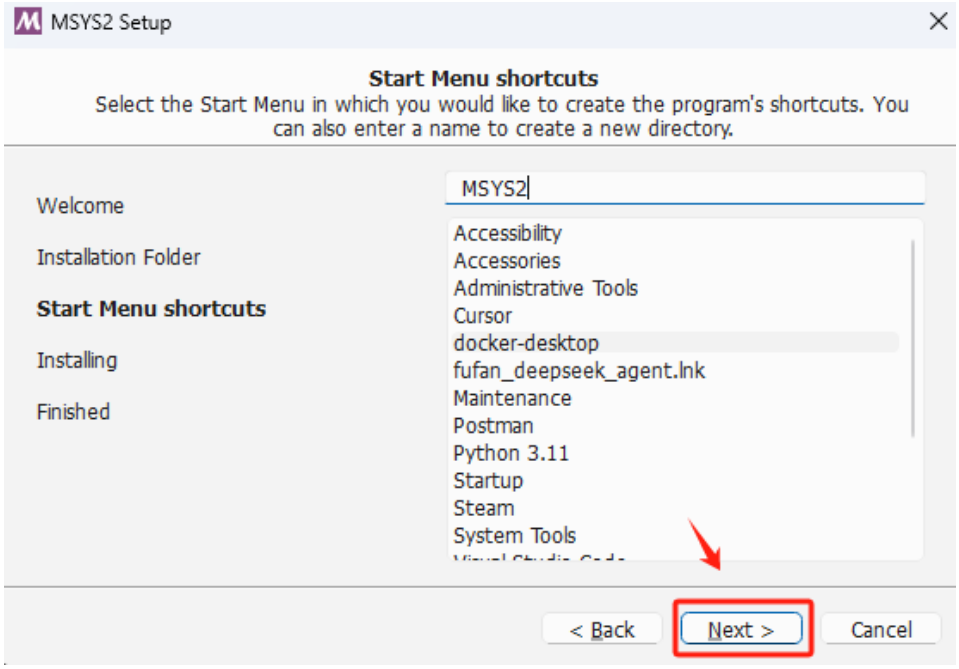
下载完成后，我们双击安装包，按照提示进行安装。



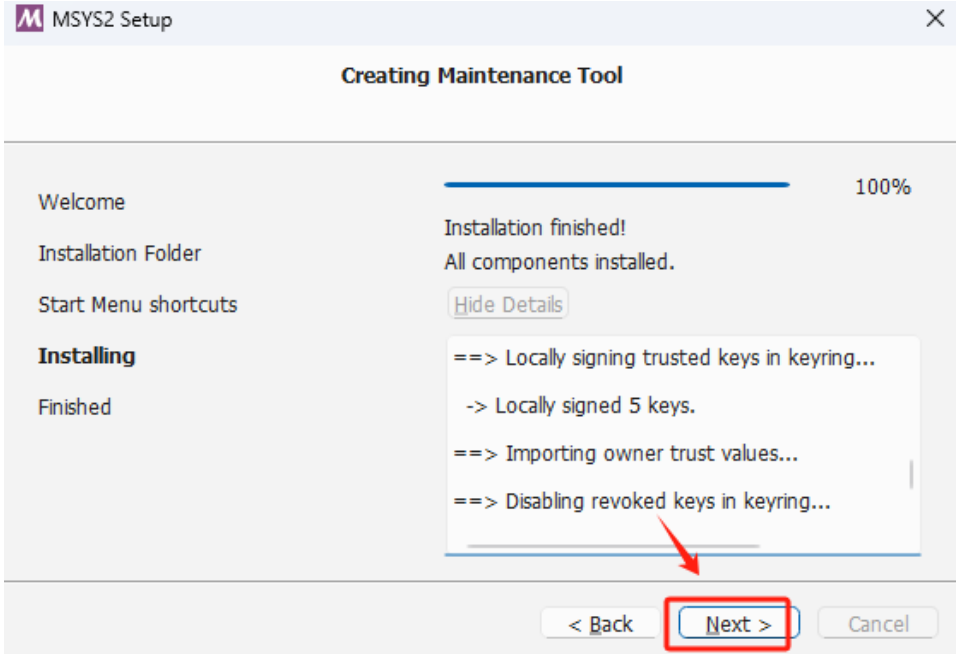
这里可以选择安装的路径，默认是`C:\msys64`，大家根据实际情况选择即可。我这里选择的是`D:\msys64`。



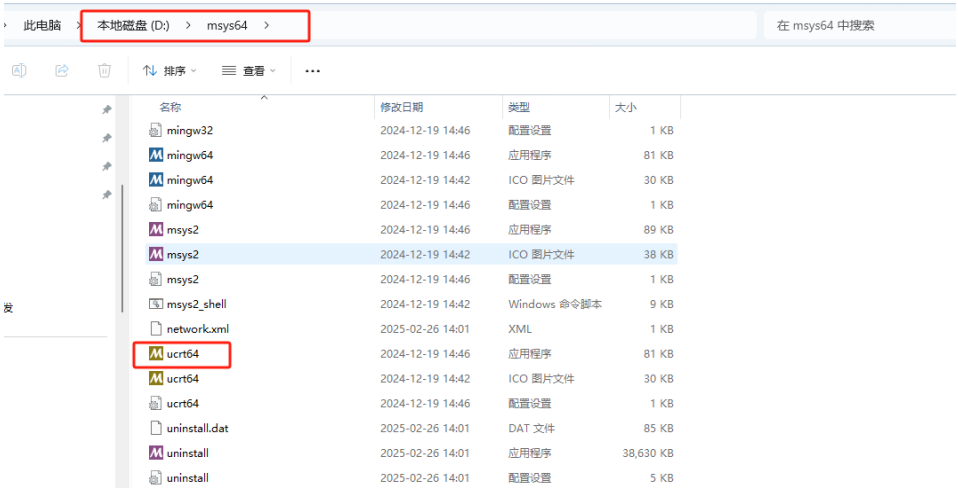
接下来选择默认的选项，点击`Next`。



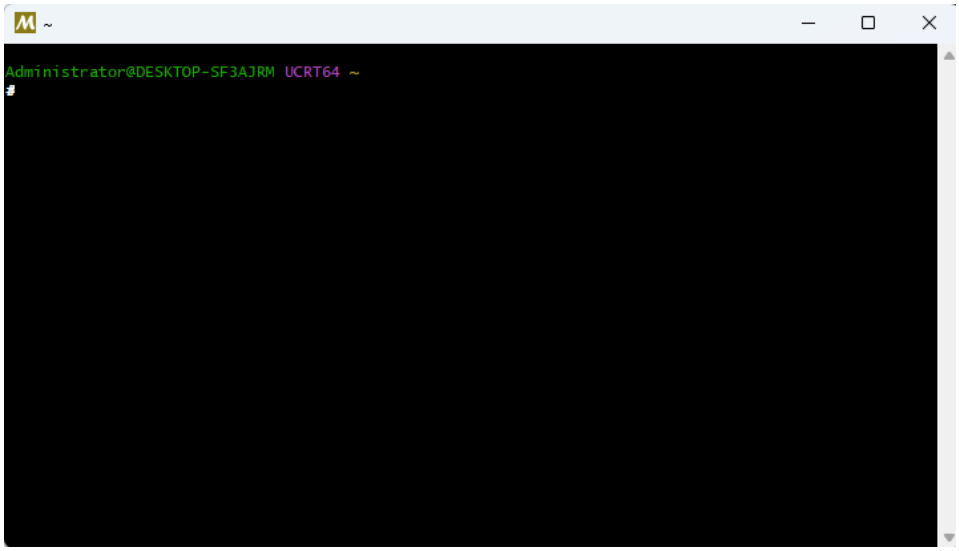
这一步的安装过程比较漫长，大家耐心等待。



等待安装后，我们可以在`D:\msys64`目录下找到`ucrt64.exe`文件，双击运行，即可打开`MSYS2`的命令行。



命令行界面如下所示，我们后续所有的操作都是在该命令行中进行。



2.2.2 Linux源码安装Redis

`Linux`系统上安装`Redis`，成功率最高的安装方式同样是源码编译安装的方式。具体的操作方法如下：

- Step 1. 下载`Redis`源码文件

首先通过`XShell`或者`FinalShell`连接到`Linux`服务器，安装编译打包需要的工具。命令如下：

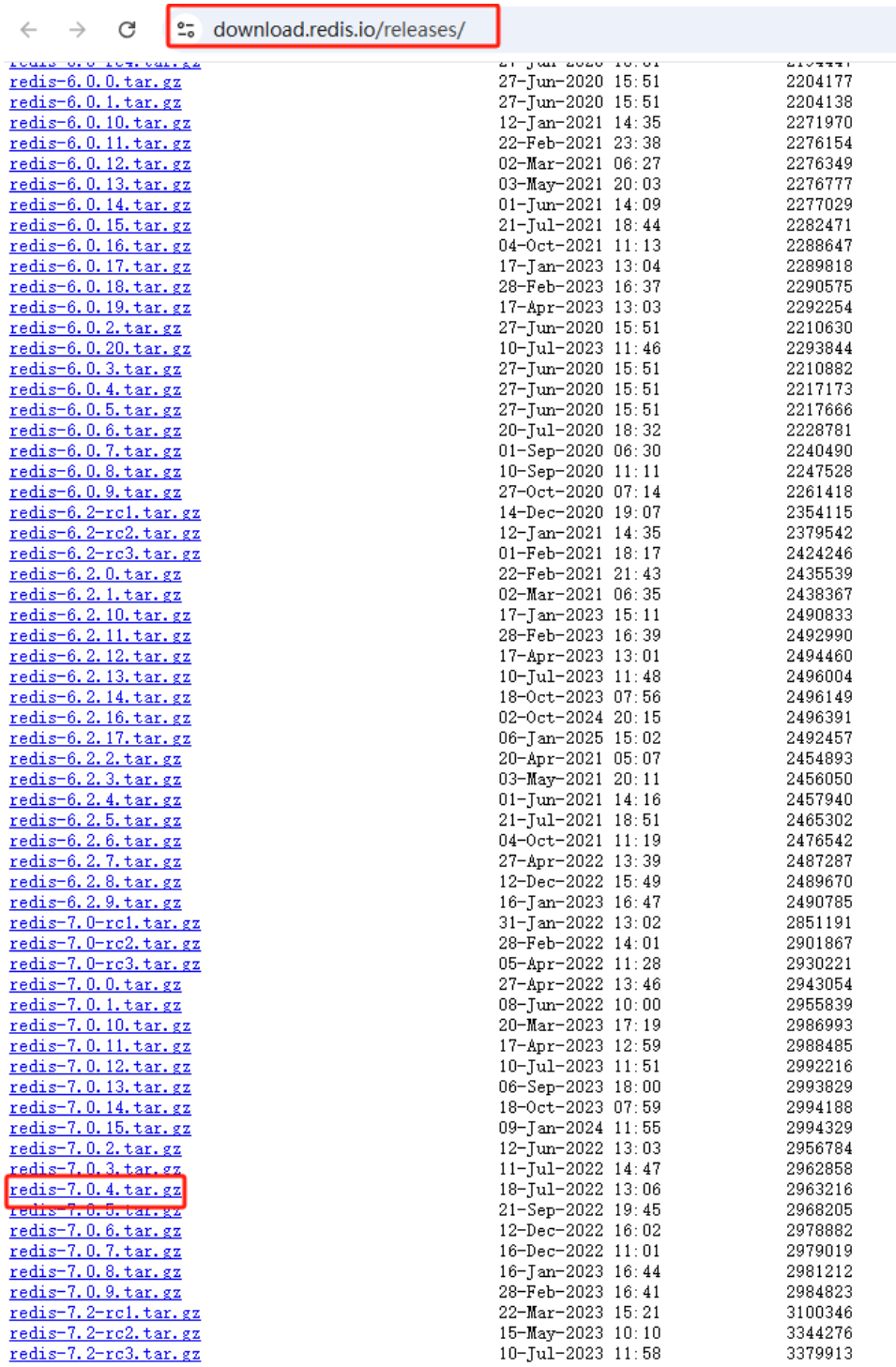
```
yum install -y gcc make tcl wget  # linux系统安装gcc、make、tcl、wget工具

apt-get install -y gcc make tcl wget  # ubuntu系统安装gcc、make、tcl、wget工具
```

我的操作系统是`Ubuntu`，所以使用的是`apt-get`命令。如果大家的操作系统是`CentOS`，则使用`yum`命令。

```
(base) root@4U:05_AssistGen# apt install -y gcc make tcl wget
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
gcc is already the newest version (4:11.2.0-1ubuntu1).
make is already the newest version (4.3-4.1build1).
tcl is already the newest version (8.6.11+1build2).
tcl set to manually installed.
wget is already the newest version (1.21.2-2ubuntu1).
0 upgraded, 0 newly installed, 0 to remove and 0 not upgraded.
(base) root@4U:05_AssistGen#
```

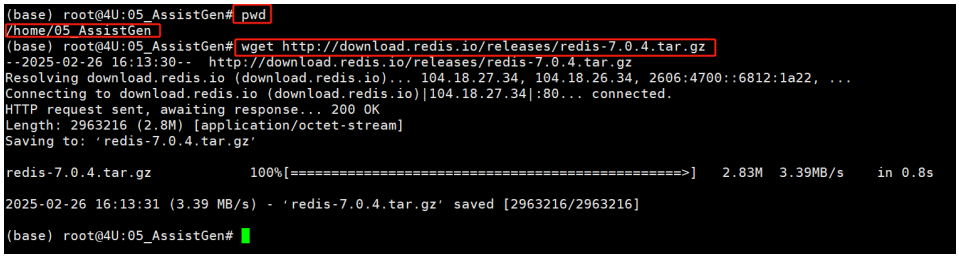
- Step 2. 下载`Redis`源码文件

`Redis`的所有发行版本官方的下载地址为Redis Releases，大家可以任意选择版本，但是不建议选择最新的版本，因为最新的版本可能存在一些问题。这里我们选择使用的版本是`7.0.4`。



在命令行界面。使用`wget`命令下载`Redis-7.0.4.tar.gz`的源码文件。命令如下：

```
pwd  # 查看当前目录
wget http://download.redis.io/releases/redis-7.0.4.tar.gz  # 下载redis源码，注意下载的位置为当前目录
```



如果大家的服务器没有外网的话，可以先下载到本地，然后上传到服务器。

- Step 3. 解压`Redis`源码文件

下载完成后，我们可以在当前目录下看到`redis-7.0.4.tar.gz`文件。使用`tar`命令解压文件。命令如下：

```
tar -zxvf redis-7.0.4.tar.gz
```

```
(base) root@4U:05_AssistGen# ll
total 2915
drwxr-xr-x  2 root root       3  2月  26 16:13 ./
drwxr-xr-x 20 root root      23  2月  24 15:05 ../
-rw-r--r--  1 root root 2963216  7月  18  2022 redis-7.0.4.tar.gz
(base) root@4U:05_AssistGen# tar -zxvf redis-7.0.4.tar.gz
redis-7.0.4/
redis-7.0.4/.codespell/
redis-7.0.4/.codespell/.codespellrc
redis-7.0.4/.codespell/requirements.txt
redis-7.0.4/.codespell/wordlist.txt
redis-7.0.4/.gitattributes
redis-7.0.4/.github/
```

- Step 4. 编译安装`Redis`

解压完成后，我们可以在当前目录下看到`redis-7.0.4`文件夹。使用`cd`命令进入`Redis`源码目录。命令如下：

```
cd redis-7.0.4
```

构建`Redis`依赖库, 执行如下命令：（如果大家有多个`CPU`，可以添加`-j`参数，比如`make -j4`，表示使用4个`CPU`进行编译。）

```
cd deps; make -j4 hiredis jemalloc linenoise lua
```

```
(base) root@4U:redis-7.0.4# cd deps; make -j4 hiredis jemalloc linenoise lua
(cd hiredis && make clean) > /dev/null || true
(cd linenoise && make clean) > /dev/null || true
(cd lua && make clean) > /dev/null || true
(cd jemalloc && [ -f Makefile ] && make distclean) > /dev/null || true
(cd hdr_histogram && make clean) > /dev/null || true
(rm -f .make-*)
(echo "" > .make-cflags)
(echo "" > .make-ldflags)
MAKE hiredis
cd hiredis && make static
MAKE jemalloc
cd jemalloc && ./configure --with-version=5.2.1-0-g0 --with-lg-quantum=3 --with-jemall
-Wall -pipe -g3 -O3 -funroll-loops " LDFLAGS=""
MAKE linenoise
cd linenoise && make
MAKE lua
```

依次执行：

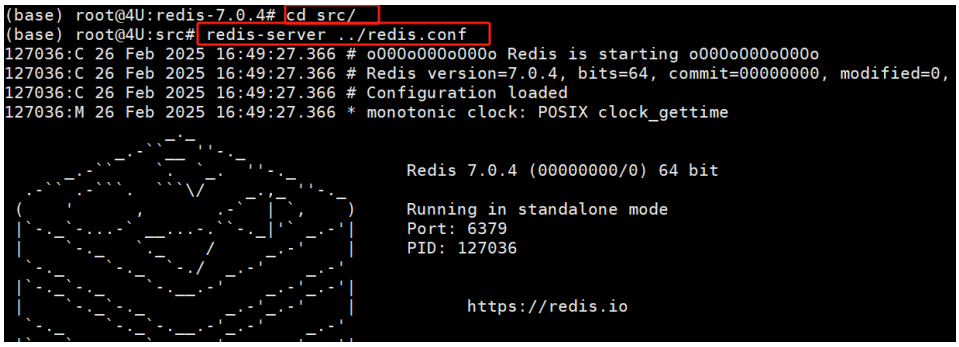
```
cd ..    # 回到redis源码目录
make install  # 安装redis
make clean  # 清理编译文件
```

```
(base) root@4U:deps# cd ..
(base) root@4U:redis-7.0.4# make install
cd src && make install
make[1]: Entering directory '/home/05_AssistGen/redis-7.0.4/src'
    CC Makefile.dep
rm -rf redis-server redis-sentinel redis-cli redis-benchmark redis-check-rdb redis-ch
edis.info lcov-html Makefile.dep
rm -f adlist.d quicklist.d ae.d anet.d dict.d server.d sds.d zmalloc.d lzf_c.d lzf_d.
.d release.d networking.d util.d object.d db.d replication.d rdb.d t_string.d t_list.
.d aof.d pubsub.d multi.d debug.d sort.d intset.d syncio.d cluster.d crc16.d endianco
rand.d memtest.d syscheck.d crcspeed.d crc64.d bitops.d sentinel.d notify.d setprocti
```

- Step 5. 启动`Redis`服务

安装完成后，通过`redis-server`命令指定`Redis`的配置文件`redis.conf`，启动`Redis`服务。命令如下：

```
cd src   # redis-server文件所在目录
redis-server ../redis.conf  # 指定redis.conf文件
```



至此，在`Linux`系统上，`Redis`的源码安装就完成了。并且成功启动了`Redis`服务。

- Step 6. 连接`Redis`服务测试

启动`Redis`服务后，后端服务需要通过`Python`的`redis`库来连接`Redis`服务进行对其相关的操作。其调用示例在官方有较为详细的文档：https://redis.readthedocs.io/en/stable/examples.html，我们这里先进行`Redis`的连接测试。

这里需要关注的关键连接信息是：`Redis`的默认连接地址是`127.0.0.1`，端口号是`6379`，数据库是`0`。在`Redis`中，数据库 0 是默认的数据库。`Redis`默认提供了 16 个逻辑数据库，编号从 0 到 15。每个数据库都是独立的，具有自己的键空间。

```
# pip install redis  # 安装redis库

import redis

r = redis.Redis(host='localhost', port=6379, db=0)
```

建立连接后，我们就可以对`Redis`进行操作了。 相关的增删改查示例都非常简单，代码如下：

```
# 增：添加数据
def add_data(key, value):
    r.set(key, value)            # set：设置键值对
    print(f"添加数据: {key} -> {value}")

# 查：获取数据
def get_data(key):
    value = r.get(key)          # get：获取键值对
    if value is not None:
        print(f"获取数据: {key} -> {value.decode('utf-8')}")
    else:
        print(f"键 {key} 不存在")

# 改：更新数据
def update_data(key, new_value):
    if r.exists(key):            # exists：判断键是否存在
        r.set(key, new_value)    # set：设置键值对，如果键存在，则更新键值对
        print(f"更新数据: {key} -> {new_value}")
```

```
    else:
        print(f"键 {key} 不存在，无法更新")

# 删：删除数据
def delete_data(key):
    if r.delete(key):    # delete：删除键值对
        print(f"删除数据: {key}")
    else:
        print(f"键 {key} 不存在，无法删除")
```

可以调用`add_data`、`get_data`、`update_data`、`delete_data`函数来对`Redis`进行操作。

```
# 添加数据
add_data("name", "juju")
add_data("age", "30")
```

```
添加数据: name -> juju
添加数据: age -> 30
```

```
# 查数据
get_data("name")
get_data("age")
```

```
获取数据: name -> juju
获取数据: age -> 30
```

```
# 改数据
update_data("age", "31")

# 查更新后的数据
get_data("age")
```

```
更新数据: age -> 31
获取数据: age -> 31
```

```
# 删数据
delete_data("name")

# 查已删除的数据
get_data("name")
```

```
删除数据: name
键 name 不存在
```

2.2.3 Redis配置项修改

`Redis`的配置文件为`redis.conf`，大家可以在`Redis`的源码目录下找到该文件。该配置文件中的内容非常多，这里我们介绍如下几个重要的配置项：

1. `bind`：指定`Redis`的监听地址，默认是`127.0.0.1`，如果需要监听所有地址，则可以设置为`0.0.0.0`。比如大家使用云服务器部署`Redis`，想要在自己的`IP`地址上进行访问，则可以设置为`0.0.0.0`，否则会无法访问到`Redis`服务。
2. `port`：指定`Redis`的监听端口号，默认是`6379`。如果大家想要修改端口号，则可以设置为其他端口号。
3. `requirepass`：指定`Redis`的密码，默认是`""`，如果大家想要设置密码，则可以设置为其他密码。
4. `protected-mode`：指定`Redis`的保护模式，默认是`yes`，如果大家想要关闭保护模式，则可以设置为`no`。

修改的方法也非常简单，大家只需要在`redis.conf`文件中找到对应的配置项，然后进行修改即可。比如在`Linux`系统上，可以通过`vim`命令修改`redis.conf`文件。命令如下：

```
vim redis.conf
```

```
# running on).
#
# IF YOU ARE SURE YOU WANT YOUR INSTANCE TO LISTEN TO ALL THE INTERFACES
# COMMENT OUT THE FOLLOWING LINE.
#
# You will also need to set a password unless you explicitly disable protected
# mode.
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
bind 0.0.0.0

# By default, outgoing connections (from replica to master, from Sentinel to
# instances, cluster bus, etc.) are not bound to a specific local address. In
# most cases, this means the operating system will handle that based on routing
# and the interface through which the connection goes out.
#
# Using bind-source-addr it is possible to configure a specific address to bind
```

找到对应的配置项，然后进行修改，然后保存退出。重新启动`Redis`服务，即可生效。

2.2 Redis 在模型调用中的缓存应用

在部署并连接了`Redis`服务后，我们就可以在模型调用中使用`Redis`的缓存机制了。

现在我们就来实现这个需求：当用户将问题（Prompt）发送到后端服务时，先检查缓存中是否可用的响应。如果检索到了，直接返回缓存中的响应。否则，再向`API`发送新的请求，得到最终的响应。

这个机制对于需要重复或者常用的查询或者问题很有效，具体的执行处理逻辑如下：

1. 首先将用户输入进来的`Prompt`生成一个唯一的哈希键，作为缓存键；
2. 使用这个缓存键在`Redis`中查找是否存在对应的缓存响应；（即某个`Prompt`是否已经调用过模型，并且将响应结果存储到了缓存中）
3. 如果有缓存命中率（找到响应），它将立即返回缓存的响应，不再去重新执行大模型服务调用；
4. 如果未找到缓存错过（不存在响应），则将`Prompt`提示发送到大模型生成响应；
5. 将生成的响应存储到`Redis`中，以便后续的请求可以快速返回；
6. 返回生成的响应给用户。

这里首先使用`DeepSeek`的`DeepSeek-Chat`模型来实现基于`Redis`的缓存机制。现在来看一下具体的代码实现。

首先，先连接`Redis`服务，代码如下：

```
import redis

# 连接到 Redis
try:
    r = redis.Redis(
        host='192.168.110.131',   # 这里的`host`和`port`需要根据实际情况进行修改
        password='g1601522830',
        port=6379,
        db=0)

    # 测试连接
    r.ping()  # 发送 PING 命令以测试连接
    print("成功连接到 Redis! ")
except redis.ConnectionError:
    print("无法连接到 Redis! 请检查 Redis 服务是否正在运行。")
except Exception as e:
    print(f"发生错误: {e}")
```

```
成功连接到 Redis!
```

如果看到`成功连接到 Redis!`，则表示`Redis`服务连接成功。否则需要检查`Redis`服务是否正常运行，以及`host`和`port`是否正确。

- 调用`DeepSeek-Chat`模型，并对输入的`Prompt`进行哈希处理，生成唯一的哈希键。

哈希的作用是：将输入的`Prompt`转换为一个唯一的哈希值。针对不同长度的`Prompt`，其生成的哈希值是等长的。虽然哈希值本身并不包含原始文本的信息，但它可以唯一标识特定的输入，避免重复存储相同的内容。。

```
import hashlib

question_1 = "你好，我是居居，很高兴认识你"

# 因为哈希函数需要处理字节数据，而不是字符串，所以需要通过 encode 转换为字节串（bytes）
md5_hash = hashlib.md5(question_1.encode('utf-8')).hexdigest()  # 返回的是一个32位的十六进制字符串
print(md5_hash)

question_2 = "哈哈，你好"
md5_hash = hashlib.md5(question_2.encode('utf-8')).hexdigest()
print(md5_hash)
```

```
f6c96a1c978407d928bd19e7f13ee18f
e522ecc43cab6ba1e2abf2eb11b59691
```

哈希值的长度是固定的，不管输入的`Prompt`有多长，其生成的哈希值的长度都是固定的。同时，同样的`Prompt`，其生成的哈希值是唯一的。

```python
import hashlib

question_1 = "你好，我是居居，很高兴认识你"
md5_hash = hashlib.md5(question_1.encode('utf-8')).hexdigest()
print(md5_hash)
```

```
f6c96a1c978407d928bd19e7f13ee18f
```

了解了基本原理，我们实际调用`DeepSeek-Chat`模型，并对输入的`Prompt`进行哈希处理，生成唯一的哈希键。代码如下：

```python
import hashlib
import redis
from openai import OpenAI
from dotenv import load_dotenv
import os

# 加载 .env 文件
load_dotenv()

# 实例化 DeepSeek API 客户端
client = OpenAI(
    api_key=os.getenv('DEEPSEEK_API_KEY'),
    base_url=os.getenv('DEEPSEEK_BASE_URL')
)


def hash_question(question):
    """使用 MD5 对问题进行哈希，并添加前缀"""
    # 计算 MD5 哈希
    md5_hash = hashlib.md5(question.encode('utf-8')).hexdigest()
    # 添加前缀
    return f"{'prefix:'}{md5_hash}"

def call_deepseek_chat(question):
    """调用 deepseek-chat 模型"""
    # 生成哈希键
    hashed_question = hash_question(question)

    # 调用模型
    response = client.chat.completions.create(
        model="deepseek-chat",
        messages=[{"role": "user", "content": question}]
    )

    # 获取模型的回答
    answer = response.choices[0].message.content

    return hashed_question, answer
```

进行模型调用，得到单次请求下哈希键和模型响应。代码如下：

```
question = "你好，我是居居，很高兴认识你"
hashed_question, answer = call_deepseek_chat(question)
print(f"哈希键: {hashed_question}")
print(f"模型响应: {answer}")
```

- 检查Redis是否缓存响应

这种情况下，因为还没有把上述的哈希键和模型响应存储到`Redis`中，所以缓存中不存在对应的响应。代码如下：

```
cached_response = r.get(hashed_question)
if cached_response:
    print(f"从缓存中获取响应: {cached_response.decode('utf-8')}")
else:
    print("缓存中不存在对应的响应")
```

```
缓存中不存在对应的响应
```

- 添加哈希键和模型响应到Redis

```
r.set(hashed_question, answer)
```

```
True
```

- 再次查询Redis中是否存在对应的缓存响应

```
cached_response = r.get(hashed_question)
if cached_response:
    print(f"从缓存中获取响应: {cached_response.decode('utf-8')}")
else:
    print("缓存中不存在对应的响应")
```

可以看到，现在缓存中已经存在对应的响应了。这样当下次提交相同的提示时，`Redis`将返回缓存的响应，而不是查询`API`，从而降低`API`延迟和成本。

因此，在了解了分步处理的过程后，我们接下来将完整的流程整合到一起，并进行测试。代码如下：

- 使用哈希键在`Redis`中查找是否存在对应的缓存响应

```
import hashlib
import redis
from openai import OpenAI
from dotenv import load_dotenv
import os
import time

# 加载 .env 文件
load_dotenv()
```

```python
# 实例化 DeepSeek API 客户端
client = OpenAI(
    api_key=os.getenv('DEEPSEEK_API_KEY'),
    base_url=os.getenv('DEEPSEEK_BASE_URL')
)

# 连接到 Redis
r = redis.Redis(
    host='192.168.110.131',   # 这里的`host`和`port`需要根据实际情况进行修改
    password='g1601522830',
    port=6379,
    db=0)

# 添加前缀
PREFIX = "deepseek:"

def hash_question(question):
    """使用 MD5 对问题进行哈希，并添加前缀"""
    md5_hash = hashlib.md5(question.encode('utf-8')).hexdigest()
    return f"{PREFIX}{md5_hash}"

def call_deepseek_chat(question):
    """调用 deepseek-chat 模型"""
    # 生成哈希键
    hashed_key = hash_question(question)

    # 开始计时
    start_time = time.time()

    # 检查缓存
    cached_response = r.get(hashed_key)
    if cached_response:
        # 解码缓存响应
        cached_response_decoded = cached_response.decode('utf-8')
        elapsed_time = time.time() - start_time  # 计算命中缓存的时间
        print(f"命中缓存: {cached_response_decoded}")
        print(f"处理时间（命中缓存）: {elapsed_time:.4f} 秒")
        return cached_response_decoded

    # 如果缓存不存在，调用模型
    model_start_time = time.time()  # 记录调用模型的开始时间
    response = client.chat.completions.create(
        model="deepseek-chat",
        messages=[{"role": "user", "content": question}]
    )

    # 获取模型的回答
    answer = response.choices[0].message.content

    # 将结果存入缓存
    r.set(hashed_key, answer)

    model_elapsed_time = time.time() - model_start_time  # 计算调用模型的时间
    total_elapsed_time = time.time() - start_time  # 计算总时间
    print(f"调用模型并缓存响应: {answer}")
    print(f"处理时间（未命中缓存）: {model_elapsed_time:.4f} 秒")
    print(f"总处理时间: {total_elapsed_time:.4f} 秒")
```

```
    return answer


# 测试调用
if __name__ == "__main__":
    question = "什么是大模型？"
    response = call_deepseek_chat(question)
    # print(f"模型响应: {response}")
```

```
命中缓存: 大模型（Large Model）通常指的是具有大量参数的机器学习模型，尤其是在自然语言处理（NLP）领域中的大规模预训练语言模型。这些模型通过在大规模数据集上进行预训练，学习到丰富的语言表示，从而在各种下游任务中表现出色。

### 大模型的特点：
1. **参数量大**：大模型的参数量通常在数亿到数千亿之间。例如，OpenAI的GPT-3模型有1750亿个参数。
2. **预训练与微调**：大模型通常在大规模文本数据上进行预训练，学习通用的语言表示。然后，可以通过微调（Fine-tuning）来适应特定的任务，如文本分类、机器翻译、问答系统等。
3. **多任务能力**：由于大模型在预训练过程中学习了广泛的语言知识，它们通常能够处理多种任务，而不需要为每个任务单独设计模型。
4. **计算资源需求高**：训练和部署大模型需要大量的计算资源，包括高性能的GPU或TPU集群，以及大量的存储和内存。

### 常见的大模型：
- **GPT系列**：由OpenAI开发，包括GPT-2、GPT-3等，广泛应用于文本生成、对话系统等任务。
- **BERT**：由Google开发，采用双向Transformer架构，适用于文本分类、问答等任务。
- **T5**：由Google开发，将各种NLP任务统一为文本到文本的转换任务。
- **PaLM**：由Google开发，具有5400亿参数，是目前最大的语言模型之一。

### 应用场景：
- **自然语言理解**：如文本分类、情感分析、命名实体识别等。
- **自然语言生成**：如文本生成、对话系统、机器翻译等。
- **问答系统**：如智能客服、知识问答等。
- **代码生成**：如GitHub Copilot等工具，利用大模型生成代码。

### 挑战：
- **计算成本**：训练和部署大模型需要大量的计算资源，成本高昂。
- **数据需求**：大模型需要大规模的高质量数据进行预训练。
- **模型解释性**：大模型的决策过程往往难以解释，存在“黑箱”问题。
- **伦理与安全**：大模型可能生成有害内容或存在偏见，需要谨慎处理。

总的来说，大模型在推动人工智能技术发展的同时，也带来了新的挑战和机遇。
处理时间（命中缓存）: 0.0020 秒
```

通过上面的示例，当用户提交相同的提示（问题）时，`Redis`将返回缓存的响应，而不是去请求大模型服务的`API`，从而使响应时间大大缩短。这就是借助`Redis`的缓存机制可以达到的效果。

但是目前实现的这个版本存在一个问题：我们通过`MD5`对`Prompt`进行哈希处理，同时也是根据这个哈希值作为缓存键进行存储和查询，所以意味着一旦`Prompt`发生变化，其哈希值就会发生变化，从而导致缓存不到结构。比如:"什么是大模型？"、"大模型是什么？"，这两个`Prompt`的哈希值是不同的，但其实意思是完全一样的。

因此，更加常用的一种方法是在匹配过程中融入动态的语义相似度匹配。

2.3 基于语义相似度匹配的的 Redis 缓存实现

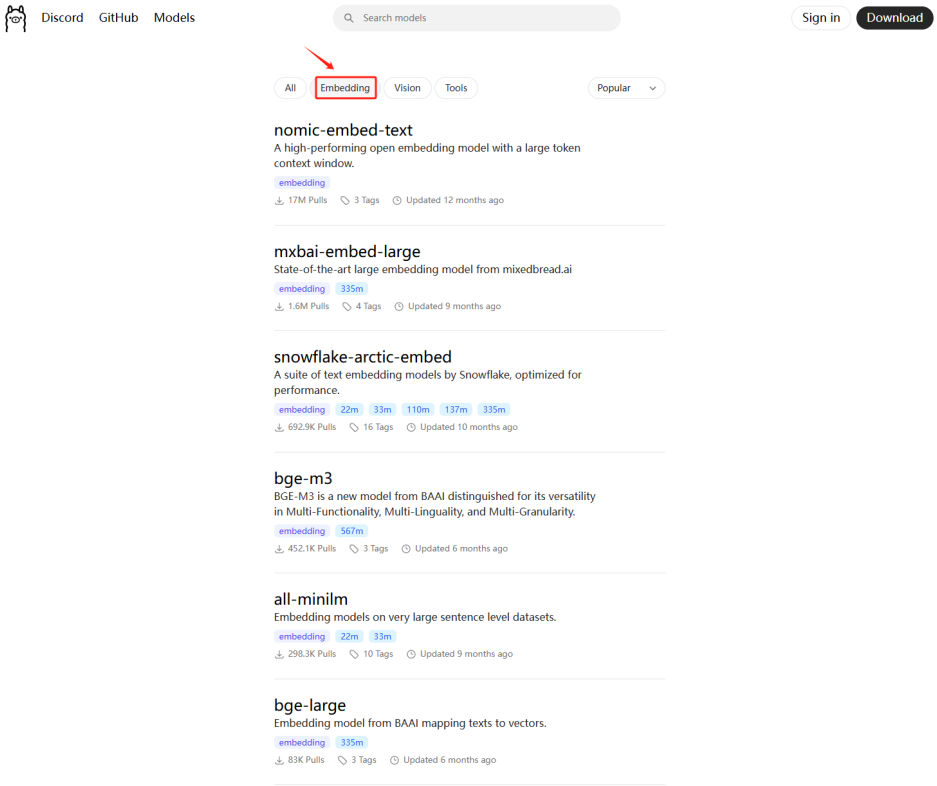
正如我们提出的这个问题：“什么是大模型？”、“大模型是什么？”、“如何理解大模型？”，这三个 `Prompt` 的语义是基本一致的，完全可以用相同的响应来进行回复。那么如何在实际的请求中将这种类似的问题进行匹配，从而达到命中缓存的效果呢？

解决的方法很容易想到：不同文本间的语义相似度可以借助 `Embedding` 来实现。通过将文本转化成向量，然后计算向量之间的余弦相似度，从而判断两个文本是否相似。如果相似，则认为它们是同一个问题，就可以直接提取已缓存问题的响应，尽管它们的 `Prompt` 不同。（哈希值也不同）

`Embedding` 模型像对话模型一样，同样可以使用在线 `API` 版本或者借助 `Ollama` 框架启动本地版本，并通过 `REST API` 来进行使用。这里我们就使用 `Ollama` 启动的本地 `Embedding` 来给大家进一步讲解实际的实现流程。

2.3.1 使用 Ollama 接入 Embedding 模型

`Ollama` 除了可以接入对话、推理类模型外，还可以接入 `Embedding` 模型。并且部署和使用方法基本与对话、推理类模型保持一致。官方下载地址：https://ollama.com/search?c=embedding，可点击链接进行查看：



关于如何选择`Embedding`模型，这个问题本质上不存在哪个`Embedding`最好的说法，同时也并没有比较通用且大家都认可的评测数据、流程等，往往还是需要结合自己的实际数据情况加上构建流程评测出来的效果，来进行综合评估。所以这里给大家推荐一个相对完善的`Embedidng`模型评估开源项目，同时也是一个`RAG`的解决方案：https://github.com/timescale/pgai?ref=timescale.ghost.io

pgai

Power your AI applications with PostgreSQL

Docs · Join the pgai Discord! · Try timescale for free! · Changelog

Supercharge your PostgreSQL database with AI capabilities. Supports:

- Automatic creation and synchronization of vector embeddings for your data
- Seamless vector and semantic search
- Retrieval Augmented Generation (RAG) directly in SQL
- Ability to call out to leading LLMs like OpenAI, Ollama, Cohere, and more via SQL.
- Built-in utilities for dataset loading and processing

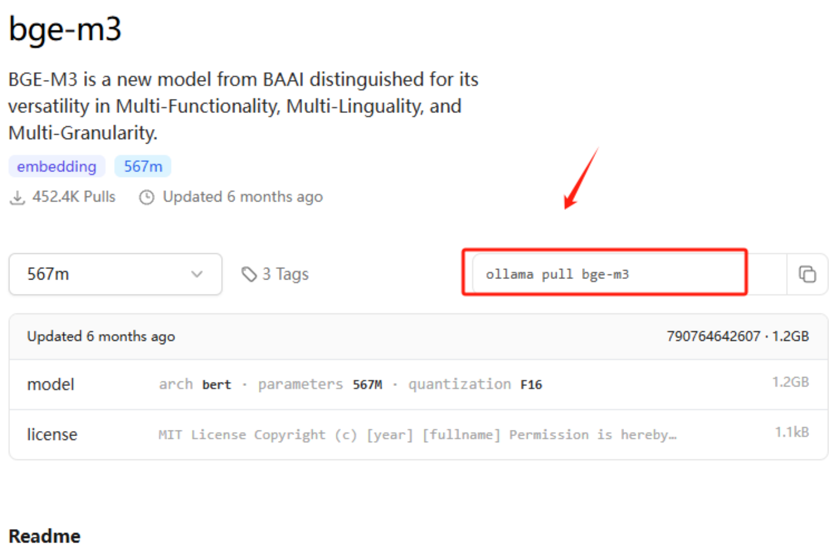
All with the reliability, scalability, and ACID compliance of PostgreSQL.

该开源项目针对`Ollama`支持的`Embedding`模型做了一些基础的评测，如下：

常规参数

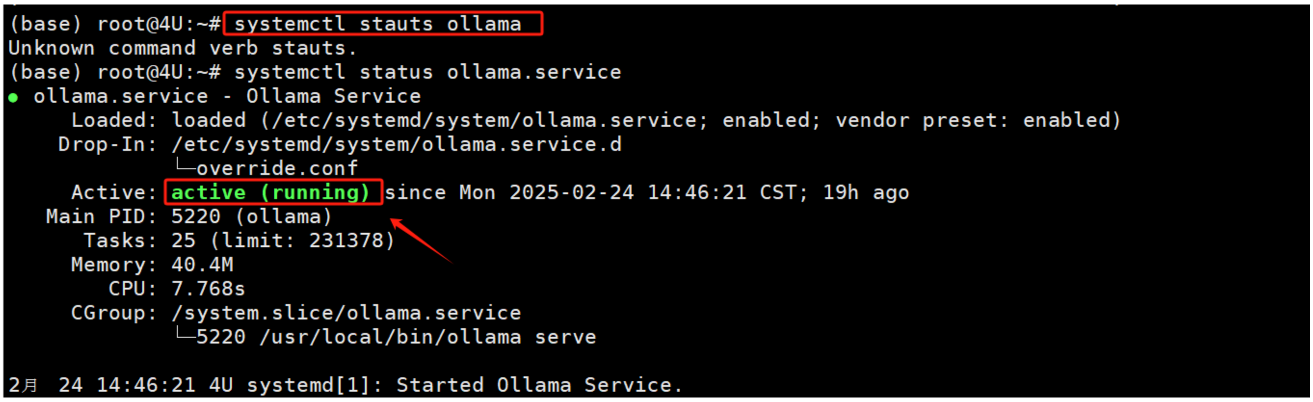
模型名称	优势	劣势
bge-m3	整体检索准确率最高	在不清楚和含糊不清的问题上表现较差，最低准确率
	在长问题上表现出色	
mxbai-embed-large	尺寸小	在简短和直接的问题上表现不如其他模型
	在上下文较重的问题上表现良好	
	长问题表现良好	
nomic-embed-text	在简短和直接的问题上表现优异	整体表现排名最后
	在长问题上也取得了较好性能	

这里我们选择`bge-m3`模型进行下载使用。点击具体的模型，可以查看模型的详细信息，如模型的大小、支持的设备、支持的参数等。



首先需要确定服务器上的`ollama`服务处于运行状态，通过如下命令查看：

```
systemctl status ollama.service
```



如果服务未处于运行状态，则可以通过如下命令启动：

```
systemctl start ollama.service
```

启动`ollama`服务后，通过如下命令下载并启动`bge-m3`模型：

```
ollama run bge-m3
```

需要说明的是，最后一行`Error: "bge-m3" does not support generation`，表示`bge-m3`模型不支持在类似对话模型启动后可以在命令行进行问答，需要通过`ollama`的`REST API`进行使用。

2.3.2 Ollama Embedding REST API 使用

`Ollama` 提供给 `Embedding` 模型使用的 `REST API` 接口为：`/api/embed`。其可以支持单个输入和多个输入的请求。具体可用的参数如下所示：

常规参数

参数名	类型	描述
model	字符串	用于生成嵌入的模型名称
input	字符串或字符串列表	要生成嵌入的文本或文本列表
truncate	布尔值	是否截断每个输入的末尾以适应上下文长度。若为 false 且超出上下文长度则返回错误。默认值为 true
options	对象	额外的模型参数，例如温度等
keep_alive	字符串	控制模型在请求后保持加载到内存中的时间（默认：5分钟）

同样，只要了解了某个服务的 `REST API` 接口以及其请求参数，就都可以通过 `Python` 的 `requests` 库进行调用。代码如下：

```python
import requests
import json

# 定义 API 端点
url = "http://192.168.110.131:11434/api/embed"    # 这里需要根据实际情况进行修改

# 单个输入的请求示例
single_input_payload = {
    "model": "bge-m3",   # 这里替换成具体的模型名称
    "input": "你好，我是居居，很高兴认识你"   # 这里替换成具体的输入文本
}

# 发送 POST 请求
response_single = requests.post(url, json=single_input_payload)

# 检查响应
if response_single.status_code == 200:
    print("Single Input Response:")
    response_data_single = response_single.json()
    print(json.dumps(response_data_single, indent=2))
else:
    print(f"Error: {response_single.status_code} - {response_single.text}")
```

```
Single Input Response:
{
  "model": "bge-m3",
  "embeddings": [
    [
      -0.01922286,
      -0.041361455,
```

-0.012615179,
-0.038346406,
-0.006061248,
-0.056141857,
0.024635639,
-0.048374478,
0.021608315,
0.028389402,
0.011415558,
-0.008346135,
-0.01575168,
0.011952146,
0.0151369395,
-0.0076897917,
0.019403733,
-0.03635689,
0.01607708,
-0.030296508,
-0.018613543,
-0.0064844512,
-0.00029064808,
-0.010478618,
0.029013928,
0.01946554,
-0.055293787,
-0.0024098635,
0.025135659,
-0.04489269,
0.027139282,
0.015564225,
0.014931622,
-0.04502287,
-0.023891093,
-0.024381446,
0.033852994,
-0.0074271495,
-0.035470333,
-0.015253574,
0.038780138,
-0.010637472,
0.050409704,
-0.047934886,
-0.0071231266,
-0.02649895,
-0.049684074,
0.01412076,
0.008318819,
-0.0024914725,
-0.014517145,
-0.006007267,
0.04962171,
-0.0011974386,
-0.045221522,
0.026687985,
0.036642727,
-0.018013492,
0.0062242304,
-0.037533417,

```
        -0.026383227,
        0.0430352,
        0.023990763,
        -0.046562776,
        0.005052609,
        0.1078952,
        -0.0252689,
        -0.036592163,
        -0.017526496,
        -0.020836545,
        -0.014016081,
        0.0073044035,
        0.018462053,
        0.009727984,
        -0.05348882,
        0.03130105,
        0.054758947,
        0.019654304,
        -0.070866816,
        -0.009440621,
        0.052959364,
        0.028022723,
        0.020885868,
        -0.0027755883,
        0.025919601,
        0.0028719697,
        -0.011645562,
        0.07296864,
        -0.040803645,
        -0.014456895,
        -0.025773542,
        -0.024834368,
        0.0054630465,
        -0.025666654,
        -0.027248958,
        -0.009498956,
        -0.02270712,
        0.03019623,
        0.04891002,
        0.03496023,
        0.032207686,
        0.022681091,
        0.0092448965,
        -0.020107258,
        -0.0018319886,
        -0.007373062,
        0.03264567,
        0.03045322,
        -0.018678445,
        -0.012370282,
        0.023889143,
        0.005763869,
        0.005892724,
        -0.036694903,
        0.0010225775,
        -0.011602502,
        -0.019340092,
        0.014320763,
```

0.0038706549,
-0.019690355,
0.0050612646,
0.042274926,
-0.0075149853,
-0.012850127,
-0.0029281029,
-0.030729791,
0.020727703,
0.0390613,
-0.009327623,
0.008809925,
0.03356942,
-0.012686815,
-0.030047603,
-0.042933784,
-0.015244054,
-0.023267841,
-0.0105991755,
0.01019861,
-0.022745306,
-0.028969925,
0.0044860677,
0.025056915,
0.0063823955,
-0.051997595,
-0.0007780814,
-0.051995378,
0.017508458,
0.0270951,
0.018337702,
-0.018740973,
-0.01517844,
0.0043489956,
-0.007722501,
0.0066714576,
0.03081158,
-0.013100324,
0.009403119,
-0.0133148255,
-0.00554864,
0.039081782,
-0.009327039,
-0.01150965,
0.0057221474,
-0.022880614,
0.03247104,
0.03626947,
0.003067755,
0.013796212,
0.007955853,
-0.04198045,
0.010368125,
-0.0032647224,
0.044448704,
0.027452532,
-0.0057273414,
0.030706208,

```
        0.07107367,
        -0.005020788,
        0.013032576,
        -0.04025715,
        -0.012651873,
        0.029171083,
        -0.008141952,
        -0.016902551,
        -0.04220599,
        0.025815476,
        -0.026521945,
        -0.04866242,
        -0.013637231,
        0.03486512,
        -0.001918829,
        -0.057317823,
        0.010962362,
        -0.020244285,
        0.03866528,
        0.010617372,
        -0.04377715,
        0.009748239,
        0.0045961896,
        -0.002983128,
        -0.004532409,
        -0.019298308,
        0.04492627,
        0.017812751,
        -0.053677455,
        -0.0078686215,
        -0.037468098,
        0.011929472,
        0.01720958,
        -0.07143857,
        -0.03330745,
        0.024078554,
        0.024672244,
        -0.0569919,
        -0.04148753,
        0.050901547,
        0.022151768,
        0.031624585,
        0.0021112408,
        0.0004811191,
        0.014531818,
        -0.008901751,
        0.034024063,
        0.02308123,
        -0.012181863,
        -0.037322327,
        -0.016135737,
        0.05219511,
        -0.00483262,
        -0.024687467,
        -0.007843138,
        0.0142526515,
        -0.0062834364,
        -0.0070895934,
```

```
        0.016342908,
        -0.014341934,
        -0.055698976,
        0.07432804,
        -0.025424266,
        0.027217498,
        -0.033968687,
        -0.019814044,
        -0.004727214,
        0.0032408026,
        -0.02595303,
        0.004324861,
        -0.0019504167,
        0.014354229,
        0.014305528,
        0.036802884,
        -0.008998354,
        -0.06352971,
        0.012633923,
        -0.033731535,
        0.028493801,
        0.00502899,
        0.014065273,
        0.024009604,
        -0.041306116,
        0.00876087,
        -0.0015096144,
        -0.0581046,
        0.009729838,
        0.035152882,
        0.0054527465,
        -0.0033466823,
        -0.007744983,
        -0.036980435,
        0.02000405,
        0.042424027,
        0.007387792,
        0.021873796,
        0.025885409,
        0.003218425,
        -0.0067987675,
        0.0067939996,
        -0.0058852965,
        -0.03311272,
        0.051925275,
        -0.038485028,
        0.017672395,
        0.02621819,
        0.042438682,
        -0.0080195675,
        0.052674558,
        0.014233261,
        -0.08290935,
        -0.00034058988,
        0.039673224,
        -0.04147254,
        -0.029645534,
        -0.009901535,
```

```
        0.033171214,
        -0.03572185,
        -0.012022275,
        0.0054901815,
        -0.0048021623,
        -0.1500592,
        0.0084257675,
        -0.008138095,
        -0.007993522,
        -0.032590296,
        -0.04603252,
        -0.06815151,
        -0.0007138535,
        -0.019334827,
        0.036172323,
        0.013635413,
        -0.013069142,
        0.014766935,
        0.0077957814,
        -0.032953933,
        0.023373697,
        0.033019446,
        -0.018915948,
        0.018228997,
        -0.0558767,
        -0.046538107,
        -0.021160431,
        0.09153756,
        -0.05778473,
        0.028429775,
        0.00013784457,
        -0.00088175054,
        -0.046915453,
        -0.050807092,
        -0.002632421,
        0.0011797646,
        0.0285695,
        -0.00094530155,
        0.015300199,
        0.03612954,
        0.027943866,
        -0.014544252,
        -0.016660748,
        -0.032984182,
        0.011127908,
        -0.008035259,
        0.021325957,
        -0.0030667118,
        0.0043126587,
        0.005962908,
        -0.00042244705,
        -0.0065966123,
        0.01868376,
        0.03252872,
        -0.021499867,
        -0.0038907684,
        0.0037884223,
        0.03002122,
```

-0.042780697,
-0.025766337,
0.011636676,
-0.007580283,
0.017507449,
-0.004628474,
0.038538925,
-0.040182322,
-0.027409792,
0.043854233,
0.048351236,
-0.03091617,
0.0027391338,
0.00036530296,
0.012099772,
0.022873694,
-0.012164523,
0.028665284,
0.01798609,
0.031127907,
0.013131891,
-0.01598989,
0.039874513,
0.0009227839,
-0.060541965,
-0.024316201,
-0.097156905,
0.0062047048,
-0.02919612,
-0.017234696,
0.032315448,
-0.0012223435,
0.03263769,
-0.0027417927,
0.036068577,
0.062720075,
0.22184578,
0.019815253,
0.01993091,
-0.043113314,
0.05674825,
-0.0155262025,
-0.020943452,
0.029482389,
-0.023183936,
-0.004876635,
0.0021091257,
-0.02630077,
-0.017170096,
0.029073019,
0.016285948,
0.049894232,
-0.0069553154,
0.0022705528,
0.03613966,
-0.031493917,
-0.0062867478,
-0.029011935,

-0.037524033,
-0.021112205,
-0.086708285,
-0.040370315,
-0.014455863,
0.042079657,
-0.0327298,
0.009579455,
-0.013825583,
0.033582326,
-0.0028413108,
0.002097946,
0.026424421,
-0.002604664,
0.03141509,
-0.042284306,
-0.020693991,
0.052990846,
-0.01812236,
-0.025489392,
0.039961986,
-0.010502693,
-0.00163735,
0.024143638,
0.008718653,
0.0010326977,
-0.013266786,
-0.006186287,
0.04246847,
-0.008784686,
0.012570915,
-0.020447822,
-0.012151716,
-0.047826536,
-0.039175574,
-0.020817315,
0.025246646,
0.027339023,
0.025236443,
0.048806693,
-0.011991174,
-0.022016289,
-0.011961466,
0.0012723304,
0.008125226,
-0.035645425,
0.011581718,
0.0074978457,
-0.05309759,
-0.010962783,
-0.016122645,
-0.0047364915,
0.010034488,
0.0030421098,
0.010316423,
0.038907442,
0.005337137,
0.057604548,

```
        -0.024876388,
        0.013173044,
        -0.062358152,
        0.055846933,
        0.019336209,
        -0.0069100326,
        0.0024203337,
        0.04047966,
        0.014952066,
        -0.025101326,
        0.025459653,
        -0.007939787,
        -0.05776767,
        0.0052549858,
        -0.012580737,
        -0.06010769,
        0.010566356,
        -0.023706116,
        -0.004249546,
        0.018867647,
        -0.028295763,
        -0.0071428707,
        0.00011242026,
        0.07138021,
        -0.0046134833,
        -0.004373349,
        0.024181986,
        0.0366532,
        -0.015330882,
        0.076670185,
        -0.044471033,
        0.0236505,
        -0.02199792,
        0.047300745,
        0.006907518,
        -0.031923965,
        -0.0044949916,
        -0.034935955,
        0.061860606,
        -0.012920158,
        -0.009882616,
        -0.017172387,
        -0.030343372,
        0.027154673,
        -0.012200059,
        0.002181584,
        0.05755524,
        0.030812616,
        -0.0052554836,
        0.030308068,
        0.011736867,
        -0.011554788,
        0.019779095,
        0.040169403,
        0.05081635,
        0.014287744,
        0.042208575,
        0.03118895,
```

```
        0.02718363,
        -0.015795866,
        -0.03593932,
        0.029518906,
        -0.0368933,
        0.021886608,
        -0.02643343,
        -0.027600609,
        -0.024287364,
        -0.021399602,
        -9.439248e-05,
        -0.0064367475,
        0.039554097,
        0.04583417,
        0.007510894,
        -0.07607565,
        -0.009225462,
        -0.006582759,
        -0.0148243345,
        -0.01328466,
        0.030134458,
        -0.027474036,
        -0.003093169,
        -0.007971443,
        0.040055204,
        0.13162026,
        -0.02346389,
        0.014800054,
        -0.047165684,
        -0.017866734,
        0.012297593,
        -0.009757697,
        -0.014321117,
        -0.01788238,
        -0.039049163,
        0.03508187,
        -0.0022837762,
        -0.007907967,
        -0.00075374526,
        -0.018473191,
        0.008784031,
        0.013422078,
        0.011996075,
        0.015082026,
        -0.039035235,
        0.02770337,
        0.00021796707,
        0.008747956,
        0.011951045,
        -0.0045034382,
        0.02824308,
        -0.036295313,
        0.00022648639,
        0.082074136,
        -0.008459332,
        -0.026773563,
        0.021175345,
        0.005603807,
```

```
        0.011868295,
        -0.027673401,
        -0.030666795,
        -0.01869547,
        0.017391281,
        0.03777573,
        -0.021297984,
        -0.0445547,
        0.0042566187,
        -0.023678964,
        0.0066275527,
        0.043544076,
        -0.023232576,
        -0.018033905,
        -0.016615305,
        0.01890017,
        -0.016875628,
        0.023675755,
        -0.013777414,
        -0.025639104,
        0.012749661,
        0.008193755,
        -0.029998936,
        -0.016121306,
        -0.013959011,
        0.021408115,
        0.023548564,
        0.0012766648,
        0.0024227598,
        -0.00060618424,
        0.033602092,
        0.02401148,
        -0.021303771,
        0.05285596,
        -0.0058503565,
        0.012856821,
        0.00946074,
        -0.018707717,
        0.034906205,
        -0.0033552044,
        -0.042422313,
        0.027284978,
        -0.014717163,
        0.009164387,
        0.025540866,
        0.0018863472,
        0.037348382,
        0.026557129,
        0.027901705,
        0.05202844,
        0.020671928,
        0.007082176,
        0.0034235918,
        -0.031215776,
        -0.040995408,
        0.003343065,
        0.027570983,
        -0.015434854,
```

-0.036407907,
0.031241637,
-0.034739457,
-0.014406115,
-0.062684685,
0.010075276,
-0.031929497,
0.0025649823,
-0.021319512,
0.005636287,
0.026131155,
0.021893157,
0.027501272,
-0.067495435,
-0.016828451,
-0.017330559,
-0.027661966,
0.026027255,
-0.04029333,
-0.07741891,
0.04245963,
0.05786638,
-0.048196528,
0.08805092,
0.019176718,
0.043642513,
-0.016361246,
-0.008850222,
-0.03140456,
-0.0018215853,
-0.02850662,
-0.017091129,
-0.0021216967,
-0.0008238167,
0.014921121,
0.00032447535,
-0.044774458,
-0.04502461,
-0.0059552435,
0.061054993,
-0.04434121,
0.014774075,
0.00038209726,
0.0702063,
-0.021992086,
0.008455862,
-0.02410905,
0.0652755,
0.005741814,
0.0019593763,
0.0047087143,
-0.024068253,
0.012610589,
-0.06671961,
0.01753201,
-0.0034162577,
0.0023024643,
-0.054187473,

-0.011831421,
0.01715048,
0.04750781,
-0.03139423,
0.02508704,
-0.010289396,
0.020973045,
-0.014849384,
-0.027357204,
-0.01407778,
0.033853255,
-0.055280462,
-0.013809495,
0.024581378,
-0.012241737,
-0.0009671642,
-0.018698305,
0.028465739,
0.0010522557,
0.00916822,
-0.04007455,
-0.025335168,
-0.011387842,
0.0055432064,
0.053396042,
0.01740828,
-0.014137389,
0.02978183,
0.018102026,
0.05603387,
0.008667747,
-0.052618537,
0.07024653,
-0.052201025,
0.029271316,
0.027836822,
0.037358917,
-0.018538367,
-0.0098952325,
0.025678122,
-0.022600275,
-0.02866699,
0.047425956,
0.041606687,
-0.045320027,
0.009275773,
-0.015545199,
-0.012686884,
-0.018841615,
0.021218903,
-0.028941818,
0.027858598,
-0.028179156,
0.029352563,
9.719478e-05,
-0.008117893,
0.023555716,
0.0057699373,

```
        0.0073017688,
        0.038897607,
        0.0060669654,
        0.038502082,
        -0.03068005,
        0.050719038,
        -0.013903436,
        -0.043558348,
        -0.01102934,
        0.017825447,
        -0.015128274,
        -0.027200924,
        -0.018548822,
        -0.043312766,
        0.046628807,
        -0.0041413624,
        -0.015847849,
        4.6269724e-05,
        -0.024507571,
        0.059471335,
        -0.017492197,
        -0.015842618,
        -0.033435818,
        0.020426262,
        -0.15537652,
        0.022720942,
        -0.022978572,
        -0.0013574777,
        -0.03353008,
        -0.02227914,
        -0.04341785,
        -0.026962942,
        0.036519103,
        0.009219965,
        -0.040424213,
        -0.017088816,
        0.01368187,
        -0.03543528,
        0.04532406,
        0.011525667,
        0.032134064,
        -0.028047284,
        0.006712295,
        0.0066740294,
        0.019953199,
        -0.019357666,
        0.088674895,
        -0.01850979,
        0.025244767,
        -0.018174196,
        -0.042392906,
        0.023810916,
        -0.042712238,
        0.016077517,
        -0.0056865416,
        -0.007365048,
        -0.0026553639,
        0.05539162,
```

```
        0.052702192,
        0.012977329,
        -0.013324897,
        -0.046663057,
        -0.004664647,
        0.0038475136,
        -0.0083794715,
        0.008683904,
        -0.028441006,
        0.00988668,
        0.019565033,
        0.038489543,
        0.021747977,
        -0.020754416,
        -0.042819865,
        -0.03930985,
        0.01374576,
        -0.035338506,
        0.030846687,
        0.036640152,
        -0.037350155,
        0.0038905449,
        -0.020976484,
        0.0058610975,
        -0.00070023915,
        0.019661274,
        -0.054429367,
        -0.03100215,
        -0.052285925,
        -0.0120779425,
        -0.074244425,
        0.0006335962,
        -0.0017926287,
        0.0014389432,
        0.022021336,
        0.036196936,
        -0.011613051,
        -0.093202114,
        0.0026990643,
        -0.023806851,
        -0.0054916134,
        0.0011963513,
        0.03746827,
        -0.022631988,
        -0.051280633,
        -0.03702279,
        0.04518093,
        0.007951412,
        -0.038437307,
        -0.034927223,
        0.03933246,
        -0.015370903,
        -0.02635214,
        -0.0041535352,
        -0.02957769,
        0.012466689,
        -0.043131832,
        -0.020760499,
```

```
        0.033076085,
        0.061528783,
        -0.040835578,
        0.0050104596,
        -0.006039951,
        -0.054586615,
        -0.046932537,
        -0.016764142,
        -0.004739649,
        -0.01928498,
        -0.031564888,
        0.0028312965,
        -0.013649793,
        0.020705067,
        0.019288259,
        -0.00771178,
        0.015912732,
        -0.0185576,
        -0.027995558,
        0.021834064,
        -0.007156878,
        -0.012114014,
        -0.04397438,
        0.013293692,
        0.031676818,
        -0.0026619898,
        -0.017556915,
        0.06203819,
        -0.015809024,
        -0.0014322706,
        -0.014704928,
        -0.046601627,
        0.010122574,
        -0.01908295,
        -0.014350051,
        0.023351008,
        -0.009993024,
        -0.036077287,
        -0.0052948133,
        -0.06703926,
        0.0069603934,
        -0.011365105,
        0.018846352,
        -0.038462013,
        -0.00936936,
        0.0022846924,
        -0.03624788,
        -0.012948129,
        -0.013476335,
        -0.0182052,
        0.0033161482,
        0.0039851656,
        0.0019273587,
        -0.018899929,
        -0.0206676,
        0.022174593,
        -0.040895447,
        0.0038407452,
```

0.008989447,
0.060425438,
0.0058986675,
-0.013137046,
0.0025831743,
0.03484262,
0.011654857,
0.006968705,
0.040688474,
-0.001316991,
-0.019467473,
0.027343286,
-0.022299161,
-0.05269345,
-0.024993394,
-0.035053432,
-0.019827059,
0.030563287,
0.008705064,
-0.0045054746,
0.0034345994,
0.07223343,
-0.0068709,
-0.0078441575,
0.05365767,
-0.015098801,
-0.0019810034,
-0.008233295,
0.08352503,
0.018159185,
0.011276068,
0.030084705,
0.014642007,
0.007945344,
0.0064324383,
-0.021965735,
0.022328451,
0.0013204975,
-0.0054936996,
-0.0153985005,
0.018929703,
0.0011742727,
-0.026166951,
-0.006397983,
0.02537267,
0.039143644,
0.0054356274,
0.046194375,
-0.0054917075,
0.01610236,
-0.06036488,
-0.03880828,
0.01609657,
-0.0067152907,
-0.012388587,
0.020276126,
-0.013779218,
-0.0008071725,

```
      -0.013861764,
      0.024183014,
      -0.014108998,
      -0.0054775397,
      -0.014219473,
      0.023931691,
      0.036456257,
      0.029436398,
      0.02639856,
      -0.009838244,
      -0.017583802,
      0.0084198555,
      -0.011142715,
      0.0033209193,
      -0.044683803,
      0.03729297,
      0.025573304,
      -0.0038417936,
      -0.076640286,
      0.016013449,
      -0.00077766576,
      -0.018147612,
      -0.04962999,
      -0.012260296,
      -0.019350484,
      0.010270535,
      0.012180092,
      0.003858917,
      0.05388526,
      0.061142597,
      0.049799904,
      -0.025325377,
      0.01596864,
      0.0037506404,
      -0.015387656,
      0.010250521
    ]
  ],
  "total_duration": 8237891641,
  "load_duration": 3850683276,
  "prompt_eval_count": 10
}
```


可以提取出`response_single.json()`中的`embeddings`字段，这个字段是一个列表，里面包含了若干个浮点数。而这些浮点数，就是`bge-m3`模型的嵌入结果，用来表示输入文本`"你好，我是居居，很高兴认识你"`这句话的语义。

```
# 提取嵌入结果
response_data_single["embeddings"]
```


```
[[-0.01922286,
  -0.041361455,
  -0.012615179,
  -0.038346406,
  -0.006061248,
  -0.056141857,
```

```
    0.024635639,
    -0.048374478,
    0.021608315,
    0.028389402,
    0.011415558,
    -0.008346135,
    -0.01575168,
    0.011952146,
    0.0151369395,
    -0.0076897917,
    0.019403733,
    -0.03635689,
    0.01607708,
    -0.030296508,
    -0.018613543,
    -0.0064844512,
    -0.00029064808,
    -0.010478618,
    0.029013928,
    0.01946554,
    -0.055293787,
    -0.0024098635,
    0.025135659,
    -0.04489269,
    0.027139282,
    0.015564225,
    0.014931622,
    -0.04502287,
    -0.023891093,
    -0.024381446,
    0.033852994,
    -0.0074271495,
    -0.035470333,
    -0.015253574,
    0.038780138,
    -0.010637472,
    0.050409704,
    -0.047934886,
    -0.0071231266,
    -0.02649895,
    -0.049684074,
    0.01412076,
    0.008318819,
    -0.0024914725,
    -0.014517145,
    -0.006007267,
    0.04962171,
    -0.0011974386,
    -0.045221522,
    0.026687985,
    0.036642727,
    -0.018013492,
    0.0062242304,
    -0.037533417,
    -0.026383227,
    0.0430352,
    0.023990763,
    -0.046562776,
```

```
    0.005052609,
    0.1078952,
    -0.0252689,
    -0.036592163,
    -0.017526496,
    -0.020836545,
    -0.014016081,
    0.0073044035,
    0.018462053,
    0.009727984,
    -0.05348882,
    0.03130105,
    0.054758947,
    0.019654304,
    -0.070866816,
    -0.009440621,
    0.052959364,
    0.028022723,
    0.020885868,
    -0.0027755883,
    0.025919601,
    0.0028719697,
    -0.011645562,
    0.07296864,
    -0.040803645,
    -0.014456895,
    -0.025773542,
    -0.024834368,
    0.0054630465,
    -0.025666654,
    -0.027248958,
    -0.009498956,
    -0.02270712,
    0.03019623,
    0.04891002,
    0.03496023,
    0.032207686,
    0.022681091,
    0.0092448965,
    -0.020107258,
    -0.0018319886,
    -0.007373062,
    0.03264567,
    0.03045322,
    -0.018678445,
    -0.012370282,
    0.023889143,
    0.005763869,
    0.005892724,
    -0.036694903,
    0.0010225775,
    -0.011602502,
    -0.019340092,
    0.014320763,
    0.0038706549,
    -0.019690355,
    0.0050612646,
    0.042274926,
```

-0.0075149853,
-0.012850127,
-0.0029281029,
-0.030729791,
0.020727703,
0.0390613,
-0.009327623,
0.008809925,
0.03356942,
-0.012686815,
-0.030047603,
-0.042933784,
-0.015244054,
-0.023267841,
-0.0105991755,
0.01019861,
-0.022745306,
-0.028969925,
0.0044860677,
0.025056915,
0.0063823955,
-0.051997595,
-0.0007780814,
-0.051995378,
0.017508458,
0.0270951,
0.018337702,
-0.018740973,
-0.01517844,
0.0043489956,
-0.007722501,
0.0066714576,
0.03081158,
-0.013100324,
0.009403119,
-0.0133148255,
-0.00554864,
0.039081782,
-0.009327039,
-0.01150965,
0.0057221474,
-0.022880614,
0.03247104,
0.03626947,
0.003067755,
0.013796212,
0.007955853,
-0.04198045,
0.010368125,
-0.0032647224,
0.044448704,
0.027452532,
-0.0057273414,
0.030706208,
0.07107367,
-0.005020788,
0.013032576,
-0.04025715,

-0.012651873,
0.029171083,
-0.008141952,
-0.016902551,
-0.04220599,
0.025815476,
-0.026521945,
-0.04866242,
-0.013637231,
0.03486512,
-0.001918829,
-0.057317823,
0.010962362,
-0.020244285,
0.03866528,
0.010617372,
-0.04377715,
0.009748239,
0.0045961896,
-0.002983128,
-0.004532409,
-0.019298308,
0.04492627,
0.017812751,
-0.053677455,
-0.0078686215,
-0.037468098,
0.011929472,
0.01720958,
-0.07143857,
-0.03330745,
0.024078554,
0.024672244,
-0.0569919,
-0.04148753,
0.050901547,
0.022151768,
0.031624585,
0.0021112408,
0.0004811191,
0.014531818,
-0.008901751,
0.034024063,
0.02308123,
-0.012181863,
-0.037322327,
-0.016135737,
0.05219511,
-0.00483262,
-0.024687467,
-0.007843138,
0.0142526515,
-0.0062834364,
-0.0070895934,
0.016342908,
-0.014341934,
-0.055698976,
0.07432804,

```
-0.025424266,
0.027217498,
-0.033968687,
-0.019814044,
-0.004727214,
0.0032408026,
-0.02595303,
0.004324861,
-0.0019504167,
0.014354229,
0.014305528,
0.036802884,
-0.008998354,
-0.06352971,
0.012633923,
-0.033731535,
0.028493801,
0.00502899,
0.014065273,
0.024009604,
-0.041306116,
0.00876087,
-0.0015096144,
-0.0581046,
0.009729838,
0.035152882,
0.0054527465,
-0.0033466823,
-0.007744983,
-0.036980435,
0.02000405,
0.042424027,
0.007387792,
0.021873796,
0.025885409,
0.003218425,
-0.0067987675,
0.0067939996,
-0.0058852965,
-0.03311272,
0.051925275,
-0.038485028,
0.017672395,
0.02621819,
0.042438682,
-0.0080195675,
0.052674558,
0.014233261,
-0.08290935,
-0.00034058988,
0.039673224,
-0.04147254,
-0.029645534,
-0.009901535,
0.033171214,
-0.03572185,
-0.012022275,
0.0054901815,
```

```
-0.0048021623,
-0.1500592,
0.0084257675,
-0.008138095,
-0.007993522,
-0.032590296,
-0.04603252,
-0.06815151,
-0.0007138535,
-0.019334827,
0.036172323,
0.013635413,
-0.013069142,
0.014766935,
0.0077957814,
-0.032953933,
0.023373697,
0.033019446,
-0.018915948,
0.018228997,
-0.0558767,
-0.046538107,
-0.021160431,
0.09153756,
-0.05778473,
0.028429775,
0.00013784457,
-0.00088175054,
-0.046915453,
-0.050807092,
-0.002632421,
0.0011797646,
0.0285695,
-0.00094530155,
0.015300199,
0.03612954,
0.027943866,
-0.014544252,
-0.016660748,
-0.032984182,
0.011127908,
-0.008035259,
0.021325957,
-0.0030667118,
0.0043126587,
0.005962908,
-0.00042244705,
-0.0065966123,
0.01868376,
0.03252872,
-0.021499867,
-0.0038907684,
0.0037884223,
0.03002122,
-0.042780697,
-0.025766337,
0.011636676,
-0.007580283,
```

```
    0.017507449,
    -0.004628474,
    0.038538925,
    -0.040182322,
    -0.027409792,
    0.043854233,
    0.048351236,
    -0.03091617,
    0.0027391338,
    0.00036530296,
    0.012099772,
    0.022873694,
    -0.012164523,
    0.028665284,
    0.01798609,
    0.031127907,
    0.013131891,
    -0.01598989,
    0.039874513,
    0.0009227839,
    -0.060541965,
    -0.024316201,
    -0.097156905,
    0.0062047048,
    -0.02919612,
    -0.017234696,
    0.032315448,
    -0.0012223435,
    0.03263769,
    -0.0027417927,
    0.036068577,
    0.062720075,
    0.22184578,
    0.019815253,
    0.01993091,
    -0.043113314,
    0.05674825,
    -0.0155262025,
    -0.020943452,
    0.029482389,
    -0.023183936,
    -0.004876635,
    0.0021091257,
    -0.02630077,
    -0.017170096,
    0.029073019,
    0.016285948,
    0.049894232,
    -0.0069553154,
    0.0022705528,
    0.03613966,
    -0.031493917,
    -0.0062867478,
    -0.029011935,
    -0.037524033,
    -0.021112205,
    -0.086708285,
    -0.040370315,
```

```
-0.014455863,
0.042079657,
-0.0327298,
0.009579455,
-0.013825583,
0.033582326,
-0.0028413108,
0.002097946,
0.026424421,
-0.002604664,
0.03141509,
-0.042284306,
-0.020693991,
0.052990846,
-0.01812236,
-0.025489392,
0.039961986,
-0.010502693,
-0.00163735,
0.024143638,
0.008718653,
0.0010326977,
-0.013266786,
-0.006186287,
0.04246847,
-0.008784686,
0.012570915,
-0.020447822,
-0.012151716,
-0.047826536,
-0.039175574,
-0.020817315,
0.025246646,
0.027339023,
0.025236443,
0.048806693,
-0.011991174,
-0.022016289,
-0.011961466,
0.0012723304,
0.008125226,
-0.035645425,
0.011581718,
0.0074978457,
-0.05309759,
-0.010962783,
-0.016122645,
-0.0047364915,
0.010034488,
0.0030421098,
0.010316423,
0.038907442,
0.005337137,
0.057604548,
-0.024876388,
0.013173044,
-0.062358152,
0.055846933,
```

0.019336209,
-0.0069100326,
0.0024203337,
0.04047966,
0.014952066,
-0.025101326,
0.025459653,
-0.007939787,
-0.05776767,
0.0052549858,
-0.012580737,
-0.06010769,
0.010566356,
-0.023706116,
-0.004249546,
0.018867647,
-0.028295763,
-0.0071428707,
0.00011242026,
0.07138021,
-0.0046134833,
-0.004373349,
0.024181986,
0.0366532,
-0.015330882,
0.076670185,
-0.044471033,
0.0236505,
-0.02199792,
0.047300745,
0.006907518,
-0.031923965,
-0.0044949916,
-0.034935955,
0.061860606,
-0.012920158,
-0.009882616,
-0.017172387,
-0.030343372,
0.027154673,
-0.012200059,
0.002181584,
0.05755524,
0.030812616,
-0.0052554836,
0.030308068,
0.011736867,
-0.011554788,
0.019779095,
0.040169403,
0.05081635,
0.014287744,
0.042208575,
0.03118895,
0.02718363,
-0.015795866,
-0.03593932,
0.029518906,

```
    -0.0368933,
    0.021886608,
    -0.02643343,
    -0.027600609,
    -0.024287364,
    -0.021399602,
    -9.439248e-05,
    -0.0064367475,
    0.039554097,
    0.04583417,
    0.007510894,
    -0.07607565,
    -0.009225462,
    -0.006582759,
    -0.0148243345,
    -0.01328466,
    0.030134458,
    -0.027474036,
    -0.003093169,
    -0.007971443,
    0.040055204,
    0.13162026,
    -0.02346389,
    0.014800054,
    -0.047165684,
    -0.017866734,
    0.012297593,
    -0.009757697,
    -0.014321117,
    -0.01788238,
    -0.039049163,
    0.03508187,
    -0.0022837762,
    -0.007907967,
    -0.00075374526,
    -0.018473191,
    0.008784031,
    0.013422078,
    0.011996075,
    0.015082026,
    -0.039035235,
    0.02770337,
    0.00021796707,
    0.008747956,
    0.011951045,
    -0.0045034382,
    0.02824308,
    -0.036295313,
    0.00022648639,
    0.082074136,
    -0.008459332,
    -0.026773563,
    0.021175345,
    0.005603807,
    0.011868295,
    -0.027673401,
    -0.030666795,
    -0.01869547,
```

```
    0.017391281,
    0.03777573,
    -0.021297984,
    -0.0445547,
    0.0042566187,
    -0.023678964,
    0.0066275527,
    0.043544076,
    -0.023232576,
    -0.018033905,
    -0.016615305,
    0.01890017,
    -0.016875628,
    0.023675755,
    -0.013777414,
    -0.025639104,
    0.012749661,
    0.008193755,
    -0.029998936,
    -0.016121306,
    -0.013959011,
    0.021408115,
    0.023548564,
    0.0012766648,
    0.0024227598,
    -0.00060618424,
    0.033602092,
    0.02401148,
    -0.021303771,
    0.05285596,
    -0.0058503565,
    0.012856821,
    0.00946074,
    -0.018707717,
    0.034906205,
    -0.0033552044,
    -0.042422313,
    0.027284978,
    -0.014717163,
    0.009164387,
    0.025540866,
    0.0018863472,
    0.037348382,
    0.026557129,
    0.027901705,
    0.05202844,
    0.020671928,
    0.007082176,
    0.0034235918,
    -0.031215776,
    -0.040995408,
    0.003343065,
    0.027570983,
    -0.015434854,
    -0.036407907,
    0.031241637,
    -0.034739457,
    -0.014406115,
```

-0.062684685,
0.010075276,
-0.031929497,
0.0025649823,
-0.021319512,
0.005636287,
0.026131155,
0.021893157,
0.027501272,
-0.067495435,
-0.016828451,
-0.017330559,
-0.027661966,
0.026027255,
-0.04029333,
-0.07741891,
0.04245963,
0.05786638,
-0.048196528,
0.08805092,
0.019176718,
0.043642513,
-0.016361246,
-0.008850222,
-0.03140456,
-0.0018215853,
-0.02850662,
-0.017091129,
-0.0021216967,
-0.0008238167,
0.014921121,
0.00032447535,
-0.044774458,
-0.04502461,
-0.0059552435,
0.061054993,
-0.04434121,
0.014774075,
0.00038209726,
0.0702063,
-0.021992086,
0.008455862,
-0.02410905,
0.0652755,
0.005741814,
0.0019593763,
0.0047087143,
-0.024068253,
0.012610589,
-0.06671961,
0.01753201,
-0.0034162577,
0.0023024643,
-0.054187473,
-0.011831421,
0.01715048,
0.04750781,
-0.03139423,

```
    0.02508704,
    -0.010289396,
    0.020973045,
    -0.014849384,
    -0.027357204,
    -0.01407778,
    0.033853255,
    -0.055280462,
    -0.013809495,
    0.024581378,
    -0.012241737,
    -0.0009671642,
    -0.018698305,
    0.028465739,
    0.0010522557,
    0.00916822,
    -0.04007455,
    -0.025335168,
    -0.011387842,
    0.0055432064,
    0.053396042,
    0.01740828,
    -0.014137389,
    0.02978183,
    0.018102026,
    0.05603387,
    0.008667747,
    -0.052618537,
    0.07024653,
    -0.052201025,
    0.029271316,
    0.027836822,
    0.037358917,
    -0.018538367,
    -0.0098952325,
    0.025678122,
    -0.022600275,
    -0.02866699,
    0.047425956,
    0.041606687,
    -0.045320027,
    0.009275773,
    -0.015545199,
    -0.012686884,
    -0.018841615,
    0.021218903,
    -0.028941818,
    0.027858598,
    -0.028179156,
    0.029352563,
    9.719478e-05,
    -0.008117893,
    0.023555716,
    0.0057699373,
    0.0073017688,
    0.038897607,
    0.0060669654,
    0.038502082,
```

```
-0.03068005,
0.050719038,
-0.013903436,
-0.043558348,
-0.01102934,
0.017825447,
-0.015128274,
-0.027200924,
-0.018548822,
-0.043312766,
0.046628807,
-0.0041413624,
-0.015847849,
4.6269724e-05,
-0.024507571,
0.059471335,
-0.017492197,
-0.015842618,
-0.033435818,
0.020426262,
-0.15537652,
0.022720942,
-0.022978572,
-0.0013574777,
-0.03353008,
-0.02227914,
-0.04341785,
-0.026962942,
0.036519103,
0.009219965,
-0.040424213,
-0.017088816,
0.01368187,
-0.03543528,
0.04532406,
0.011525667,
0.032134064,
-0.028047284,
0.006712295,
0.0066740294,
0.019953199,
-0.019357666,
0.088674895,
-0.01850979,
0.025244767,
-0.018174196,
-0.042392906,
0.023810916,
-0.042712238,
0.016077517,
-0.0056865416,
-0.007365048,
-0.0026553639,
0.05539162,
0.052702192,
0.012977329,
-0.013324897,
-0.046663057,
```

```
-0.004664647,
0.0038475136,
-0.0083794715,
0.008683904,
-0.028441006,
0.00988668,
0.019565033,
0.038489543,
0.021747977,
-0.020754416,
-0.042819865,
-0.03930985,
0.01374576,
-0.035338506,
0.030846687,
0.036640152,
-0.037350155,
0.0038905449,
-0.020976484,
0.0058610975,
-0.00070023915,
0.019661274,
-0.054429367,
-0.03100215,
-0.052285925,
-0.0120779425,
-0.074244425,
0.0006335962,
-0.0017926287,
0.0014389432,
0.022021336,
0.036196936,
-0.011613051,
-0.093202114,
0.0026990643,
-0.023806851,
-0.0054916134,
0.0011963513,
0.03746827,
-0.022631988,
-0.051280633,
-0.03702279,
0.04518093,
0.007951412,
-0.038437307,
-0.034927223,
0.03933246,
-0.015370903,
-0.02635214,
-0.0041535352,
-0.02957769,
0.012466689,
-0.043131832,
-0.020760499,
0.033076085,
0.061528783,
-0.040835578,
0.0050104596,
```

```
-0.006039951,
-0.054586615,
-0.046932537,
-0.016764142,
-0.004739649,
-0.01928498,
-0.031564888,
0.0028312965,
-0.013649793,
0.020705067,
0.019288259,
-0.00771178,
0.015912732,
-0.0185576,
-0.027995558,
0.021834064,
-0.007156878,
-0.012114014,
-0.04397438,
0.013293692,
0.031676818,
-0.0026619898,
-0.017556915,
0.06203819,
-0.015809024,
-0.0014322706,
-0.014704928,
-0.046601627,
0.010122574,
-0.01908295,
-0.014350051,
0.023351008,
-0.009993024,
-0.036077287,
-0.0052948133,
-0.06703926,
0.0069603934,
-0.011365105,
0.018846352,
-0.038462013,
-0.00936936,
0.0022846924,
-0.03624788,
-0.012948129,
-0.013476335,
-0.0182052,
0.0033161482,
0.0039851656,
0.0019273587,
-0.018899929,
-0.0206676,
0.022174593,
-0.040895447,
0.0038407452,
0.008989447,
0.060425438,
0.0058986675,
-0.013137046,
```

```
    0.0025831743,
    0.03484262,
    0.011654857,
    0.006968705,
    0.040688474,
    -0.001316991,
    -0.019467473,
    0.027343286,
    -0.022299161,
    -0.05269345,
    -0.024993394,
    -0.035053432,
    -0.019827059,
    0.030563287,
    0.008705064,
    -0.0045054746,
    0.0034345994,
    0.07223343,
    -0.0068709,
    -0.0078441575,
    0.05365767,
    -0.015098801,
    -0.0019810034,
    -0.008233295,
    0.08352503,
    0.018159185,
    0.011276068,
    0.030084705,
    0.014642007,
    0.007945344,
    0.0064324383,
    -0.021965735,
    0.022328451,
    0.0013204975,
    -0.0054936996,
    -0.0153985005,
    0.018929703,
    0.0011742727,
    -0.026166951,
    -0.006397983,
    0.02537267,
    0.039143644,
    0.0054356274,
    0.046194375,
    -0.0054917075,
    0.01610236,
    -0.06036488,
    -0.03880828,
    0.01609657,
    -0.0067152907,
    -0.012388587,
    0.020276126,
    -0.013779218,
    -0.0008071725,
    -0.013861764,
    0.024183014,
    -0.014108998,
    -0.0054775397,
```

```
  -0.014219473,
  0.023931691,
  0.036456257,
  0.029436398,
  0.02639856,
  -0.009838244,
  -0.017583802,
  0.0084198555,
  ...]]
```

具体用多少个浮点数来表示输入文本的语义，取决于具体的模型。比如我们使用的`bge-m3`模型，其返回的嵌入结果就是一个包含`1024`个浮点数的列表。

```
len(response_data_single["embeddings"][0])
```

```
1024
```

大家可以在`huggingface`或者`modelscope`等平台上，找到`bge-m3`模型的详细架构、评测信息及论文。我们这里不作为重点展开讲解。`bge-m3`模型在`modelscope`平台上的地址为：https://modelscope.cn/models/BAAI/bge-m3

Model Name	Dimension	Sequence Length	Introduction
BAAI/bge-m3	1024	8192	multilingual; unified fine-tuning (dense, sparse, and colbert) from bge-m3-unsupervised
BAAI/bge-m3-unsupervised	1024	8192	multilingual; contrastive learning from bge-m3-retromae
BAAI/bge-m3-retromae	--	8192	multilingual; extend the max_length of xlm-roberta to 8192 and further pretrained via retromae
BAAI/bge-large-en-v1.5	1024	512	English model
BAAI/bge-base-en-v1.5	768	512	English model
BAAI/bge-small-en-v1.5	384	512	English model

所以大家可以简单理解为，经过`bge-m3`模型处理后的结果，就是将输入的文本转换为一个包含`1024`个浮点数的列表，并且用这个列表来表示输入文本的语义。即 response_data_single["embeddings"][0] = "你好，我是居居，很高兴认识你"

同时，`bge-m3`模型还支持多个输入的请求，即可以一次性输入多个文本，然后返回每个文本的嵌入结果。代码如下：

```
import requests
import json

# 定义 API 端点
url = "http://192.168.110.131:11434/api/embed"     # 这里需要根据实际情况进行修改


# 多个输入的请求示例
multiple_input_payload = {
```

```
    "model": "bge-m3",
    "input": ["天为什么是蓝色的？",
              "草为什么是绿色的？"]
}

# 发送 POST 请求
response_multiple = requests.post(url, json=multiple_input_payload)

# 检查响应
if response_multiple.status_code == 200:
    print("\nMultiple Input Response:")
    response_data_multiple = response_multiple.json()
    print(json.dumps(response_data_multiple, indent=2))

    # 提取嵌入结果
    multiple_embeddings = response_data_multiple.get("embeddings", [])
else:
    print(f"Error: {response_multiple.status_code} - {response_multiple.text}")
```

```
Multiple Input Response:
{
  "model": "bge-m3",
  "embeddings": [
    [
      -0.042137332,
      0.005296808,
      -0.037715282,
      -0.06215416,
      -0.030467484,
      -0.007836851,
      -0.043893967,
      0.0022080292,
      0.016335009,
      -0.0117954165,
      -0.004529275,
      0.025638063,
      0.007796983,
      0.0063098017,
      0.018519185,
      -0.022322154,
      0.029461581,
      0.022574248,
      0.00038702495,
      0.019533236,
      -0.009881311,
      -0.0383717,
      -0.032535516,
      0.016021805,
      -0.006125824,
      0.0075718816,
      0.016597407,
      -0.017723264,
      -0.0002764759,
      0.001488903,
      0.004165976,
      0.0616205,
      -0.029179199,
```

```
        -0.022410607,
        0.0003814827,
        -0.017742423,
        -0.005474743,
        0.0049243644,
        -0.040230345,
        -0.032896325,
        0.009739351,
        -0.03913332,
        0.017520385,
        -0.00029088696,
        0.027668685,
        -0.046231378,
        -0.011650233,
        -0.03293465,
        -0.04470256,
        -0.0478838,
        0.008582121,
        0.0038168635,
        0.088327065,
        -0.016265068,
        0.011740456,
        -0.0029550614,
        0.028078863,
        -0.030943036,
        -0.043092407,
        -0.028814875,
        -0.014810867,
        -0.0020290688,
        -0.041463394,
        0.008612065,
        -0.013520107,
        0.08999495,
        0.02035601,
        -0.0044486164,
        -0.020074788,
        -0.004833883,
        -0.028073747,
        0.00032702446,
        0.0047799097,
        -0.0044234022,
        -0.046205215,
        -0.01817478,
        -0.0045246906,
        -0.0123508675,
        -0.040190786,
        0.04064541,
        0.09014224,
        0.014359349,
        -0.02341912,
        0.014465041,
        -0.0053078295,
        0.05538501,
        -0.034623098,
        0.041395552,
        -0.029607566,
        -0.037130527,
        -0.002067333,
```

```
        0.042320542,
        0.017242383,
        -0.056331713,
        -0.023820797,
        -0.0020812673,
        -0.034105197,
        0.0056840624,
        0.027562784,
        0.026438197,
        0.015122961,
        -0.029478177,
        -0.010766563,
        -0.00533037,
        -0.020935973,
        0.00032368215,
        0.051139206,
        0.009538334,
        -0.02400277,
        0.019776152,
        0.0584742,
        0.04907174,
        0.016894652,
        0.031076962,
        -0.03176524,
        0.009510206,
        -0.02218482,
        0.018387897,
        0.0031903163,
        0.010497863,
        -0.011635472,
        0.07276368,
        -0.0064129275,
        -0.045014057,
        -0.013379973,
        -0.028407896,
        0.016923483,
        0.04445915,
        -0.016673427,
        0.039915126,
        -0.025509529,
        0.019542592,
        -0.047511388,
        0.028730273,
        -0.054099925,
        0.0014411394,
        0.025573539,
        -0.010792686,
        0.031407714,
        -0.013306026,
        0.052538764,
        0.016519766,
        -0.027834127,
        -0.011570048,
        0.035803452,
        0.02007018,
        0.0024324271,
        -0.031974126,
        -0.0026232207,
```

-0.0049809525,
-0.0052737803,
0.023195265,
0.060987324,
-0.002575687,
-0.076468945,
-0.006767327,
-0.053074267,
0.009725925,
-0.05268764,
0.032771066,
0.038590457,
-0.0039342837,
-0.045914114,
0.02513468,
0.0021120834,
-0.03730243,
0.02977119,
0.06439825,
-0.008863884,
-0.020030886,
-0.02747345,
-0.0073932316,
-0.0048822267,
-0.0067345817,
-0.019864578,
0.015068145,
0.024992965,
-0.013167202,
0.027090143,
0.004015794,
-0.08040249,
-0.03474101,
0.0045948057,
-0.030831112,
-0.012045554,
0.02483818,
-0.015777161,
-0.009927092,
0.018526495,
0.0118054245,
0.0211727,
-0.037752703,
-0.001975264,
-0.0012090072,
0.027399225,
0.004032278,
0.022544393,
0.029868845,
0.00067225663,
-0.004111884,
-0.0016398347,
-0.0004749759,
0.041708183,
-0.017336763,
-0.0136280665,
0.019985555,
-0.09729808,

-0.08953787,
0.04107954,
-0.011029402,
-0.018815925,
0.011755434,
0.022931322,
0.014722142,
-0.021672897,
-0.023385474,
0.021395497,
0.024507977,
0.04235797,
0.015848517,
-0.02279181,
0.006232009,
-0.031116417,
-0.044243816,
0.048988286,
0.0065010227,
0.011547768,
0.0549056,
-0.04781384,
0.0029705032,
-0.026103448,
0.024926864,
0.012209596,
-0.003091049,
0.02546409,
-0.038644683,
-0.007874508,
-0.0073739677,
0.036731694,
0.024388144,
-0.044597827,
-0.021421095,
-0.0059583816,
0.05000934,
0.009413394,
0.040630322,
-0.03523095,
-0.004540993,
0.0022993265,
-0.023829076,
0.02254886,
-0.017369289,
-0.070587724,
0.0075032255,
0.031555068,
0.008978053,
0.009435259,
-0.017410524,
0.042360272,
0.03270071,
0.010293682,
-0.024027297,
-0.008167552,
0.03461271,
0.013336475,

-0.009405411,
-0.0094858855,
0.0004914884,
0.0026585604,
0.0040770057,
-0.024476796,
-0.0046170247,
-0.015485166,
-0.0057349876,
0.01156721,
-0.0007533831,
0.003609351,
-0.01711615,
-0.0011313154,
-0.019069092,
-0.0044528176,
-0.023705354,
0.015393332,
0.029787816,
-0.0032745432,
-0.038109522,
0.019249914,
0.0008609008,
-0.004534888,
0.031711407,
-0.020981707,
-0.0397812,
0.04040954,
0.020614095,
0.019030828,
0.02284263,
-0.035594705,
-0.16899337,
0.018541625,
0.014923138,
-0.007641145,
0.0075152004,
-0.024446165,
-0.0076794797,
-0.018340291,
0.00046605608,
-0.03280913,
-0.01234831,
-0.021243177,
-0.031833846,
-0.00918224,
-0.019617198,
-0.03971306,
0.01629078,
-0.035811763,
0.027423305,
-0.027777009,
-0.015663758,
-0.024587302,
0.05940994,
-0.002367637,
0.017544618,
0.013158937,

-0.027401974,
0.024363685,
-0.041683983,
-0.07853095,
0.04206934,
0.020653142,
-0.03020147,
0.032506283,
0.028300753,
-0.019385654,
0.030426295,
-0.048927456,
-0.0049383417,
0.02057371,
-0.007289443,
0.02098331,
-0.024779443,
0.030403,
-0.032178733,
0.0074722483,
-0.026696924,
-0.0057991054,
0.0046112672,
0.010001329,
0.0052110488,
0.003479432,
0.0062448625,
0.00091557216,
-0.030177973,
0.01767895,
-0.0018189346,
0.005169421,
0.022286616,
-0.0017267654,
-0.015250424,
-0.049802978,
0.025996951,
0.015015538,
0.011650537,
-0.015945734,
0.013876556,
0.031699087,
0.027250074,
0.0066137933,
0.019414512,
0.027010594,
0.0034491217,
-0.04149347,
0.03261539,
0.05666244,
-0.035440873,
-0.038834725,
-0.00057862076,
-0.094313756,
-0.018587569,
-0.013786597,
0.013335928,
-0.026358103,

```
        0.018386213,
        -0.034061395,
        -0.029979499,
        0.03101407,
        0.007732108,
        0.2302883,
        0.053795822,
        -0.04571347,
        -0.024319816,
        0.08307778,
        -0.038508948,
        0.0057074195,
        -0.005989547,
        -0.03456581,
        -0.028869793,
        -0.04034803,
        -0.016712103,
        -0.0066257915,
        -0.00882439,
        -0.0077595403,
        0.009169582,
        -0.028189782,
        -0.0033542505,
        0.056126468,
        -0.023846501,
        -0.0066834236,
        -0.0039994065,
        0.025167191,
        0.0070630596,
        0.0010949638,
        -0.022675293,
        0.01223915,
        0.07312742,
        0.00886766,
        -0.006981496,
        -0.028843962,
        -0.016596032,
        -0.014287267,
        0.011711063,
        -0.023769658,
        -0.007999069,
        0.018134821,
        0.013600745,
        0.06240502,
        -0.027478902,
        -0.039845858,
        -0.0061672656,
        0.046465103,
        0.033261463,
        0.08106834,
        0.015202136,
        0.03412636,
        -0.062328212,
        -0.02208452,
        -0.008780762,
        -0.02163691,
        -0.010332136,
        -0.053890415,
```

```
        0.030947028,
        -0.08494406,
        -0.029017588,
        -0.03273794,
        -0.021868378,
        0.005420783,
        0.014095832,
        0.03149366,
        0.016462496,
        0.0173673,
        -0.043623086,
        -0.0020759876,
        -0.03034031,
        0.04324677,
        -0.07422934,
        0.017031752,
        0.006030616,
        -0.014976466,
        -0.016536182,
        -0.014204306,
        0.012176292,
        0.035498124,
        0.048893087,
        0.033731975,
        0.05893236,
        -0.04438582,
        -0.019384483,
        0.042567324,
        -0.035785895,
        -0.027839916,
        0.010269249,
        -0.0029354466,
        0.034064908,
        0.042051334,
        -0.02377663,
        0.00018843584,
        -0.035804797,
        0.010596093,
        0.040386043,
        -0.07433216,
        -0.012972507,
        0.0020062684,
        0.00087044924,
        -0.025378624,
        -0.037642613,
        -0.015890175,
        -0.0076076183,
        -0.02487156,
        -0.029907221,
        -0.013884005,
        0.0052179443,
        -0.0051465426,
        -0.029555114,
        0.020426996,
        0.040726706,
        -0.027832404,
        0.042246837,
        -0.0033184467,
```

```
        0.001172159,
        -0.0557444,
        0.03280031,
        0.03765769,
        0.042062365,
        0.013983726,
        0.0004090046,
        0.0014503683,
        0.04151597,
        -0.027207311,
        -0.03811464,
        0.04106502,
        0.039898492,
        -0.018392889,
        -0.00559719,
        0.0068726214,
        0.0076489043,
        0.0006230939,
        0.033178918,
        0.01079643,
        -0.041384447,
        -0.02852786,
        0.016783055,
        0.055578187,
        0.0037498036,
        0.03197655,
        0.062441908,
        0.002617525,
        0.05285312,
        0.030939603,
        -0.038866654,
        -0.0042523113,
        -0.045382734,
        -0.009775364,
        -0.034698278,
        0.048598345,
        -0.06818018,
        -0.02444149,
        -0.00949042,
        0.04478301,
        -0.021221437,
        0.014549152,
        -0.03238616,
        0.012591191,
        0.0032901457,
        0.009852417,
        -0.014907928,
        -0.044620603,
        -0.011357798,
        -0.026407277,
        -0.005224016,
        0.039710406,
        0.10613283,
        0.008799815,
        0.019459253,
        -0.027430058,
        0.040994428,
        0.06255134,
```

-0.0154246325,
0.020769319,
0.010436879,
0.009245113,
0.039452195,
0.0220522,
-0.036322378,
-0.022910323,
-0.026929844,
-0.06490222,
0.010154379,
0.036104105,
0.01758559,
0.016539378,
0.045427833,
0.013154021,
0.009437805,
-0.018297847,
-0.037246346,
0.023390587,
-0.020376557,
0.01932413,
0.095888935,
0.00019674169,
-0.007847683,
0.07149473,
0.032855626,
-0.012178423,
0.017537348,
-0.0023458574,
-0.019856326,
0.06914988,
-0.053961936,
0.0052872654,
0.012509775,
0.02971595,
-0.02288805,
0.027495164,
-0.000319835,
-0.012909792,
-0.0010698951,
-0.03811992,
-0.02886807,
-0.03195666,
-0.00096493005,
0.008205409,
-0.011321164,
-0.03252674,
0.0045426735,
-0.00061344256,
-0.043989014,
0.0035914248,
-0.01197999,
0.045499917,
-0.043762688,
-0.0055888724,
0.06170304,
-0.030959107,

-0.010835247,
0.009822792,
-0.02960603,
0.0030709354,
0.01912414,
-0.03826324,
-0.009199489,
0.016952321,
0.019091725,
0.027671369,
0.035184704,
0.005566593,
-0.004224325,
-0.023847291,
0.020012423,
0.028360244,
0.059198152,
-0.00022528647,
0.005460027,
0.011523086,
0.016808476,
0.027759623,
0.02160105,
-0.004696154,
0.0938206,
0.02182346,
-0.012023529,
-0.02564297,
-0.0050284294,
0.018947987,
0.027122408,
-0.047256127,
-0.0041446886,
-0.0007468413,
0.010792308,
0.012999448,
0.0009446954,
0.0041484386,
-0.009155726,
0.038582344,
0.017999995,
-0.01361257,
-0.029035058,
-0.015319298,
0.018348051,
-0.037154764,
-0.021188546,
-0.030136174,
0.01371045,
-0.0252252,
0.036648184,
-0.03341299,
-0.015977893,
-0.009471925,
0.07168737,
-0.06356442,
-0.00022004705,
-0.024585567,

```
        0.026660563,
        0.014813339,
        -0.011186062,
        0.02635014,
        -0.01858569,
        -0.011048534,
        -0.0037236298,
        -0.013637644,
        0.019399958,
        -0.020726122,
        -0.005873677,
        -0.05457326,
        -0.0010754383,
        -0.032321937,
        0.024502382,
        -0.017577903,
        0.05186453,
        -0.041495655,
        -0.034449518,
        -0.022555443,
        -0.003808817,
        -0.0064183106,
        -0.029723026,
        -0.010237768,
        0.024252448,
        0.0057710805,
        -0.02287218,
        0.046242468,
        -0.033879757,
        -0.016698122,
        -0.027835608,
        -0.031813916,
        -0.023639053,
        -0.06083357,
        0.0026263874,
        0.023786323,
        -0.0026732201,
        0.014249788,
        0.015631843,
        0.0615537,
        -0.008399952,
        -0.073089555,
        -0.013478393,
        -0.06011397,
        0.020935731,
        -0.038242858,
        0.010878561,
        -0.008781969,
        -0.013526375,
        0.023306286,
        -0.0076850546,
        -0.004033106,
        -0.01432605,
        0.024405079,
        -0.022795958,
        0.020023946,
        0.062027246,
        0.041369814,
```

-0.034915254,
0.05650035,
-0.033667177,
0.0017321882,
0.028220907,
0.02724711,
-0.011040877,
0.009466217,
0.06299186,
-0.032366026,
-0.031843692,
0.013810198,
0.036252715,
-0.012403185,
-0.03576739,
-0.03492998,
0.036709167,
-0.02419729,
0.065110564,
-0.019976975,
0.010940757,
-0.010370927,
0.0048032184,
0.00061093795,
0.042379238,
-0.02673291,
-0.0049099107,
0.0005162485,
0.083228916,
0.004646965,
-0.03581079,
-0.033165492,
0.016541002,
-0.04952429,
-0.021216413,
0.019397082,
0.009300314,
-0.008058479,
-0.002323441,
-0.008978433,
-0.010337525,
-0.005220179,
-0.031576138,
0.008024545,
-0.016106777,
0.024416858,
-0.012258581,
-0.027702363,
0.018079732,
-0.058961343,
-0.022970779,
-0.14390658,
0.0037366857,
-0.0032212595,
-0.01208507,
0.0030887993,
-0.009112169,
-0.023670638,

-0.004746355,
-0.003891055,
-0.04284588,
0.01887782,
0.04493371,
-0.04388891,
-0.023415668,
0.025585275,
0.03896533,
0.013118991,
0.009227519,
0.003197749,
0.050215654,
0.0037740078,
-0.003807509,
-0.020907104,
0.0070870104,
0.02154187,
-0.034253534,
0.0051916186,
0.013102805,
-0.05013027,
-0.011375772,
-0.034398254,
-0.012215769,
0.008936165,
0.013042285,
-0.0076083266,
0.020103235,
0.0015896345,
0.0013469604,
0.018086607,
0.0064340974,
-0.009976818,
0.0202195,
-0.04537388,
-0.0037930352,
0.00905707,
0.056803208,
-0.05901939,
-0.0025249731,
-0.026721176,
0.029771479,
-0.009664647,
0.03187292,
0.048367433,
0.044916652,
-0.030181149,
0.022398885,
0.017789925,
0.0104143135,
0.04470509,
0.043254152,
0.0026467436,
-0.015192392,
0.00457335,
-0.023504077,
-0.013324363,

```
        0.018283904,
        -0.09116972,
        0.023821378,
        -0.013131093,
        0.0167326,
        -0.032134093,
        0.024058864,
        0.01773737,
        0.0029694063,
        0.015400375,
        0.021107076,
        0.040144313,
        -0.018088255,
        -0.018179972,
        -0.0061101853,
        0.0333514,
        -0.051985517,
        0.0018389969,
        -0.042181797,
        0.010265013,
        0.030729504,
        -0.0021242746,
        -0.030638387,
        -0.017514488,
        -0.043879565,
        -0.07925115,
        -0.009313998,
        -0.05037993,
        0.0062779468,
        -0.021197595,
        0.009778766,
        -0.026689053,
        0.008420625,
        -0.013590769,
        -0.025906648,
        -0.019427065,
        -0.0040084007,
        -0.0051668566,
        -0.03791699,
        -0.030223792,
        0.021360977,
        0.05214987,
        -0.001195285,
        0.056878682,
        0.021255476,
        0.030073736,
        -0.009004809,
        -0.00044008184,
        0.018010443,
        0.05210606,
        -0.009380206,
        0.014023647,
        -0.027397914,
        0.0024216578,
        0.028537884,
        0.0041281483,
        0.004196239,
        -0.042434014,
```

-0.01951087,
0.010246227,
0.019259237,
0.044823032,
-0.015578959,
0.055282183,
-0.029959694,
-0.002444019,
-0.030419223,
-0.0015842384,
0.019480113,
0.016508475,
0.002307729,
0.027967995,
-0.00045019708,
0.027565105,
-0.027752383,
0.015513533,
0.0042188945,
-0.020022955,
-0.020090383,
-0.02833768,
-0.008205656,
0.018416148,
0.013030916,
0.0043110526,
0.019754246,
0.02399727,
-0.04217088,
-0.029771129,
-0.01836493,
0.030383382,
-0.011573766,
0.0028558928,
-0.031085022,
0.013471276,
0.015004199,
0.018147502,
0.009186545,
-0.028823266,
-0.01669298,
-0.07682935,
0.0026605711,
-0.011341015,
-0.052770227,
0.008115369,
0.044612564,
0.0425546,
0.032530054,
-0.056424312,
-0.009478192,
0.04076847,
-0.027977053,
0.026075717,
0.03094319,
0.0077947774,
-0.010361255,
0.0379331,

```
        0.022605725,
        0.04959574,
        0.01899966,
        0.007897847,
        0.023821533,
        -0.026519801,
        -0.0055518188,
        0.028449029,
        0.012021892,
        0.0055719055,
        0.0065661324,
        -0.021502847,
        -0.039742764,
        0.045740023,
        -0.006839873,
        0.0048939977,
        0.00086617976,
        -0.021977283,
        0.006932612,
        -0.011130043,
        -0.018758686,
        -0.013963125,
        -0.008874906,
        -0.00250807,
        -0.03496783,
        0.0045763687,
        -0.0135702705,
        -0.04224915,
        0.057138246,
        -0.0035801234,
        -0.039638594,
        -0.0042460803,
        0.003417834,
        -0.038236164,
        -0.014725476,
        0.010474428,
        -0.007192687,
        0.012550738,
        0.049059138,
        -0.02168995,
        0.014943881,
        -0.02944271,
        0.015512776,
        -0.006800395,
        0.016161487,
        -0.05883302,
        0.028153684,
        -0.008107896,
        -0.009384435,
        -0.013420283,
        0.05274957,
        -0.023136541,
        0.013694571,
        -0.018405624,
        -0.012455675,
        -0.007547505,
        0.02333343,
        0.013213835,
```

```
      -0.044884287,
      0.024978612,
      0.007462058,
      0.017579155,
      -0.04433104
    ],
    [
      -0.039806955,
      0.018106725,
      -0.02714153,
      -0.033833414,
      -0.010921911,
      -0.08787385,
      -0.048164885,
      0.031297795,
      0.032188654,
      -0.004530212,
      -0.021804372,
      0.024625015,
      -0.007042427,
      0.009888381,
      0.009500467,
      -0.025422748,
      0.011871752,
      -0.00732404,
      0.018392378,
      -0.004384966,
      0.0001918134,
      -0.04004225,
      0.010167631,
      0.00806233,
      0.03110558,
      0.04625598,
      -0.012658598,
      -0.028965505,
      -0.021548064,
      0.056681987,
      -0.018512,
      0.0221357,
      -0.03510766,
      -0.0047702184,
      -0.016245274,
      -0.015167115,
      0.023144336,
      -0.0524422,
      -0.031156112,
      -0.027200451,
      0.009988742,
      -0.020565098,
      0.014143164,
      -0.015100003,
      0.033231463,
      -0.044389643,
      -0.004136181,
      -0.025562724,
      -0.058502603,
      -0.014722696,
      -0.011400933,
```

```
        0.014331452,
        0.07658601,
        -0.014814307,
        0.0031236135,
        -0.0033029092,
        0.019509245,
        0.017938983,
        -0.024790436,
        -0.015248284,
        -0.011639638,
        0.00501442,
        -0.0055467803,
        0.008408418,
        -0.024580697,
        0.0754493,
        0.019086119,
        -0.022098487,
        -0.008814615,
        -0.021437159,
        -0.030012729,
        0.0044206856,
        -0.036046855,
        -0.012681524,
        -0.024640366,
        -0.019347185,
        -0.0052223173,
        0.015137555,
        -0.048428133,
        0.046845492,
        0.065149285,
        0.002647668,
        0.0007207915,
        0.02435114,
        0.044494085,
        0.0182996,
        -0.029606061,
        0.037432224,
        -0.033808608,
        0.0022643944,
        0.009645186,
        -0.022846015,
        0.024837328,
        -0.047011603,
        -0.035402983,
        -0.0046847723,
        -0.039559603,
        0.0074818777,
        0.02330356,
        0.025716277,
        -0.014304545,
        0.023733906,
        -0.006246653,
        0.008509405,
        -0.011293758,
        -0.025966475,
        -0.01006334,
        0.008181945,
        -0.01438242,
```

```
        0.009540608,
        0.06885038,
        0.04916865,
        0.0018908386,
        0.003854289,
        -0.03604726,
        -0.0146035105,
        -0.022960719,
        0.019865872,
        -0.0058088247,
        0.010324922,
        -0.012516908,
        0.015785128,
        0.051367365,
        -0.025630152,
        -0.017501123,
        -0.01962036,
        -0.00045100553,
        0.044624988,
        -0.021584759,
        0.03565252,
        -0.026582485,
        0.027201204,
        -0.059125,
        0.01317713,
        -0.012250775,
        -0.025474267,
        0.027487954,
        0.016126018,
        0.022486532,
        -0.024704559,
        0.037192255,
        0.017422775,
        -0.019903207,
        -0.036879085,
        0.011453935,
        -0.00235618,
        0.013886989,
        -0.023742465,
        0.014823697,
        -0.0013640152,
        0.007340349,
        0.025479008,
        0.029555285,
        -0.017932586,
        -0.03817522,
        0.0007982303,
        -0.012220362,
        0.00021680194,
        -0.036326665,
        -0.0062908116,
        0.032608576,
        0.012039895,
        -0.020109933,
        0.002184031,
        0.014053472,
        -0.03662654,
        0.0044956626,
```

```
        0.07524842,
        -0.0152201755,
        0.01951614,
        -0.015556351,
        -0.013448956,
        0.048647966,
        0.0122912,
        -0.016437223,
        0.021207852,
        0.030076804,
        -0.016976235,
        0.0035818152,
        -0.0107744625,
        -0.061285947,
        -0.021923166,
        -0.017293802,
        -0.08759088,
        -0.039984487,
        -0.00043348063,
        0.010560561,
        -0.021260453,
        0.010690529,
        -0.00076214195,
        -0.0052662776,
        -0.07078946,
        -0.020219805,
        -0.0030916699,
        0.0038979545,
        -0.0032112007,
        0.003736981,
        0.06447358,
        0.054365616,
        0.020362832,
        -0.011558257,
        0.027246028,
        0.030825194,
        -0.03900765,
        -0.038409732,
        -0.018651864,
        -0.036674645,
        -0.04837994,
        0.041603886,
        -0.03700395,
        -0.003150507,
        -0.05434538,
        0.013208883,
        -0.0087698875,
        -0.010625307,
        -0.018957065,
        0.029379977,
        -0.019123154,
        0.0066628377,
        0.00064306625,
        -0.0045157797,
        -0.012306675,
        -0.01965497,
        -0.03788315,
        0.021389931,
```

-0.0076195886,
-0.02418394,
0.07190492,
-0.058766495,
-0.04346301,
-0.011424005,
0.08432645,
0.010501825,
-0.023052415,
-0.010896898,
-0.041609973,
0.034540582,
-0.025880303,
-0.0044130436,
0.022009451,
-0.0456012,
-0.026753843,
-0.00026467428,
0.03633925,
0.021008782,
0.021594657,
-0.033761226,
-0.003918439,
-0.0036088256,
-0.030125177,
0.06014599,
-0.0052329483,
-0.01676937,
0.027992217,
0.0020218696,
0.01681878,
0.031245762,
0.0074888594,
0.039434277,
0.0072179767,
0.023201365,
-0.01086617,
-0.011059847,
0.036365334,
0.012356821,
-0.004479378,
-0.029226903,
0.028809851,
0.004290806,
0.016059155,
-0.042729758,
-0.015265739,
-0.023315625,
-0.0048190304,
0.036037758,
-0.0037047488,
0.0058000875,
0.004225492,
0.030909065,
-0.020124217,
0.012732193,
-0.049268555,
-0.010288543,

```
        0.032568026,
        0.004937674,
        -0.019989511,
        -0.00376297,
        -0.021907778,
        0.0032743274,
        0.021680184,
        0.0031568797,
        -0.030931003,
        0.036583208,
        0.008307239,
        -0.035271116,
        0.021804618,
        -0.03550608,
        -0.17422535,
        -0.0055584437,
        0.016568558,
        -0.042162247,
        0.009234726,
        -0.021887384,
        0.0044160085,
        -0.032132085,
        -0.02526897,
        -0.03257029,
        -0.061538894,
        -0.032033436,
        -0.05803612,
        0.019335447,
        -0.048609257,
        -0.0123426365,
        0.0047235885,
        -0.053109802,
        0.0016916203,
        -0.037515607,
        -0.003102214,
        -0.0014287904,
        0.07933123,
        -0.03516418,
        -0.02318397,
        -0.016031312,
        -0.017235648,
        -0.018351972,
        -0.04145283,
        -0.031296615,
        0.027566928,
        -0.04918952,
        -0.02337211,
        0.023168243,
        -0.03689156,
        -0.007357466,
        0.015730577,
        -0.02637404,
        -0.0011485332,
        0.013727407,
        -0.032218985,
        0.008697318,
        -0.053799678,
        0.020576116,
```

-0.011187619,
-0.0036567596,
-0.009576436,
-0.009067609,
-0.0059511457,
0.047911406,
0.0032387092,
0.032790948,
0.0044828868,
0.045609824,
-0.044275556,
0.032871027,
0.019262772,
0.02776002,
0.025117772,
0.007967346,
-0.011527599,
-0.035411373,
0.02457575,
-0.025654942,
-0.0071522677,
0.029986482,
0.023516832,
0.030535799,
0.045038912,
0.011724758,
0.0022815238,
0.019235903,
-0.0058248453,
-0.007677952,
0.044396043,
0.044792693,
-0.027284242,
-0.0034480433,
-0.0031397224,
-0.09265861,
0.02985109,
-0.03535556,
0.008404779,
-0.03251267,
0.00832229,
-0.01800912,
-0.015365692,
-0.006667201,
0.005070796,
0.2429714,
0.034296818,
-0.061447185,
0.011751609,
0.03563779,
-0.045496084,
0.033172783,
0.016725417,
-0.013325212,
-0.04710844,
-0.01763304,
0.014165556,
0.020341875,

-0.022885386,
0.01367139,
0.062560104,
-0.011401764,
-0.01932027,
0.06305335,
-0.010825633,
-0.0067342655,
0.0017462067,
0.021622764,
0.040241074,
-0.060613107,
-0.04294679,
-0.0018433991,
0.031839903,
0.007688521,
0.014064593,
-0.021199869,
-0.04950631,
0.03862481,
0.059081998,
0.019803561,
-0.018866148,
0.00047172944,
0.03317995,
0.041107073,
-0.03278018,
-0.05122741,
-0.010277512,
0.021450896,
0.0075447876,
0.0232653,
-0.016950065,
0.021917164,
-0.014481437,
-0.0060826093,
-0.005029667,
-0.01805828,
0.03047291,
-0.04469266,
0.023791978,
-0.029562002,
-0.03152938,
-0.015890347,
-0.035565127,
-0.037068065,
0.05428713,
0.053404607,
0.00011869666,
0.015649194,
-0.026654938,
-0.040659923,
0.04302314,
0.025084848,
-0.056900717,
0.01611657,
0.061198022,
-0.010592068,

```
        -0.0117429765,
        -0.04247498,
        0.010234422,
        0.046964835,
        0.049254086,
        0.024806071,
        0.043843374,
        -0.022250427,
        -0.0021817428,
        0.0040312316,
        -0.021270124,
        -0.04342302,
        0.009678571,
        0.041352488,
        0.055536613,
        0.022192832,
        -0.007861454,
        0.025131604,
        -0.025723273,
        0.0131876785,
        0.026801726,
        -0.06978861,
        0.00529091,
        -0.026728887,
        -0.004692603,
        -0.029864728,
        -0.020519625,
        -0.0012926307,
        -0.031000795,
        0.0042536897,
        -0.046934154,
        -0.008083166,
        0.044085227,
        0.03775721,
        -0.029693477,
        -0.011002456,
        0.05681608,
        -0.04142204,
        -0.01024627,
        0.011809906,
        -0.03645116,
        -0.01609644,
        0.020747473,
        0.05280843,
        0.021996047,
        0.027929071,
        -0.013967343,
        0.013427497,
        0.024700826,
        -0.025466755,
        -0.03191814,
        0.018574893,
        0.015539701,
        0.0111877015,
        0.0011065699,
        0.020807378,
        -0.0022950503,
        -0.04080207,
```

```
        0.022003127,
        0.009040768,
        -0.029722787,
        -0.016087228,
        0.0021534252,
        0.05132611,
        -0.0043423655,
        0.03197384,
        0.014612478,
        -0.0075006653,
        0.03795542,
        0.020516044,
        -0.000753212,
        0.0140319625,
        -0.036455624,
        -0.0062314887,
        -0.020856056,
        0.028727965,
        -0.040752828,
        -0.03098454,
        -0.03632487,
        0.057478394,
        0.02651742,
        -0.015345521,
        -0.03778333,
        -0.01893185,
        0.043988228,
        0.008374845,
        0.03578321,
        -0.0336425,
        -0.0014386997,
        -0.01712994,
        -0.0057694726,
        0.06346051,
        0.079468384,
        -0.0037914685,
        0.044671357,
        -0.021962574,
        0.048231363,
        0.026963266,
        0.011023811,
        0.046371337,
        -0.012278799,
        -0.013266753,
        0.024954408,
        0.03089167,
        -0.02914379,
        -0.0147609105,
        0.034018278,
        -0.03385961,
        0.02879893,
        0.046497975,
        0.019660087,
        0.02409784,
        0.020308346,
        -0.0067614387,
        0.046343904,
        -0.0005784594,
```

```
        -0.021660347,
        0.054527197,
        -0.022563111,
        0.028653856,
        0.080075465,
        -0.010942759,
        0.008333316,
        0.07266343,
        -0.0063063973,
        -0.015116326,
        -0.001291006,
        0.007033947,
        -0.055138774,
        -0.013635839,
        -0.058863103,
        0.018013697,
        -0.003587027,
        0.03016848,
        0.004340696,
        -0.02231035,
        -0.0021904027,
        -0.018317793,
        -0.034449887,
        -0.044973347,
        -0.042687934,
        -0.012501527,
        0.008154034,
        0.008486107,
        -0.027704634,
        -0.016208507,
        -0.017212354,
        0.0154763,
        -0.057456914,
        0.058836646,
        -0.010927607,
        0.028967118,
        -0.023743823,
        0.018120123,
        0.008504358,
        -0.02019317,
        0.027913263,
        -0.033391576,
        -0.05849971,
        0.0018396962,
        0.0028468056,
        0.012038477,
        0.019820383,
        0.033899166,
        -0.0034281653,
        0.010240003,
        0.007385701,
        -0.022350889,
        -0.030605549,
        0.017475165,
        -0.010327821,
        0.027654938,
        0.05355119,
        -0.014949391,
```

```
        0.0056421943,
        -0.006985141,
        0.06490143,
        -0.0019786155,
        0.06540899,
        0.00030945623,
        0.06651776,
        0.031434633,
        0.029223382,
        0.005312344,
        0.02692162,
        0.0080550695,
        0.002547737,
        -0.046926796,
        -0.0055953916,
        -0.027216688,
        0.0192472,
        0.019051598,
        0.027252058,
        -0.002625347,
        0.011415791,
        0.011703875,
        -0.042730447,
        -0.017118221,
        -0.04256073,
        -0.01781555,
        0.025160633,
        -0.009056251,
        -0.0028699443,
        6.267121e-05,
        0.0029575485,
        -0.03786012,
        0.032860763,
        -0.03924916,
        0.018752553,
        -0.0035671783,
        0.037145987,
        -0.07339691,
        -0.026845668,
        -0.024349527,
        -0.010001643,
        -0.008324767,
        0.02444552,
        -0.0069263037,
        -0.039972678,
        -0.011900961,
        0.0012417778,
        0.017389448,
        0.020070942,
        0.02678553,
        0.015119207,
        -0.052110363,
        0.021610133,
        -0.015621386,
        0.017013198,
        -0.016733378,
        0.050118178,
        -0.036132984,
```

-0.028919403,
-0.029918974,
0.000518692,
0.013747376,
-0.024373489,
-0.0054474077,
0.019377526,
0.024695145,
-0.004618123,
0.041052736,
0.026499253,
0.04656181,
-0.029189281,
0.00806193,
-0.011962786,
-0.048480418,
0.0014013203,
0.003932788,
0.016817138,
0.011343118,
-0.022587297,
-0.0015305758,
0.000102333805,
0.02695947,
0.011637096,
-0.049498096,
0.018794134,
-0.0237115,
0.029978462,
-0.015226518,
-0.012754388,
0.050404325,
-0.035898246,
-0.03161909,
-0.0103445435,
0.049694147,
-0.006435172,
-0.02847327,
0.08106319,
0.0809235,
-0.009108299,
0.028060187,
-0.038263198,
0.0008455384,
0.030499715,
-0.006457368,
0.004814054,
0.016831778,
0.010108291,
-0.03908352,
-0.043085005,
0.018404072,
0.0045124697,
0.0065787556,
-0.05353358,
-0.01287885,
0.010193927,
-0.027740126,

```
        0.06827365,
        -0.034389623,
        0.028805405,
        -0.0023430856,
        0.009260239,
        0.005530282,
        -0.025005303,
        -0.043685574,
        0.012968424,
        0.012462829,
        0.04922187,
        0.011033892,
        -0.020876283,
        -0.015727568,
        -0.006340249,
        -0.04240463,
        -0.0387011,
        0.031318422,
        -0.010669236,
        -0.038257945,
        0.025644235,
        -0.036942434,
        -0.040654384,
        0.026557662,
        -0.038425166,
        0.010972449,
        -0.0026865592,
        0.04297902,
        0.013841087,
        -0.03508849,
        0.022146288,
        -0.059955824,
        -0.011984481,
        -0.14203429,
        0.033375036,
        0.025635537,
        -0.014800292,
        0.021314079,
        -0.004364666,
        -0.04288085,
        0.03326454,
        -0.017278133,
        -0.015381165,
        0.039260007,
        0.030663054,
        -0.009398082,
        0.0016157854,
        0.0058923406,
        0.038551237,
        0.029295184,
        0.026982203,
        0.0046961685,
        0.020004513,
        -0.0002823051,
        0.029769791,
        -0.06499312,
        0.0027809944,
        0.033780456,
```

-0.054227453,
0.015405423,
0.014887773,
-0.06223483,
0.019264648,
-0.02459092,
-0.028076837,
-0.0007747023,
0.0050467625,
0.0054090386,
0.044855643,
0.014965291,
0.0039615836,
0.0364642,
-0.0151543515,
-0.010991144,
0.014066373,
-0.020417051,
-0.015295324,
0.0032331492,
0.029506348,
-0.05176175,
-0.012887155,
-0.049193565,
-0.025174212,
-0.0057166666,
-0.00039162458,
0.00026217767,
0.07197446,
-0.008801158,
0.023327848,
-0.013825299,
0.019054411,
0.026037995,
0.029275186,
0.0014251247,
0.0050785993,
0.014754311,
-0.02803208,
-0.020658525,
0.009343213,
-0.038446713,
-0.0021920993,
0.030721895,
0.0060682576,
-0.0029885068,
0.0290252,
-0.006677525,
-0.007283009,
0.009740268,
0.021990651,
-0.024688762,
-0.04346818,
-0.029574063,
-0.007917782,
0.05176251,
-0.03194417,
-0.010803628,

```
        -0.03710345,
        0.01654293,
        0.035606004,
        0.00021680047,
        -0.01902328,
        0.017944654,
        -0.014634677,
        -0.08091542,
        -0.0005132855,
        -0.0646727,
        -0.044102196,
        -0.05027943,
        0.023779303,
        0.0013730207,
        -0.00033895578,
        -0.021345852,
        -0.01240465,
        -0.025230281,
        0.02300386,
        0.0059149317,
        -0.006499429,
        -0.04890679,
        0.049107686,
        0.041578628,
        -0.021378888,
        0.023876995,
        -0.012183695,
        0.013455828,
        -0.005322552,
        -0.015810125,
        0.025522828,
        -0.010038473,
        -0.006050725,
        0.019967675,
        -0.003245995,
        -0.025224932,
        0.0439206,
        -0.0012436587,
        0.031723216,
        -0.02448605,
        -0.015330694,
        0.026584158,
        -0.0051061427,
        0.066581406,
        -0.0038770763,
        0.020848919,
        -0.028565202,
        -0.046525814,
        -0.058668178,
        0.034280267,
        -0.0088918265,
        -0.026706906,
        0.018157527,
        0.0033295872,
        0.005130095,
        0.019549852,
        -0.008804092,
        0.0069478713,
```

```
        0.009930967,
        -0.032567203,
        -0.042641874,
        -0.025361327,
        -0.02444165,
        0.060229223,
        -0.00029296556,
        -0.032582972,
        0.008229873,
        0.00873001,
        -0.00815309,
        -0.016171323,
        -0.029477486,
        0.048059177,
        1.3053809e-06,
        0.0039567174,
        -0.040332206,
        0.00093539315,
        0.055955756,
        0.008039591,
        -0.098208874,
        -0.024455974,
        0.012274973,
        -0.026007501,
        -0.0077049667,
        0.0053886627,
        -0.019571463,
        -0.010535064,
        0.03599966,
        0.034453247,
        0.042971645,
        -0.047814555,
        -0.006743227,
        0.033782616,
        -0.03835373,
        0.02963017,
        0.057553414,
        0.012134137,
        -0.01209421,
        0.008088897,
        0.029915666,
        0.024700345,
        0.03985199,
        0.018612375,
        -0.007387687,
        -0.028683772,
        0.015096653,
        0.020524723,
        -0.0025782587,
        0.00621819,
        -0.018208023,
        -0.044911884,
        -0.006225681,
        0.011868896,
        -0.00073360204,
        -0.00060059247,
        0.025633158,
        0.010927056,
```

```
      0.028156856,
      0.013699513,
      -0.01723093,
      -0.007418714,
      0.001602008,
      -0.006724474,
      -0.042085167,
      -0.044802777,
      -0.0005950661,
      -0.024677,
      0.0152367465,
      -0.035699733,
      -0.021288736,
      -0.015362072,
      0.04180836,
      -0.033493835,
      -0.00041006325,
      -0.030007878,
      -0.0003501069,
      0.025993193,
      0.030151015,
      -0.014684397,
      0.036694344,
      -0.049206484,
      0.05902542,
      -0.0063680233,
      0.027811198,
      -0.054977603,
      0.04245663,
      -0.044224508,
      0.0074903113,
      0.019687254,
      -0.045596678,
      -0.01462281,
      -0.0051411195,
      0.026682809,
      -0.0054900898,
      0.016229948,
      0.025660323,
      0.031181056,
      -0.052291203,
      0.03282995,
      -0.014796475,
      -0.0059239697,
      0.012493835
    ]
  ],
  "total_duration": 269259423,
  "load_duration": 5524644,
  "prompt_eval_count": 14
}
```

```
print("Total embeddings:", len(multiple_embeddings))
print("First embedding:", len(multiple_embeddings[0]), multiple_embeddings[0])  # 表示天为什么是蓝色的？
print("Second embedding:", len(multiple_embeddings[1]), multiple_embeddings[1])  # 表示草为什么是绿色的？
```

```
Total embeddings: 2
```

```
First embedding: 1024 [-0.042137332, 0.005296808, -0.037715282, -0.06215416,
-0.030467484, -0.007836851, -0.043893967, 0.0022080292, 0.016335009, -0.0117954165,
-0.004529275, 0.025638063, 0.007796983, 0.0063098017, 0.018519185, -0.022322154,
0.029461581, 0.022574248, 0.00038702495, 0.019533236, -0.009881311, -0.0383717,
-0.032535516, 0.016021805, -0.006125824, 0.0075718816, 0.016597407, -0.017723264,
-0.0002764759, 0.001488903, 0.004165976, 0.0616205, -0.029179199, -0.022410607,
0.0003814827, -0.017742423, -0.005474743, 0.0049243644, -0.040230345, -0.032896325,
0.009739351, -0.03913332, 0.017520385, -0.00029088696, 0.027668685, -0.046231378,
-0.011650233, -0.03293465, -0.04470256, -0.0478838, 0.008582121, 0.0038168635,
0.088327065, -0.016265068, 0.011740456, -0.0029550614, 0.028078863, -0.030943036,
-0.043092407, -0.028814875, -0.014810867, -0.0020290688, -0.041463394, 0.008612065,
-0.013520107, 0.08999495, 0.02035601, -0.0044486164, -0.020074788, -0.004833883,
-0.028073747, 0.00032702446, 0.0047799097, -0.0044234022, -0.046205215, -0.01817478,
-0.0045246906, -0.0123508675, -0.040190786, 0.04064541, 0.09014224, 0.014359349,
-0.02341912, 0.014465041, -0.0053078295, 0.05538501, -0.034623098, 0.041395552,
-0.029607566, -0.037130527, -0.002067333, 0.042320542, 0.017242383, -0.056331713,
-0.023820797, -0.0020812673, -0.034105197, 0.0056840624, 0.027562784, 0.026438197,
0.015122961, -0.029478177, -0.010766563, -0.00533037, -0.020935973, 0.00032368215,
0.051139206, 0.009538334, -0.02400277, 0.019776152, 0.0584742, 0.04907174, 0.016894652,
0.031076962, -0.03176524, 0.009510206, -0.02218482, 0.018387897, 0.0031903163,
0.010497863, -0.011635472, 0.07276368, -0.0064129275, -0.045014057, -0.013379973,
-0.028407896, 0.016923483, 0.04445915, -0.016673427, 0.039915126, -0.025509529,
0.019542592, -0.047511388, 0.028730273, -0.054099925, 0.0014411394, 0.025573539,
-0.010792686, 0.031407714, -0.013306026, 0.052538764, 0.016519766, -0.027834127,
-0.011570048, 0.035803452, 0.02007018, 0.0024324271, -0.031974126, -0.0026232207,
-0.0049809525, -0.0052737803, 0.023195265, 0.060987324, -0.002575687, -0.076468945,
-0.006767327, -0.053074267, 0.009725925, -0.05268764, 0.032771066, 0.038590457,
-0.0039342837, -0.045914114, 0.02513468, 0.0021120834, -0.03730243, 0.02977119,
0.06439825, -0.008863884, -0.020030886, -0.02747345, -0.0073932316, -0.0048822267,
-0.0067345817, -0.019864578, 0.015068145, 0.024992965, -0.013167202, 0.027090143,
0.004015794, -0.08040249, -0.03474101, 0.0045948057, -0.030831112, -0.012045554,
0.02483818, -0.015777161, -0.009927092, 0.018526495, 0.0118054245, 0.0211727,
-0.037752703, -0.001975264, -0.0012090072, 0.027399225, 0.004032278, 0.022544393,
0.029868845, 0.00067225663, -0.004111884, -0.0016398347, -0.0004749759, 0.041708183,
-0.017336763, -0.0136280665, 0.019985555, -0.09729808, -0.08953787, 0.04107954,
-0.011029402, -0.018815925, 0.011755434, 0.022931322, 0.014722142, -0.021672897,
-0.023385474, 0.021395497, 0.024507977, 0.04235797, 0.015848517, -0.02279181,
0.006232009, -0.031116417, -0.044243816, 0.048988286, 0.0065010227, 0.011547768,
0.0549056, -0.04781384, 0.0029705032, -0.026103448, 0.024926864, 0.012209596,
-0.003091049, 0.02546409, -0.038644683, -0.007874508, -0.0073739677, 0.036731694,
0.024388144, -0.044597827, -0.021421095, -0.0059583816, 0.05000934, 0.009413394,
0.040630322, -0.03523095, -0.004540993, 0.0022993265, -0.023829076, 0.02254886,
-0.017369289, -0.070587724, 0.0075032255, 0.031555068, 0.008978053, 0.009435259,
-0.017410524, 0.042360272, 0.03270071, 0.010293682, -0.024027297, -0.008167552,
0.03461271, 0.013336475, -0.009405411, -0.0094858855, 0.0004914884, 0.0026585604,
0.0040770057, -0.024476796, -0.0046170247, -0.015485166, -0.0057349876, 0.01156721,
-0.0007533831, 0.003609351, -0.01711615, -0.0011313154, -0.019069092, -0.0044528176,
-0.023705354, 0.015393332, 0.029787816, -0.0032745432, -0.038109522, 0.019249914,
0.0008609008, -0.004534888, 0.031711407, -0.020981707, -0.0397812, 0.04040954,
0.020614095, 0.019030828, 0.02284263, -0.035594705, -0.16899337, 0.018541625,
0.014923138, -0.007641145, 0.0075152004, -0.024446165, -0.0076794797, -0.018340291,
0.00046605608, -0.03280913, -0.01234831, -0.021243177, -0.031833846, -0.00918224,
-0.019617198, -0.03971306, 0.01629078, -0.035811763, 0.027423305, -0.027777009,
-0.015663758, -0.024587302, 0.05940994, -0.002367637, 0.017544618, 0.013158937,
-0.027401974, 0.024363685, -0.041683983, -0.07853095, 0.04206934, 0.020653142,
-0.03020147, 0.032506283, 0.028300753, -0.019385654, 0.030426295, -0.048927456,
-0.0049383417, 0.02057371, -0.007289443, 0.02098331, -0.024779443, 0.030403,
-0.032178733, 0.0074722483, -0.026696924, -0.0057991054, 0.0046112672, 0.010001329,
```

0.0052110488, 0.003479432, 0.0062448625, 0.00091557216, -0.030177973, 0.01767895,
-0.0018189346, 0.005169421, 0.022286616, -0.0017267654, -0.015250424, -0.049802978,
0.025996951, 0.015015538, 0.011650537, -0.015945734, 0.013876556, 0.031699087,
0.027250074, 0.0066137933, 0.019414512, 0.027010594, 0.0034491217, -0.04149347,
0.03261539, 0.05666244, -0.035440873, -0.038834725, -0.00057862076, -0.094313756,
-0.018587569, -0.013786597, 0.013335928, -0.026358103, 0.018386213, -0.034061395,
-0.029979499, 0.03101407, 0.007732108, 0.2302883, 0.053795822, -0.04571347, -0.024319816,
0.08307778, -0.038508948, 0.0057074195, -0.005989547, -0.03456581, -0.028869793,
-0.04034803, -0.016712103, -0.0066257915, -0.00882439, -0.0077595403, 0.009169582,
-0.028189782, -0.0033542505, 0.056126468, -0.023846501, -0.0066834236, -0.0039994065,
0.025167191, 0.0070630596, 0.0010949638, -0.022675293, 0.01223915, 0.07312742,
0.00886766, -0.006981496, -0.028843962, -0.016596032, -0.014287267, 0.011711063,
-0.023769658, -0.007999069, 0.018134821, 0.013600745, 0.06240502, -0.027478902,
-0.039845858, -0.0061672656, 0.046465103, 0.033261463, 0.08106834, 0.015202136,
0.03412636, -0.062328212, -0.02208452, -0.008780762, -0.02163691, -0.010332136,
-0.053890415, 0.030947028, -0.08494406, -0.029017588, -0.03273794, -0.021868378,
0.005420783, 0.014095832, 0.03149366, 0.016462496, 0.0173673, -0.043623086,
-0.0020759876, -0.03034031, 0.04324677, -0.07422934, 0.017031752, 0.006030616,
-0.014976466, -0.016536182, -0.014204306, 0.012176292, 0.035498124, 0.048893087,
0.033731975, 0.05893236, -0.04438582, -0.019384483, 0.042567324, -0.035785895,
-0.027839916, 0.010269249, -0.0029354466, 0.034064908, 0.042051334, -0.02377663,
0.00018843584, -0.035804797, 0.010596093, 0.040386043, -0.07433216, -0.012972507,
0.0020062684, 0.00087044924, -0.025378624, -0.037642613, -0.015890175, -0.0076076183,
-0.02487156, -0.029907221, -0.013884005, 0.0052179443, -0.0051465426, -0.029555114,
0.020426996, 0.040726706, -0.027832404, 0.042246837, -0.0033184467, 0.001172159,
-0.0557444, 0.03280031, 0.03765769, 0.042062365, 0.013983726, 0.0004090046, 0.0014503683,
0.04151597, -0.027207311, -0.03811464, 0.04106502, 0.039898492, -0.018392889,
-0.00559719, 0.0068726214, 0.0076489043, 0.0006230939, 0.033178918, 0.01079643,
-0.041384447, -0.02852786, 0.016783055, 0.055578187, 0.0037498036, 0.03197655,
0.062441908, 0.002617525, 0.05285312, 0.030939603, -0.038866654, -0.0042523113,
-0.045382734, -0.009775364, -0.034698278, 0.048598345, -0.06818018, -0.02444149,
-0.00949042, 0.04478301, -0.021221437, 0.014549152, -0.03238616, 0.012591191,
0.0032901457, 0.009852417, -0.014907928, -0.044620603, -0.011357798, -0.026407277,
-0.005224016, 0.039710406, 0.10613283, 0.008799815, 0.019459253, -0.027430058,
0.040994428, 0.06255134, -0.0154246325, 0.020769319, 0.010436879, 0.009245113,
0.039452195, 0.0220522, -0.036322378, -0.022910323, -0.026929844, -0.06490222,
0.010154379, 0.036104105, 0.01758559, 0.016539378, 0.045427833, 0.013154021, 0.009437805,
-0.018297847, -0.037246346, 0.023390587, -0.020376557, 0.01932413, 0.095888935,
0.00019674169, -0.007847683, 0.07149473, 0.032855626, -0.012178423, 0.017537348,
-0.0023458574, -0.019856326, 0.06914988, -0.053961936, 0.0052872654, 0.012509775,
0.02971595, -0.02288805, 0.027495164, -0.000319835, -0.012909792, -0.0010698951,
-0.03811992, -0.02886807, -0.03195666, -0.00096493005, 0.008205409, -0.011321164,
-0.03252674, 0.0045426735, -0.00061344256, -0.043989014, 0.0035914248, -0.01197999,
0.045499917, -0.043762688, -0.0055888724, 0.06170304, -0.030959107, -0.010835247,
0.009822792, -0.02960603, 0.0030709354, 0.01912414, -0.03826324, -0.009199489,
0.016952321, 0.019091725, 0.027671369, 0.035184704, 0.005566593, -0.004224325,
-0.023847291, 0.020012423, 0.028360244, 0.059198152, -0.00022528647, 0.005460027,
0.011523086, 0.016808476, 0.027759623, 0.02160105, -0.004696154, 0.0938206, 0.02182346,
-0.012023529, -0.02564297, -0.0050284294, 0.018947987, 0.027122408, -0.047256127,
-0.0041446886, -0.0007468413, 0.010792308, 0.012999448, 0.0009446954, 0.0041484386,
-0.009155726, 0.038582344, 0.017999995, -0.01361257, -0.029035058, -0.015319298,
0.018348051, -0.037154764, -0.021188546, -0.030136174, 0.01371045, -0.0252252,
0.036648184, -0.03341299, -0.015977893, -0.009471925, 0.07168737, -0.06356442,
-0.00022004705, -0.024585567, 0.026660563, 0.014813339, -0.011186062, 0.02635014,
-0.01858569, -0.011048534, -0.0037236298, -0.013637644, 0.019399958, -0.020726122,
-0.005873677, -0.05457326, -0.0010754383, -0.032321937, 0.024502382, -0.017577903,
0.05186453, -0.041495655, -0.034449518, -0.022555443, -0.003808817, -0.0064183106,
-0.029723026, -0.010237768, 0.024252448, 0.0057710805, -0.02287218, 0.046242468,

-0.033879757, -0.016698122, -0.027835608, -0.031813916, -0.023639053, -0.06083357,
0.0026263874, 0.023786323, -0.0026732201, 0.014249788, 0.015631843, 0.0615537,
-0.008399952, -0.073089555, -0.013478393, -0.06011397, 0.020935731, -0.038242858,
0.010878561, -0.008781969, -0.013526375, 0.023306286, -0.0076850546, -0.004033106,
-0.01432605, 0.024405079, -0.022795958, 0.020023946, 0.062027246, 0.041369814,
-0.034915254, 0.05650035, -0.033667177, 0.0017321882, 0.028220907, 0.02724711,
-0.011040877, 0.009466217, 0.06299186, -0.032366026, -0.031843692, 0.013810198,
0.036252715, -0.012403185, -0.03576739, -0.03492998, 0.036709167, -0.02419729,
0.065110564, -0.019976975, 0.010940757, -0.010370927, 0.0048032184, 0.00061093795,
0.042379238, -0.02673291, -0.0049099107, 0.0005162485, 0.083228916, 0.004646965,
-0.03581079, -0.033165492, 0.016541002, -0.04952429, -0.021216413, 0.019397082,
0.009300314, -0.008058479, -0.002323441, -0.008978433, -0.010337525, -0.005220179,
-0.031576138, 0.008024545, -0.016106777, 0.024416858, -0.012258581, -0.027702363,
0.018079732, -0.058961343, -0.022970779, -0.14390658, 0.0037366857, -0.0032212595,
-0.01208507, 0.0030887993, -0.009112169, -0.023670638, -0.004746355, -0.003891055,
-0.04284588, 0.01887782, 0.04493371, -0.04388891, -0.023415668, 0.025585275, 0.03896533,
0.013118991, 0.009227519, 0.003197749, 0.050215654, 0.0037740078, -0.003807509,
-0.020907104, 0.0070870104, 0.02154187, -0.034253534, 0.0051916186, 0.013102805,
-0.05013027, -0.011375772, -0.034398254, -0.012215769, 0.008936165, 0.013042285,
-0.0076083266, 0.020103235, 0.0015896345, 0.0013469604, 0.018086607, 0.0064340974,
-0.009976818, 0.0202195, -0.04537388, -0.0037930352, 0.00905707, 0.056803208,
-0.05901939, -0.0025249731, -0.026721176, 0.029771479, -0.009664647, 0.03187292,
0.048367433, 0.044916652, -0.030181149, 0.022398885, 0.017789925, 0.0104143135,
0.04470509, 0.043254152, 0.0026467436, -0.015192392, 0.00457335, -0.023504077,
-0.013324363, 0.018283904, -0.09116972, 0.023821378, -0.013131093, 0.0167326,
-0.032134093, 0.024058864, 0.01773737, 0.0029694063, 0.015400375, 0.021107076,
0.040144313, -0.018088255, -0.018179972, -0.0061101853, 0.0333514, -0.051985517,
0.0018389969, -0.042181797, 0.010265013, 0.030729504, -0.0021242746, -0.030638387,
-0.017514488, -0.043879565, -0.07925115, -0.009313998, -0.05037993, 0.0062779468,
-0.021197595, 0.009778766, -0.026689053, 0.008420625, -0.013590769, -0.025906648,
-0.019427065, -0.0040084007, -0.0051668566, -0.03791699, -0.030223792, 0.021360977,
0.05214987, -0.001195285, 0.056878682, 0.021255476, 0.030073736, -0.009004809,
-0.00044008184, 0.018010443, 0.05210606, -0.009380206, 0.014023647, -0.027397914,
0.0024216578, 0.028537884, 0.0041281483, 0.004196239, -0.042434014, -0.01951087,
0.010246227, 0.019259237, 0.044823032, -0.015578959, 0.055282183, -0.029959694,
-0.002444019, -0.030419223, -0.0015842384, 0.019480113, 0.016508475, 0.002307729,
0.027967995, -0.00045019708, 0.027565105, -0.027752383, 0.015513533, 0.0042188945,
-0.020022955, -0.020090383, -0.02833768, -0.008205656, 0.018416148, 0.013030916,
0.0043110526, 0.019754246, 0.02399727, -0.04217088, -0.029771129, -0.01836493,
0.030383382, -0.011573766, 0.0028558928, -0.031085022, 0.013471276, 0.015004199,
0.018147502, 0.009186545, -0.028823266, -0.01669298, -0.07682935, 0.0026605711,
-0.011341015, -0.052770227, 0.008115369, 0.044612564, 0.0425546, 0.032530054,
-0.056424312, -0.009478192, 0.04076847, -0.027977053, 0.026075717, 0.03094319,
0.0077947774, -0.010361255, 0.0379331, 0.022605725, 0.04959574, 0.01899966, 0.007897847,
0.023821533, -0.026519801, -0.0055518188, 0.028449029, 0.012021892, 0.0055719055,
0.0065661324, -0.021502847, -0.039742764, 0.045740023, -0.006839873, 0.0048939977,
0.00086617976, -0.021977283, 0.006932612, -0.011130043, -0.018758686, -0.013963125,
-0.008874906, -0.00250807, -0.03496783, 0.0045763687, -0.0135702705, -0.04224915,
0.057138246, -0.0035801234, -0.039638594, -0.0042460803, 0.003417834, -0.038236164,
-0.014725476, 0.010474428, -0.007192687, 0.012550738, 0.049059138, -0.02168995,
0.014943881, -0.02944271, 0.015512776, -0.006800395, 0.016161487, -0.05883302,
0.028153684, -0.008107896, -0.009384435, -0.013420283, 0.05274957, -0.023136541,
0.013694571, -0.018405624, -0.012455675, -0.007547505, 0.02333343, 0.013213835,
-0.044884287, 0.024978612, 0.007462058, 0.017579155, -0.04433104]

```
Second embedding: 1024 [-0.039806955, 0.018106725, -0.02714153, -0.033833414,
-0.010921911, -0.08787385, -0.048164885, 0.031297795, 0.032188654, -0.004530212,
-0.021804372, 0.024625015, -0.007042427, 0.009888381, 0.009500467, -0.025422748,
0.011871752, -0.00732404, 0.018392378, -0.004384966, 0.0001918134, -0.04004225,
0.010167631, 0.00806233, 0.03110558, 0.04625598, -0.012658598, -0.028965505,
-0.021548064, 0.056681987, -0.018512, 0.0221357, -0.03510766, -0.0047702184,
-0.016245274, -0.015167115, 0.023144336, -0.0524422, -0.031156112, -0.027200451,
0.009988742, -0.020565098, 0.014143164, -0.015100003, 0.033231463, -0.044389643,
-0.004136181, -0.025562724, -0.058502603, -0.014722696, -0.011400933, 0.014331452,
0.07658601, -0.014814307, 0.0031236135, -0.0033029092, 0.019509245, 0.017938983,
-0.024790436, -0.015248284, -0.011639638, 0.00501442, -0.0055467803, 0.008408418,
-0.024580697, 0.0754493, 0.019086119, -0.022098487, -0.008814615, -0.021437159,
-0.030012729, 0.0044206856, -0.036046855, -0.012681524, -0.024640366, -0.019347185,
-0.0052223173, 0.015137555, -0.048428133, 0.046845492, 0.065149285, 0.002647668,
0.0007207915, 0.02435114, 0.044494085, 0.0182996, -0.029606061, 0.037432224,
-0.033808608, 0.0022643944, 0.009645186, -0.022846015, 0.024837328, -0.047011603,
-0.035402983, -0.0046847723, -0.039559603, 0.0074818777, 0.02330356, 0.025716277,
-0.014304545, 0.023733906, -0.006246653, 0.008509405, -0.011293758, -0.025966475,
-0.01006334, 0.008181945, -0.01438242, 0.009540608, 0.06885038, 0.04916865, 0.0018908386,
0.003854289, -0.03604726, -0.0146035105, -0.022960719, 0.019865872, -0.0058088247,
0.010324922, -0.012516908, 0.015785128, 0.051367365, -0.025630152, -0.017501123,
-0.01962036, -0.00045100553, 0.044624988, -0.021584759, 0.03565252, -0.026582485,
0.027201204, -0.059125, 0.01317713, -0.012250775, -0.025474267, 0.027487954, 0.016126018,
0.022486532, -0.024704559, 0.037192255, 0.017422775, -0.019903207, -0.036879085,
0.011453935, -0.00235618, 0.013886989, -0.023742465, 0.014823697, -0.0013640152,
0.007340349, 0.025479008, 0.029555285, -0.017932586, -0.03817522, 0.0007982303,
-0.012220362, 0.00021680194, -0.036326665, -0.0062908116, 0.032608576, 0.012039895,
-0.020109933, 0.002184031, 0.014053472, -0.03662654, 0.0044956626, 0.07524842,
-0.0152201755, 0.01951614, -0.015556351, -0.013448956, 0.048647966, 0.0122912,
-0.016437223, 0.021207852, 0.030076804, -0.016976235, 0.0035818152, -0.0107744625,
-0.061285947, -0.021923166, -0.017293802, -0.08759088, -0.039984487, -0.00043348063,
0.010560561, -0.021260453, 0.010690529, -0.00076214195, -0.0052662776, -0.07078946,
-0.020219805, -0.0030916699, 0.0038979545, -0.0032112007, 0.003736981, 0.06447358,
0.054365616, 0.020362832, -0.011558257, 0.027246028, 0.030825194, -0.03900765,
-0.038409732, -0.018651864, -0.036674645, -0.04837994, 0.041603886, -0.03700395,
-0.003150507, -0.05434538, 0.013208883, -0.0087698875, -0.010625307, -0.018957065,
0.029379977, -0.019123154, 0.0066628377, 0.00064306625, -0.0045157797, -0.012306675,
-0.01965497, -0.03788315, 0.021389931, -0.0076195886, -0.02418394, 0.07190492,
-0.058766495, -0.04346301, -0.011424005, 0.08432645, 0.010501825, -0.023052415,
-0.010896898, -0.041609973, 0.034540582, -0.025880303, -0.0044130436, 0.022009451,
-0.0456012, -0.026753843, -0.00026467428, 0.03633925, 0.021008782, 0.021594657,
-0.033761226, -0.003918439, -0.0036088256, -0.030125177, 0.06014599, -0.0052329483,
-0.01676937, 0.027992217, 0.0020218696, 0.01681878, 0.031245762, 0.0074888594,
0.039434277, 0.0072179767, 0.023201365, -0.01086617, -0.011059847, 0.036365334,
0.012356821, -0.004479378, -0.029226903, 0.028809851, 0.004290806, 0.016059155,
-0.042729758, -0.015265739, -0.023315625, -0.0048190304, 0.036037758, -0.0037047488,
0.0058000875, 0.004225492, 0.030909065, -0.020124217, 0.012732193, -0.049268555,
-0.010288543, 0.032568026, 0.004937674, -0.019989511, -0.00376297, -0.021907778,
0.0032743274, 0.021680184, 0.0031568797, -0.030931003, 0.036583208, 0.008307239,
-0.035271116, 0.021804618, -0.03550608, -0.17422535, -0.0055584437, 0.016568558,
-0.042162247, 0.009234726, -0.021887384, 0.0044160085, -0.032132085, -0.02526897,
-0.03257029, -0.061538894, -0.032033436, -0.05803612, 0.019335447, -0.048609257,
-0.0123426365, 0.0047235885, -0.053109802, 0.0016916203, -0.037515607, -0.003102214,
-0.0014287904, 0.07933123, -0.03516418, -0.02318397, -0.016031312, -0.017235648,
-0.018351972, -0.04145283, -0.031296615, 0.027566928, -0.04918952, -0.02337211,
0.023168243, -0.03689156, -0.007357466, 0.015730577, -0.02637404, -0.0011485332,
0.013727407, -0.032218985, 0.008697318, -0.053799678, 0.020576116, -0.011187619,
-0.0036567596, -0.009576436, -0.009067609, -0.0059511457, 0.047911406, 0.0032387092,
```

0.032790948, 0.0044828868, 0.045609824, -0.044275556, 0.032871027, 0.019262772,
0.02776002, 0.025117772, 0.007967346, -0.011527599, -0.035411373, 0.02457575,
-0.025654942, -0.0071522677, 0.029986482, 0.023516832, 0.030535799, 0.045038912,
0.011724758, 0.0022815238, 0.019235903, -0.0058248453, -0.007677952, 0.044396043,
0.044792693, -0.027284242, -0.0034480433, -0.0031397224, -0.09265861, 0.02985109,
-0.03535556, 0.008404779, -0.03251267, 0.00832229, -0.01800912, -0.015365692,
-0.006667201, 0.005070796, 0.2429714, 0.034296818, -0.061447185, 0.011751609, 0.03563779,
-0.045496084, 0.033172783, 0.016725417, -0.013325212, -0.04710844, -0.01763304,
0.014165556, 0.020341875, -0.022885386, 0.01367139, 0.062560104, -0.011401764,
-0.01932027, 0.06305335, -0.010825633, -0.0067342655, 0.0017462067, 0.021622764,
0.040241074, -0.060613107, -0.04294679, -0.0018433991, 0.031839903, 0.007688521,
0.014064593, -0.021199869, -0.04950631, 0.03862481, 0.059081998, 0.019803561,
-0.018866148, 0.00047172944, 0.03317995, 0.041107073, -0.03278018, -0.05122741,
-0.010277512, 0.021450896, 0.0075447876, 0.0232653, -0.016950065, 0.021917164,
-0.014481437, -0.0060826093, -0.005029667, -0.01805828, 0.03047291, -0.04469266,
0.023791978, -0.029562002, -0.03152938, -0.015890347, -0.035565127, -0.037068065,
0.05428713, 0.053404607, 0.00011869666, 0.015649194, -0.026654938, -0.040659923,
0.04302314, 0.025084848, -0.056900717, 0.01611657, 0.061198022, -0.010592068,
-0.0117429765, -0.04247498, 0.010234422, 0.046964835, 0.049254086, 0.024806071,
0.043843374, -0.022250427, -0.0021817428, 0.0040312316, -0.021270124, -0.04342302,
0.009678571, 0.041352488, 0.055536613, 0.022192832, -0.007861454, 0.025131604,
-0.025723273, 0.0131876785, 0.026801726, -0.06978861, 0.00529091, -0.026728887,
-0.004692603, -0.029864728, -0.020519625, -0.0012926307, -0.031000795, 0.0042536897,
-0.046934154, -0.008083166, 0.044085227, 0.03775721, -0.029693477, -0.011002456,
0.05681608, -0.04142204, -0.01024627, 0.011809906, -0.03645116, -0.01609644, 0.020747473,
0.05280843, 0.021996047, 0.027929071, -0.013967343, 0.013427497, 0.024700826,
-0.025466755, -0.03191814, 0.018574893, 0.015539701, 0.0111877015, 0.0011065699,
0.020807378, -0.0022950503, -0.04080207, 0.022003127, 0.009040768, -0.029722787,
-0.016087228, 0.0021534252, 0.05132611, -0.0043423655, 0.03197384, 0.014612478,
-0.0075006653, 0.03795542, 0.020516044, -0.000753212, 0.0140319625, -0.036455624,
-0.0062314887, -0.020856056, 0.028727965, -0.040752828, -0.03098454, -0.03632487,
0.057478394, 0.02651742, -0.015345521, -0.03778333, -0.01893185, 0.043988228,
0.008374845, 0.03578321, -0.0336425, -0.0014386997, -0.01712994, -0.0057694726,
0.06346051, 0.079468384, -0.0037914685, 0.044671357, -0.021962574, 0.048231363,
0.026963266, 0.011023811, 0.046371337, -0.012278799, -0.013266753, 0.024954408,
0.03089167, -0.02914379, -0.0147609105, 0.034018278, -0.03385961, 0.02879893,
0.046497975, 0.019660087, 0.02409784, 0.020308346, -0.0067614387, 0.046343904,
-0.0005784594, -0.021660347, 0.054527197, -0.022563111, 0.028653856, 0.080075465,
-0.010942759, 0.008333316, 0.07266343, -0.0063063973, -0.015116326, -0.001291006,
0.007033947, -0.055138774, -0.013635839, -0.058863103, 0.018013697, -0.003587027,
0.03016848, 0.004340696, -0.02231035, -0.0021904027, -0.018317793, -0.034449887,
-0.044973347, -0.042687934, -0.012501527, 0.008154034, 0.008486107, -0.027704634,
-0.016208507, -0.017212354, 0.0154763, -0.057456914, 0.058836646, -0.010927607,
0.028967118, -0.023743823, 0.018120123, 0.008504358, -0.02019317, 0.027913263,
-0.033391576, -0.05849971, 0.0018396962, 0.0028468056, 0.012038477, 0.019820383,
0.033899166, -0.0034281653, 0.010240003, 0.007385701, -0.022350889, -0.030605549,
0.017475165, -0.010327821, 0.027654938, 0.05355119, -0.014949391, 0.0056421943,
-0.006985141, 0.06490143, -0.0019786155, 0.06540899, 0.00030945623, 0.06651776,
0.031434633, 0.029223382, 0.005312344, 0.02692162, 0.0080550695, 0.002547737,
-0.046926796, -0.0055953916, -0.027216688, 0.0192472, 0.019051598, 0.027252058,
-0.002625347, 0.011415791, 0.011703875, -0.042730447, -0.017118221, -0.04256073,
-0.01781555, 0.025160633, -0.009056251, -0.0028699443, 6.267121e-05, 0.0029575485,
-0.03786012, 0.032860763, -0.03924916, 0.018752553, -0.0035671783, 0.037145987,
-0.07339691, -0.026845668, -0.024349527, -0.010001643, -0.008324767, 0.02444552,
-0.0069263037, -0.039972678, -0.011900961, 0.0012417778, 0.017389448, 0.020070942,
0.02678553, 0.015119207, -0.052110363, 0.021610133, -0.015621386, 0.017013198,
-0.016733378, 0.050118178, -0.036132984, -0.028919403, -0.029918974, 0.000518692,
0.013747376, -0.024373489, -0.0054474077, 0.019377526, 0.024695145, -0.004618123,

0.041052736, 0.026499253, 0.04656181, -0.029189281, 0.00806193, -0.011962786,
-0.048480418, 0.0014013203, 0.003932788, 0.016817138, 0.011343118, -0.022587297,
-0.0015305758, 0.000102333805, 0.02695947, 0.011637096, -0.049498096, 0.018794134,
-0.0237115, 0.029978462, -0.015226518, -0.012754388, 0.050404325, -0.035898246,
-0.03161909, -0.0103445435, 0.049694147, -0.006435172, -0.02847327, 0.08106319,
0.0809235, -0.009108299, 0.028060187, -0.038263198, 0.0008455384, 0.030499715,
-0.006457368, 0.004814054, 0.016831778, 0.010108291, -0.03908352, -0.043085005,
0.018404072, 0.0045124697, 0.0065787556, -0.05353358, -0.01287885, 0.010193927,
-0.027740126, 0.06827365, -0.034389623, 0.028805405, -0.0023430856, 0.009260239,
0.005530282, -0.025005303, -0.043685574, 0.012968424, 0.012462829, 0.04922187,
0.011033892, -0.020876283, -0.015727568, -0.006340249, -0.04240463, -0.0387011,
0.031318422, -0.010669236, -0.038257945, 0.025644235, -0.036942434, -0.040654384,
0.026557662, -0.038425166, 0.010972449, -0.0026865592, 0.04297902, 0.013841087,
-0.03508849, 0.022146288, -0.059955824, -0.011984481, -0.14203429, 0.033375036,
0.025635537, -0.014800292, 0.021314079, -0.004364666, -0.04288085, 0.03326454,
-0.017278133, -0.015381165, 0.039260007, 0.030663054, -0.009398082, 0.0016157854,
0.0058923406, 0.038551237, 0.029295184, 0.026982203, 0.0046961685, 0.020004513,
-0.0002823051, 0.029769791, -0.06499312, 0.0027809944, 0.033780456, -0.054227453,
0.015405423, 0.014887773, -0.06223483, 0.019264648, -0.02459092, -0.028076837,
-0.0007747023, 0.0050467625, 0.0054090386, 0.044855643, 0.014965291, 0.0039615836,
0.0364642, -0.0151543515, -0.010991144, 0.014066373, -0.020417051, -0.015295324,
0.0032331492, 0.029506348, -0.05176175, -0.012887155, -0.049193565, -0.025174212,
-0.0057166666, -0.00039162458, 0.00026217767, 0.07197446, -0.008801158, 0.023327848,
-0.013825299, 0.019054411, 0.026037995, 0.029275186, 0.0014251247, 0.0050785993,
0.014754311, -0.02803208, -0.020658525, 0.009343213, -0.038446713, -0.0021920993,
0.030721895, 0.0060682576, -0.0029885068, 0.0290252, -0.006677525, -0.007283009,
0.009740268, 0.021990651, -0.024688762, -0.04346818, -0.029574063, -0.007917782,
0.05176251, -0.03194417, -0.010803628, -0.03710345, 0.01654293, 0.035606004,
0.00021680047, -0.01902328, 0.017944654, -0.014634677, -0.08091542, -0.0005132855,
-0.0646727, -0.044102196, -0.05027943, 0.023779303, 0.0013730207, -0.00033895578,
-0.021345852, -0.01240465, -0.025230281, 0.02300386, 0.0059149317, -0.006499429,
-0.04890679, 0.049107686, 0.041578628, -0.021378888, 0.023876995, -0.012183695,
0.013455828, -0.005322552, -0.015810125, 0.025522828, -0.010038473, -0.006050725,
0.019967675, -0.003245995, -0.025224932, 0.0439206, -0.0012436587, 0.031723216,
-0.02448605, -0.015330694, 0.026584158, -0.0051061427, 0.066581406, -0.0038770763,
0.020848919, -0.028565202, -0.046525814, -0.058668178, 0.034280267, -0.0088918265,
-0.026706906, 0.018157527, 0.0033295872, 0.005130095, 0.019549852, -0.008804092,
0.0069478713, 0.009930967, -0.032567203, -0.042641874, -0.025361327, -0.02444165,
0.060229223, -0.00029296556, -0.032582972, 0.008229873, 0.00873001, -0.00815309,
-0.016171323, -0.029477486, 0.048059177, 1.3053809e-06, 0.0039567174, -0.040332206,
0.00093539315, 0.055955756, 0.008039591, -0.098208874, -0.024455974, 0.012274973,
-0.026007501, -0.0077049667, 0.0053886627, -0.019571463, -0.010535064, 0.03599966,
0.034453247, 0.042971645, -0.047814555, -0.006743227, 0.033782616, -0.03835373,
0.02963017, 0.057553414, 0.012134137, -0.01209421, 0.008088897, 0.029915666, 0.024700345,
0.03985199, 0.018612375, -0.007387687, -0.028683772, 0.015096653, 0.020524723,
-0.0025782587, 0.00621819, -0.018208023, -0.044911884, -0.006225681, 0.011868896,
-0.00073360204, -0.00060059247, 0.025633158, 0.010927056, 0.028156856, 0.013699513,
-0.01723093, -0.007418714, 0.001602008, -0.006724474, -0.042085167, -0.044802777,
-0.0005950661, -0.024677, 0.0152367465, -0.035699733, -0.021288736, -0.015362072,
0.04180836, -0.033493835, -0.00041006325, -0.030007878, -0.0003501069, 0.025993193,
0.030151015, -0.014684397, 0.036694344, -0.049206484, 0.05902542, -0.0063680233,
0.027811198, -0.054977603, 0.04245663, -0.044224508, 0.0074903113, 0.019687254,
-0.045596678, -0.01462281, -0.0051411195, 0.026682809, -0.0054900898, 0.016229948,
0.025660323, 0.031181056, -0.052291203, 0.03282995, -0.014796475, -0.0059239697,
0.012493835]

2.3.3 Ollama 使用 OpenAI 兼容的 Embedding 模型接口

`Ollama` 启动的 `Embedding REST API` 同样是兼容了 `OpenAI` 的接口规范，因此也可以用如下方式进行调用：

```
from openai import OpenAI
client = OpenAI(base_url="http://192.168.110.131:11434/v1")  # 需要根据自己的实际情况调整

response = client.embeddings.create(
  model="bge-m3",  # 这里替换成 bge-m3 模型
  input="你好，我是居居，很高兴认识你",
)

print(response)
```

CreateEmbeddingResponse(data=[Embedding(embedding=[-0.01922286, -0.041361455,
-0.012615179, -0.038346406, -0.006061248, -0.056141857, 0.024635639, -0.048374478,
0.021608315, 0.028389402, 0.011415558, -0.008346135, -0.01575168, 0.011952146,
0.0151369395, -0.0076897917, 0.019403733, -0.03635689, 0.01607708, -0.030296508,
-0.018613543, -0.0064844512, -0.00029064808, -0.010478618, 0.029013928, 0.01946554,
-0.055293787, -0.0024098635, 0.025135659, -0.04489269, 0.027139282, 0.015564225,
0.014931622, -0.04502287, -0.023891093, -0.024381446, 0.033852994, -0.0074271495,
-0.035470333, -0.015253574, 0.038780138, -0.010637472, 0.050409704, -0.047934886,
-0.0071231266, -0.02649895, -0.049684074, 0.01412076, 0.008318819, -0.0024914725,
-0.014517145, -0.006007267, 0.04962171, -0.0011974386, -0.045221522, 0.026687985,
0.036642727, -0.018013492, 0.0062242304, -0.037533417, -0.026383227, 0.0430352,
0.023990763, -0.046562776, 0.005052609, 0.1078952, -0.0252689, -0.036592163,
-0.017526496, -0.020836545, -0.014016081, 0.0073044035, 0.018462053, 0.009727984,
-0.05348882, 0.03130105, 0.054758947, 0.019654304, -0.070866816, -0.009440621,
0.052959364, 0.028022723, 0.020885868, -0.0027755883, 0.025919601, 0.0028719697,
-0.011645562, 0.07296864, -0.040803645, -0.014456895, -0.025773542, -0.024834368,
0.0054630465, -0.025666654, -0.027248958, -0.009498956, -0.02270712, 0.03019623,
0.04891002, 0.03496023, 0.032207686, 0.022681091, 0.0092448965, -0.020107258,
-0.0018319886, -0.007373062, 0.03264567, 0.03045322, -0.018678445, -0.012370282,
0.023889143, 0.005763869, 0.005892724, -0.036694903, 0.0010225775, -0.011602502,
-0.019340092, 0.014320763, 0.0038706549, -0.019690355, 0.0050612646, 0.042274926,
-0.0075149853, -0.012850127, -0.0029281029, -0.030729791, 0.020727703, 0.0390613,
-0.009327623, 0.008809925, 0.03356942, -0.012686815, -0.030047603, -0.042933784,
-0.015244054, -0.023267841, -0.0105991755, 0.01019861, -0.022745306, -0.028969925,
0.0044860677, 0.025056915, 0.0063823955, -0.051997595, -0.0007780814, -0.051995378,
0.017508458, 0.0270951, 0.018337702, -0.018740973, -0.01517844, 0.0043489956,
-0.007722501, 0.0066714576, 0.03081158, -0.013100324, 0.009403119, -0.0133148255,
-0.00554864, 0.039081782, -0.009327039, -0.01150965, 0.0057221474, -0.022880614,
0.03247104, 0.03626947, 0.003067755, 0.013796212, 0.007955853, -0.04198045, 0.010368125,
-0.0032647224, 0.044448704, 0.027452532, -0.0057273414, 0.030706208, 0.07107367,
-0.005020788, 0.013032576, -0.04025715, -0.012651873, 0.029171083, -0.008141952,
-0.016902551, -0.04220599, 0.025815476, -0.026521945, -0.04866242, -0.013637231,
0.03486512, -0.001918829, -0.057317823, 0.010962362, -0.020244285, 0.03866528,
0.010617372, -0.04377715, 0.009748239, 0.0045961896, -0.002983128, -0.004532409,
-0.019298308, 0.04492627, 0.017812751, -0.053677455, -0.0078686215, -0.037468098,
0.011929472, 0.01720958, -0.07143857, -0.03330745, 0.024078554, 0.024672244, -0.0569919,
-0.04148753, 0.050901547, 0.022151768, 0.031624585, 0.0021112408, 0.0004811191,
0.014531818, -0.008901751, 0.034024063, 0.02308123, -0.012181863, -0.037322327,
-0.016135737, 0.05219511, -0.00483262, -0.024687467, -0.007843138, 0.0142526515,
-0.0062834364, -0.0070895934, 0.016342908, -0.014341934, -0.055698976, 0.07432804,
-0.025424266, 0.027217498, -0.033968687, -0.019814044, -0.004727214, 0.0032408026,
-0.02595303, 0.004324861, -0.0019504167, 0.014354229, 0.014305528, 0.036802884,
-0.008998354, -0.06352971, 0.012633923, -0.033731535, 0.028493801, 0.00502899,
0.014065273, 0.024009604, -0.041306116, 0.00876087, -0.0015096144, -0.0581046,
0.009729838, 0.035152882, 0.0054527465, -0.0033466823, -0.007744983, -0.036980435,
0.02000405, 0.042424027, 0.007387792, 0.021873796, 0.025885409, 0.003218425,
-0.0067987675, 0.0067939996, -0.0058852965, -0.03311272, 0.051925275, -0.038485028,
0.017672395, 0.02621819, 0.042438682, -0.0080195675, 0.052674558, 0.014233261,
-0.08290935, -0.00034058988, 0.039673224, -0.04147254, -0.029645534, -0.009901535,
0.033171214, -0.03572185, -0.012022275, 0.0054901815, -0.0048021623, -0.1500592,
0.0084257675, -0.008138095, -0.007993522, -0.032590296, -0.04603252, -0.06815151,
-0.0007138535, -0.019334827, 0.036172323, 0.013635413, -0.013069142, 0.014766935,
0.0077957814, -0.032953933, 0.023373697, 0.033019446, -0.018915948, 0.018228997,
-0.0558767, -0.046538107, -0.021160431, 0.09153756, -0.05778473, 0.028429775,
0.00013784457, -0.00088175054, -0.046915453, -0.050807092, -0.002632421, 0.0011797646,
0.0285695, -0.00094530155, 0.015300199, 0.03612954, 0.027943866, -0.014544252,
-0.016660748, -0.032984182, 0.011127908, -0.008035259, 0.021325957, -0.0030667118,
0.0043126587, 0.005962908, -0.00042244705, -0.0065966123, 0.01868376, 0.03252872,

-0.021499867, -0.0038907684, 0.0037884223, 0.03002122, -0.042780697, -0.025766337,
0.011636676, -0.007580283, 0.017507449, -0.004628474, 0.038538925, -0.040182322,
-0.027409792, 0.043854233, 0.048351236, -0.03091617, 0.0027391338, 0.00036530296,
0.012099772, 0.022873694, -0.012164523, 0.028665284, 0.01798609, 0.031127907,
0.013131891, -0.01598989, 0.039874513, 0.0009227839, -0.060541965, -0.024316201,
-0.097156905, 0.0062047048, -0.02919612, -0.017234696, 0.032315448, -0.0012223435,
0.03263769, -0.0027417927, 0.036068577, 0.062720075, 0.22184578, 0.019815253, 0.01993091,
-0.043113314, 0.05674825, -0.0155262025, -0.020943452, 0.029482389, -0.023183936,
-0.004876635, 0.0021091257, -0.02630077, -0.017170096, 0.029073019, 0.016285948,
0.049894232, -0.0069553154, 0.0022705528, 0.03613966, -0.031493917, -0.0062867478,
-0.029011935, -0.037524033, -0.021112205, -0.086708285, -0.040370315, -0.014455863,
0.042079657, -0.0327298, 0.009579455, -0.013825583, 0.033582326, -0.0028413108,
0.002097946, 0.026424421, -0.002604664, 0.03141509, -0.042284306, -0.020693991,
0.052990846, -0.01812236, -0.025489392, 0.039961986, -0.010502693, -0.00163735,
0.024143638, 0.008718653, 0.0010326977, -0.013266786, -0.006186287, 0.04246847,
-0.008784686, 0.012570915, -0.020447822, -0.012151716, -0.047826536, -0.039175574,
-0.020817315, 0.025246646, 0.027339023, 0.025236443, 0.048806693, -0.011991174,
-0.022016289, -0.011961466, 0.0012723304, 0.008125226, -0.035645425, 0.011581718,
0.0074978457, -0.05309759, -0.010962783, -0.016122645, -0.0047364915, 0.010034488,
0.0030421098, 0.010316423, 0.038907442, 0.005337137, 0.057604548, -0.024876388,
0.013173044, -0.062358152, 0.055846933, 0.019336209, -0.0069100326, 0.0024203337,
0.04047966, 0.014952066, -0.025101326, 0.025459653, -0.007939787, -0.05776767,
0.0052549858, -0.012580737, -0.06010769, 0.010566356, -0.023706116, -0.004249546,
0.018867647, -0.028295763, -0.0071428707, 0.00011242026, 0.07138021, -0.0046134833,
-0.004373349, 0.024181986, 0.0366532, -0.015330882, 0.076670185, -0.044471033, 0.0236505,
-0.02199792, 0.047300745, 0.006907518, -0.031923965, -0.0044949916, -0.034935955,
0.061860606, -0.012920158, -0.009882616, -0.017172387, -0.030343372, 0.027154673,
-0.012200059, 0.002181584, 0.05755524, 0.030812616, -0.0052554836, 0.030308068,
0.011736867, -0.011554788, 0.019779095, 0.040169403, 0.05081635, 0.014287744,
0.042208575, 0.03118895, 0.02718363, -0.015795866, -0.03593932, 0.029518906, -0.0368933,
0.021886608, -0.02643343, -0.027600609, -0.024287364, -0.021399602, -9.439248e-05,
-0.0064367475, 0.039554097, 0.04583417, 0.007510894, -0.07607565, -0.009225462,
-0.006582759, -0.0148243345, -0.01328466, 0.030134458, -0.027474036, -0.003093169,
-0.007971443, 0.040055204, 0.13162026, -0.02346389, 0.014800054, -0.047165684,
-0.017866734, 0.012297593, -0.009757697, -0.014321117, -0.01788238, -0.039049163,
0.03508187, -0.0022837762, -0.007907967, -0.00075374526, -0.018473191, 0.008784031,
0.013422078, 0.011996075, 0.015082026, -0.039035235, 0.02770337, 0.00021796707,
0.008747956, 0.011951045, -0.0045034382, 0.02824308, -0.036295313, 0.00022648639,
0.082074136, -0.008459332, -0.026773563, 0.021175345, 0.005603807, 0.011868295,
-0.027673401, -0.030666795, -0.01869547, 0.017391281, 0.03777573, -0.021297984,
-0.0445547, 0.0042566187, -0.023678964, 0.0066275527, 0.043544076, -0.023232576,
-0.018033905, -0.016615305, 0.01890017, -0.016875628, 0.023675755, -0.013777414,
-0.025639104, 0.012749661, 0.008193755, -0.029998936, -0.016121306, -0.013959011,
0.021408115, 0.023548564, 0.0012766648, 0.0024227598, -0.00060618424, 0.033602092,
0.02401148, -0.021303771, 0.05285596, -0.0058503565, 0.012856821, 0.00946074,
-0.018707717, 0.034906205, -0.0033552044, -0.042422313, 0.027284978, -0.014717163,
0.009164387, 0.025540866, 0.0018863472, 0.037348382, 0.026557129, 0.027901705,
0.05202844, 0.020671928, 0.007082176, 0.0034235918, -0.031215776, -0.040995408,
0.003343065, 0.027570983, -0.015434854, -0.036407907, 0.031241637, -0.034739457,
-0.014406115, -0.062684685, 0.010075276, -0.031929497, 0.0025649823, -0.021319512,
0.005636287, 0.026131155, 0.021893157, 0.027501272, -0.067495435, -0.016828451,
-0.017330559, -0.027661966, 0.026027255, -0.04029333, -0.07741891, 0.04245963,
0.05786638, -0.048196528, 0.08805092, 0.019176718, 0.043642513, -0.016361246,
-0.008850222, -0.03140456, -0.0018215853, -0.02850662, -0.017091129, -0.0021216967,
-0.0008238167, 0.014921121, 0.00032447535, -0.044774458, -0.04502461, -0.0059552435,
0.061054993, -0.04434121, 0.014774075, 0.00038209726, 0.0702063, -0.021992086,
0.008455862, -0.02410905, 0.0652755, 0.005741814, 0.0019593763, 0.0047087143,
-0.024068253, 0.012610589, -0.06671961, 0.01753201, -0.0034162577, 0.0023024643,

-0.054187473, -0.011831421, 0.01715048, 0.04750781, -0.03139423, 0.02508704,
-0.010289396, 0.020973045, -0.014849384, -0.027357204, -0.01407778, 0.033853255,
-0.055280462, -0.013809495, 0.024581378, -0.012241737, -0.0009671642, -0.018698305,
0.028465739, 0.0010522557, 0.00916822, -0.04007455, -0.025335168, -0.011387842,
0.0055432064, 0.053396042, 0.01740828, -0.014137389, 0.02978183, 0.018102026, 0.05603387,
0.008667747, -0.052618537, 0.07024653, -0.052201025, 0.029271316, 0.027836822,
0.037358917, -0.018538367, -0.0098952325, 0.025678122, -0.022600275, -0.02866699,
0.047425956, 0.041606687, -0.045320027, 0.009275773, -0.015545199, -0.012686884,
-0.018841615, 0.021218903, -0.028941818, 0.027858598, -0.028179156, 0.029352563,
9.719478e-05, -0.008117893, 0.023555716, 0.0057699373, 0.0073017688, 0.038897607,
0.0060669654, 0.038502082, -0.03068005, 0.050719038, -0.013903436, -0.043558348,
-0.01102934, 0.017825447, -0.015128274, -0.027200924, -0.018548822, -0.043312766,
0.046628807, -0.0041413624, -0.015847849, 4.6269724e-05, -0.024507571, 0.059471335,
-0.017492197, -0.015842618, -0.033435818, 0.020426262, -0.15537652, 0.022720942,
-0.022978572, -0.0013574777, -0.03353008, -0.02227914, -0.04341785, -0.026962942,
0.036519103, 0.009219965, -0.040424213, -0.017088816, 0.01368187, -0.03543528,
0.04532406, 0.011525667, 0.032134064, -0.028047284, 0.006712295, 0.0066740294,
0.019953199, -0.019357666, 0.088674895, -0.01850979, 0.025244767, -0.018174196,
-0.042392906, 0.023810916, -0.042712238, 0.016077517, -0.0056865416, -0.007365048,
-0.0026553639, 0.05539162, 0.052702192, 0.012977329, -0.013324897, -0.046663057,
-0.004664647, 0.0038475136, -0.0083794715, 0.008683904, -0.028441006, 0.00988668,
0.019565033, 0.038489543, 0.021747977, -0.020754416, -0.042819865, -0.03930985,
0.01374576, -0.035338506, 0.030846687, 0.036640152, -0.037350155, 0.0038905449,
-0.020976484, 0.0058610975, -0.00070023915, 0.019661274, -0.054429367, -0.03100215,
-0.052285925, -0.0120779425, -0.074244425, 0.0006335962, -0.0017926287, 0.0014389432,
0.022021336, 0.036196936, -0.011613051, -0.093202114, 0.0026990643, -0.023806851,
-0.0054916134, 0.0011963513, 0.03746827, -0.022631988, -0.051280633, -0.03702279,
0.04518093, 0.007951412, -0.038437307, -0.034927223, 0.03933246, -0.015370903,
-0.02635214, -0.0041535352, -0.02957769, 0.012466689, -0.043131832, -0.020760499,
0.033076085, 0.061528783, -0.040835578, 0.0050104596, -0.006039951, -0.054586615,
-0.046932537, -0.016764142, -0.004739649, -0.01928498, -0.031564888, 0.0028312965,
-0.013649793, 0.020705067, 0.019288259, -0.00771178, 0.015912732, -0.0185576,
-0.027995558, 0.021834064, -0.007156878, -0.012114014, -0.04397438, 0.013293692,
0.031676818, -0.0026619898, -0.017556915, 0.06203819, -0.015809024, -0.0014322706,
-0.014704928, -0.046601627, 0.010122574, -0.01908295, -0.014350051, 0.023351008,
-0.009993024, -0.036077287, -0.0052948133, -0.06703926, 0.0069603934, -0.011365105,
0.018846352, -0.038462013, -0.00936936, 0.0022846924, -0.03624788, -0.012948129,
-0.013476335, -0.0182052, 0.0033161482, 0.0039851656, 0.0019273587, -0.018899929,
-0.0206676, 0.022174593, -0.040895447, 0.0038407452, 0.008989447, 0.060425438,
0.0058986675, -0.013137046, 0.0025831743, 0.03484262, 0.011654857, 0.006968705,
0.040688474, -0.001316991, -0.019467473, 0.027343286, -0.022299161, -0.05269345,
-0.024993394, -0.035053432, -0.019827059, 0.030563287, 0.008705064, -0.0045054746,
0.0034345994, 0.07223343, -0.0068709, -0.0078441575, 0.05365767, -0.015098801,
-0.0019810034, -0.008233295, 0.08352503, 0.018159185, 0.011276068, 0.030084705,
0.014642007, 0.007945344, 0.0064324383, -0.021965735, 0.022328451, 0.0013204975,
-0.0054936996, -0.0153985005, 0.018929703, 0.0011742727, -0.026166951, -0.006397983,
0.02537267, 0.039143644, 0.0054356274, 0.046194375, -0.0054917075, 0.01610236,
-0.06036488, -0.03880828, 0.01609657, -0.0067152907, -0.012388587, 0.020276126,
-0.013779218, -0.0008071725, -0.013861764, 0.024183014, -0.014108998, -0.0054775397,
-0.014219473, 0.023931691, 0.036456257, 0.029436398, 0.02639856, -0.009838244,
-0.017583802, 0.0084198555, -0.011142715, 0.0033209193, -0.044683803, 0.03729297,
0.025573304, -0.0038417936, -0.076640286, 0.016013449, -0.00077766576, -0.018147612,
-0.04962999, -0.012260296, -0.019350484, 0.010270535, 0.012180092, 0.003858917,
0.05388526, 0.061142597, 0.049799904, -0.025325377, 0.01596864, 0.0037506404,
-0.015387656, 0.010250521], index=0, object='embedding')], model='bge-m3', object='list',
usage=Usage(prompt_tokens=10, total_tokens=10))

```python
print(response.data[0].embedding)
```

[-0.01922286, -0.041361455, -0.012615179, -0.038346406, -0.006061248, -0.056141857,
0.024635639, -0.048374478, 0.021608315, 0.028389402, 0.011415558, -0.008346135,
-0.01575168, 0.011952146, 0.0151369395, -0.0076897917, 0.019403733, -0.03635689,
0.01607708, -0.030296508, -0.018613543, -0.0064844512, -0.00029064808, -0.010478618,
0.029013928, 0.01946554, -0.055293787, -0.0024098635, 0.025135659, -0.04489269,
0.027139282, 0.015564225, 0.014931622, -0.04502287, -0.023891093, -0.024381446,
0.033852994, -0.0074271495, -0.035470333, -0.015253574, 0.038780138, -0.010637472,
0.050409704, -0.047934886, -0.0071231266, -0.02649895, -0.049684074, 0.01412076,
0.008318819, -0.0024914725, -0.014517145, -0.006007267, 0.04962171, -0.0011974386,
-0.045221522, 0.026687985, 0.036642727, -0.018013492, 0.0062242304, -0.037533417,
-0.026383227, 0.0430352, 0.023990763, -0.046562776, 0.005052609, 0.1078952, -0.0252689,
-0.036592163, -0.017526496, -0.020836545, -0.014016081, 0.0073044035, 0.018462053,
0.009727984, -0.05348882, 0.03130105, 0.054758947, 0.019654304, -0.070866816,
-0.009440621, 0.052959364, 0.028022723, 0.020885868, -0.0027755883, 0.025919601,
0.0028719697, -0.011645562, 0.07296864, -0.040803645, -0.014456895, -0.025773542,
-0.024834368, 0.0054630465, -0.025666654, -0.027248958, -0.009498956, -0.02270712,
0.03019623, 0.04891002, 0.03496023, 0.032207686, 0.022681091, 0.0092448965, -0.020107258,
-0.0018319886, -0.007373062, 0.03264567, 0.03045322, -0.018678445, -0.012370282,
0.023889143, 0.005763869, 0.005892724, -0.036694903, 0.0010225775, -0.011602502,
-0.019340092, 0.014320763, 0.0038706549, -0.019690355, 0.0050612646, 0.042274926,
-0.0075149853, -0.012850127, -0.0029281029, -0.030729791, 0.020727703, 0.0390613,
-0.009327623, 0.008809925, 0.03356942, -0.012686815, -0.030047603, -0.042933784,
-0.015244054, -0.023267841, -0.0105991755, 0.01019861, -0.022745306, -0.028969925,
0.0044860677, 0.025056915, 0.0063823955, -0.051997595, -0.0007780814, -0.051995378,
0.017508458, 0.0270951, 0.018337702, -0.018740973, -0.01517844, 0.0043489956,
-0.007722501, 0.0066714576, 0.03081158, -0.013100324, 0.009403119, -0.0133148255,
-0.00554864, 0.039081782, -0.009327039, -0.01150965, 0.0057221474, -0.022880614,
0.03247104, 0.03626947, 0.003067755, 0.013796212, 0.007955853, -0.04198045, 0.010368125,
-0.0032647224, 0.044448704, 0.027452532, -0.0057273414, 0.030706208, 0.07107367,
-0.005020788, 0.013032576, -0.04025715, -0.012651873, 0.029171083, -0.008141952,
-0.016902551, -0.04220599, 0.025815476, -0.026521945, -0.04866242, -0.013637231,
0.03486512, -0.001918829, -0.057317823, 0.010962362, -0.020244285, 0.03866528,
0.010617372, -0.04377715, 0.009748239, 0.0045961896, -0.002983128, -0.004532409,
-0.019298308, 0.04492627, 0.017812751, -0.053677455, -0.0078686215, -0.037468098,
0.011929472, 0.01720958, -0.07143857, -0.03330745, 0.024078554, 0.024672244, -0.0569919,
-0.04148753, 0.050901547, 0.022151768, 0.031624585, 0.0021112408, 0.0004811191,
0.014531818, -0.008901751, 0.034024063, 0.02308123, -0.012181863, -0.037322327,
-0.016135737, 0.05219511, -0.00483262, -0.024687467, -0.007843138, 0.0142526515,
-0.0062834364, -0.0070895934, 0.016342908, -0.014341934, -0.055698976, 0.07432804,
-0.025424266, 0.027217498, -0.033968687, -0.019814044, -0.004727214, 0.0032408026,
-0.02595303, 0.004324861, -0.0019504167, 0.014354229, 0.014305528, 0.036802884,
-0.008998354, -0.06352971, 0.012633923, -0.033731535, 0.028493801, 0.00502899,
0.014065273, 0.024009604, -0.041306116, 0.00876087, -0.0015096144, -0.0581046,
0.009729838, 0.035152882, 0.0054527465, -0.0033466823, -0.007744983, -0.036980435,
0.02000405, 0.042424027, 0.007387792, 0.021873796, 0.025885409, 0.003218425,
-0.0067987675, 0.0067939996, -0.0058852965, -0.03311272, 0.051925275, -0.038485028,
0.017672395, 0.02621819, 0.042438682, -0.0080195675, 0.052674558, 0.014233261,
-0.08290935, -0.00034058988, 0.039673224, -0.04147254, -0.029645534, -0.009901535,
0.033171214, -0.03572185, -0.012022275, 0.0054901815, -0.0048021623, -0.1500592,
0.0084257675, -0.008138095, -0.007993522, -0.032590296, -0.04603252, -0.06815151,
-0.0007138535, -0.019334827, 0.036172323, 0.013635413, -0.013069142, 0.014766935,
0.0077957814, -0.032953933, 0.023373697, 0.033019446, -0.018915948, 0.018228997,
-0.0558767, -0.046538107, -0.021160431, 0.09153756, -0.05778473, 0.028429775,
0.00013784457, -0.00088175054, -0.046915453, -0.050807092, -0.002632421, 0.0011797646,
0.0285695, -0.00094530155, 0.015300199, 0.03612954, 0.027943866, -0.014544252,
-0.016660748, -0.032984182, 0.011127908, -0.008035259, 0.021325957, -0.0030667118,
0.0043126587, 0.005962908, -0.00042244705, -0.0065966123, 0.01868376, 0.03252872,
-0.021499867, -0.0038907684, 0.0037884223, 0.03002122, -0.042780697, -0.025766337,

0.011636676, -0.007580283, 0.017507449, -0.004628474, 0.038538925, -0.040182322,
-0.027409792, 0.043854233, 0.048351236, -0.03091617, 0.0027391338, 0.00036530296,
0.012099772, 0.022873694, -0.012164523, 0.028665284, 0.01798609, 0.031127907,
0.013131891, -0.01598989, 0.039874513, 0.0009227839, -0.060541965, -0.024316201,
-0.097156905, 0.0062047048, -0.02919612, -0.017234696, 0.032315448, -0.0012223435,
0.03263769, -0.0027417927, 0.036068577, 0.062720075, 0.22184578, 0.019815253, 0.01993091,
-0.043113314, 0.05674825, -0.0155262025, -0.020943452, 0.029482389, -0.023183936,
-0.004876635, 0.0021091257, -0.02630077, -0.017170096, 0.029073019, 0.016285948,
0.049894232, -0.0069553154, 0.0022705528, 0.03613966, -0.031493917, -0.0062867478,
-0.029011935, -0.037524033, -0.021112205, -0.086708285, -0.040370315, -0.014455863,
0.042079657, -0.0327298, 0.009579455, -0.013825583, 0.033582326, -0.0028413108,
0.002097946, 0.026424421, -0.002604664, 0.03141509, -0.042284306, -0.020693991,
0.052990846, -0.01812236, -0.025489392, 0.039961986, -0.010502693, -0.00163735,
0.024143638, 0.008718653, 0.0010326977, -0.013266786, -0.006186287, 0.04246847,
-0.008784686, 0.012570915, -0.020447822, -0.012151716, -0.047826536, -0.039175574,
-0.020817315, 0.025246646, 0.027339023, 0.025236443, 0.048806693, -0.011991174,
-0.022016289, -0.011961466, 0.0012723304, 0.008125226, -0.035645425, 0.011581718,
0.0074978457, -0.05309759, -0.010962783, -0.016122645, -0.0047364915, 0.010034488,
0.0030421098, 0.010316423, 0.038907442, 0.005337137, 0.057604548, -0.024876388,
0.013173044, -0.062358152, 0.055846933, 0.019336209, -0.0069100326, 0.0024203337,
0.04047966, 0.014952066, -0.025101326, 0.025459653, -0.007939787, -0.05776767,
0.0052549858, -0.012580737, -0.06010769, 0.010566356, -0.023706116, -0.004249546,
0.018867647, -0.028295763, -0.0071428707, 0.00011242026, 0.07138021, -0.0046134833,
-0.004373349, 0.024181986, 0.0366532, -0.015330882, 0.076670185, -0.044471033, 0.0236505,
-0.02199792, 0.047300745, 0.006907518, -0.031923965, -0.0044949916, -0.034935955,
0.061860606, -0.012920158, -0.009882616, -0.017172387, -0.030343372, 0.027154673,
-0.012200059, 0.002181584, 0.05755524, 0.030812616, -0.0052554836, 0.030308068,
0.011736867, -0.011554788, 0.019779095, 0.040169403, 0.05081635, 0.014287744,
0.042208575, 0.03118895, 0.02718363, -0.015795866, -0.03593932, 0.029518906, -0.0368933,
0.021886608, -0.02643343, -0.027600609, -0.024287364, -0.021399602, -9.439248e-05,
-0.0064367475, 0.039554097, 0.04583417, 0.007510894, -0.07607565, -0.009225462,
-0.006582759, -0.0148243345, -0.01328466, 0.030134458, -0.027474036, -0.003093169,
-0.007971443, 0.040055204, 0.13162026, -0.02346389, 0.014800054, -0.047165684,
-0.017866734, 0.012297593, -0.009757697, -0.014321117, -0.01788238, -0.039049163,
0.03508187, -0.0022837762, -0.007907967, -0.00075374526, -0.018473191, 0.008784031,
0.013422078, 0.011996075, 0.015082026, -0.039035235, 0.02770337, 0.00021796707,
0.008747956, 0.011951045, -0.0045034382, 0.02824308, -0.036295313, 0.00022648639,
0.082074136, -0.008459332, -0.026773563, 0.021175345, 0.005603807, 0.011868295,
-0.027673401, -0.030666795, -0.01869547, 0.017391281, 0.03777573, -0.021297984,
-0.0445547, 0.0042566187, -0.023678964, 0.0066275527, 0.043544076, -0.023232576,
-0.018033905, -0.016615305, 0.01890017, -0.016875628, 0.023675755, -0.013777414,
-0.025639104, 0.012749661, 0.008193755, -0.029998936, -0.016121306, -0.013959011,
0.021408115, 0.023548564, 0.0012766648, 0.0024227598, -0.00060618424, 0.033602092,
0.02401148, -0.021303771, 0.05285596, -0.0058503565, 0.012856821, 0.00946074,
-0.018707717, 0.034906205, -0.0033552044, -0.042422313, 0.027284978, -0.014717163,
0.009164387, 0.025540866, 0.0018863472, 0.037348382, 0.026557129, 0.027901705,
0.05202844, 0.020671928, 0.007082176, 0.0034235918, -0.031215776, -0.040995408,
0.003343065, 0.027570983, -0.015434854, -0.036407907, 0.031241637, -0.034739457,
-0.014406115, -0.062684685, 0.010075276, -0.031929497, 0.0025649823, -0.021319512,
0.005636287, 0.026131155, 0.021893157, 0.027501272, -0.067495435, -0.016828451,
-0.017330559, -0.027661966, 0.026027255, -0.04029333, -0.07741891, 0.04245963,
0.05786638, -0.048196528, 0.08805092, 0.019176718, 0.043642513, -0.016361246,
-0.008850222, -0.03140456, -0.0018215853, -0.02850662, -0.017091129, -0.0021216967,
-0.0008238167, 0.014921121, 0.00032447535, -0.044774458, -0.04502461, -0.0059552435,
0.061054993, -0.04434121, 0.014774075, 0.00038209726, 0.0702063, -0.021992086,
0.008455862, -0.02410905, 0.0652755, 0.005741814, 0.0019593763, 0.0047087143,
-0.024068253, 0.012610589, -0.06671961, 0.01753201, -0.0034162577, 0.0023024643,
-0.054187473, -0.011831421, 0.01715048, 0.04750781, -0.03139423, 0.02508704,

-0.010289396, 0.020973045, -0.014849384, -0.027357204, -0.01407778, 0.033853255,
-0.055280462, -0.013809495, 0.024581378, -0.012241737, -0.0009671642, -0.018698305,
0.028465739, 0.0010522557, 0.00916822, -0.04007455, -0.025335168, -0.011387842,
0.0055432064, 0.053396042, 0.01740828, -0.014137389, 0.02978183, 0.018102026, 0.05603387,
0.008667747, -0.052618537, 0.07024653, -0.052201025, 0.029271316, 0.027836822,
0.037358917, -0.018538367, -0.0098952325, 0.025678122, -0.022600275, -0.02866699,
0.047425956, 0.041606687, -0.045320027, 0.009275773, -0.015545199, -0.012686884,
-0.018841615, 0.021218903, -0.028941818, 0.027858598, -0.028179156, 0.029352563,
9.719478e-05, -0.008117893, 0.023555716, 0.0057699373, 0.0073017688, 0.038897607,
0.0060669654, 0.038502082, -0.03068005, 0.050719038, -0.013903436, -0.043558348,
-0.01102934, 0.017825447, -0.015128274, -0.027200924, -0.018548822, -0.043312766,
0.046628807, -0.0041413624, -0.015847849, 4.6269724e-05, -0.024507571, 0.059471335,
-0.017492197, -0.015842618, -0.033435818, 0.020426262, -0.15537652, 0.022720942,
-0.022978572, -0.0013574777, -0.03353008, -0.02227914, -0.04341785, -0.026962942,
0.036519103, 0.009219965, -0.040424213, -0.017088816, 0.01368187, -0.03543528,
0.04532406, 0.011525667, 0.032134064, -0.028047284, 0.006712295, 0.0066740294,
0.019953199, -0.019357666, 0.088674895, -0.01850979, 0.025244767, -0.018174196,
-0.042392906, 0.023810916, -0.042712238, 0.016077517, -0.0056865416, -0.007365048,
-0.0026553639, 0.05539162, 0.052702192, 0.012977329, -0.013324897, -0.046663057,
-0.004664647, 0.0038475136, -0.0083794715, 0.008683904, -0.028441006, 0.00988668,
0.019565033, 0.038489543, 0.021747977, -0.020754416, -0.042819865, -0.03930985,
0.01374576, -0.035338506, 0.030846687, 0.036640152, -0.037350155, 0.0038905449,
-0.020976484, 0.0058610975, -0.00070023915, 0.019661274, -0.054429367, -0.03100215,
-0.052285925, -0.0120779425, -0.074244425, 0.0006335962, -0.0017926287, 0.0014389432,
0.022021336, 0.036196936, -0.011613051, -0.093202114, 0.0026990643, -0.023806851,
-0.0054916134, 0.0011963513, 0.03746827, -0.022631988, -0.051280633, -0.03702279,
0.04518093, 0.007951412, -0.038437307, -0.034927223, 0.03933246, -0.015370903,
-0.02635214, -0.0041535352, -0.02957769, 0.012466689, -0.043131832, -0.020760499,
0.033076085, 0.061528783, -0.040835578, 0.0050104596, -0.006039951, -0.054586615,
-0.046932537, -0.016764142, -0.004739649, -0.01928498, -0.031564888, 0.0028312965,
-0.013649793, 0.020705067, 0.019288259, -0.00771178, 0.015912732, -0.0185576,
-0.027995558, 0.021834064, -0.007156878, -0.012114014, -0.04397438, 0.013293692,
0.031676818, -0.0026619898, -0.017556915, 0.06203819, -0.015809024, -0.0014322706,
-0.014704928, -0.046601627, 0.010122574, -0.01908295, -0.014350051, 0.023351008,
-0.009993024, -0.036077287, -0.0052948133, -0.06703926, 0.0069603934, -0.011365105,
0.018846352, -0.038462013, -0.00936936, 0.0022846924, -0.03624788, -0.012948129,
-0.013476335, -0.0182052, 0.0033161482, 0.0039851656, 0.0019273587, -0.018899929,
-0.0206676, 0.022174593, -0.040895447, 0.0038407452, 0.008989447, 0.060425438,
0.0058986675, -0.013137046, 0.0025831743, 0.03484262, 0.011654857, 0.006968705,
0.040688474, -0.001316991, -0.019467473, 0.027343286, -0.022299161, -0.05269345,
-0.024993394, -0.035053432, -0.019827059, 0.030563287, 0.008705064, -0.0045054746,
0.0034345994, 0.07223343, -0.0068709, -0.0078441575, 0.05365767, -0.015098801,
-0.0019810034, -0.008233295, 0.08352503, 0.018159185, 0.011276068, 0.030084705,
0.014642007, 0.007945344, 0.0064324383, -0.021965735, 0.022328451, 0.0013204975,
-0.0054936996, -0.0153985005, 0.018929703, 0.0011742727, -0.026166951, -0.006397983,
0.02537267, 0.039143644, 0.0054356274, 0.046194375, -0.0054917075, 0.01610236,
-0.06036488, -0.03880828, 0.01609657, -0.0067152907, -0.012388587, 0.020276126,
-0.013779218, -0.0008071725, -0.013861764, 0.024183014, -0.014108998, -0.0054775397,
-0.014219473, 0.023931691, 0.036456257, 0.029436398, 0.02639856, -0.009838244,
-0.017583802, 0.0084198555, -0.011142715, 0.0033209193, -0.044683803, 0.03729297,
0.025573304, -0.0038417936, -0.076640286, 0.016013449, -0.00077766576, -0.018147612,
-0.04962999, -0.012260296, -0.019350484, 0.010270535, 0.012180092, 0.003858917,
0.05388526, 0.061142597, 0.049799904, -0.025325377, 0.01596864, 0.0037506404,
-0.015387656, 0.010250521]

`Ollama` 兼容的 `OpenAI` 接口，其返回的结果与 `Ollama` 的 `REST API` 接口返回的结果是一样的。同时 `input` 参数也同样支持单个输入和多个输入。如果需要嵌入多个文本，则需要将 `input` 参数设置为列表，如下所示：

```
from openai import OpenAI
client = OpenAI(base_url="http://192.168.110.131:11434/v1")

batch_input = ["天为什么是蓝色的？", "草为什么是绿色的？"]


response = client.embeddings.create(
  model="bge-m3",
  input=batch_input
)

print(response)
```

```
CreateEmbeddingResponse(data=[Embedding(embedding=[-0.042137332, 0.005296808,
-0.037715282, -0.06215416, -0.030467484, -0.007836851, -0.043893967, 0.0022080292,
0.016335009, -0.0117954165, -0.004529275, 0.025638063, 0.007796983, 0.0063098017,
0.018519185, -0.022322154, 0.029461581, 0.022574248, 0.00038702495, 0.019533236,
-0.009881311, -0.0383717, -0.032535516, 0.016021805, -0.006125824, 0.0075718816,
0.016597407, -0.017723264, -0.0002764759, 0.001488903, 0.004165976, 0.0616205,
-0.029179199, -0.022410607, 0.0003814827, -0.017742423, -0.005474743, 0.0049243644,
-0.040230345, -0.032896325, 0.009739351, -0.03913332, 0.017520385, -0.00029088696,
0.027668685, -0.046231378, -0.011650233, -0.03293465, -0.04470256, -0.0478838,
0.008582121, 0.0038168635, 0.088327065, -0.016265068, 0.011740456, -0.0029550614,
0.028078863, -0.030943036, -0.043092407, -0.028814875, -0.014810867, -0.0020290688,
-0.041463394, 0.008612065, -0.013520107, 0.08999495, 0.02035601, -0.0044486164,
-0.020074788, -0.004833883, -0.028073747, 0.00032702446, 0.0047799097, -0.0044234022,
-0.046205215, -0.01817478, -0.0045246906, -0.0123508675, -0.040190786, 0.04064541,
0.09014224, 0.014359349, -0.02341912, 0.014465041, -0.0053078295, 0.05538501,
-0.034623098, 0.041395552, -0.029607566, -0.037130527, -0.002067333, 0.042320542,
0.017242383, -0.056331713, -0.023820797, -0.0020812673, -0.034105197, 0.0056840624,
0.027562784, 0.026438197, 0.015122961, -0.029478177, -0.010766563, -0.00533037,
-0.020935973, 0.00032368215, 0.051139206, 0.009538334, -0.02400277, 0.019776152,
0.0584742, 0.04907174, 0.016894652, 0.031076962, -0.03176524, 0.009510206, -0.02218482,
0.018387897, 0.0031903163, 0.010497863, -0.011635472, 0.07276368, -0.0064129275,
-0.045014057, -0.013379973, -0.028407896, 0.016923483, 0.04445915, -0.016673427,
0.039915126, -0.025509529, 0.019542592, -0.047511388, 0.028730273, -0.054099925,
0.0014411394, 0.025573539, -0.010792686, 0.031407714, -0.013306026, 0.052538764,
0.016519766, -0.027834127, -0.011570048, 0.035803452, 0.02007018, 0.0024324271,
-0.031974126, -0.0026232207, -0.0049809525, -0.0052737803, 0.023195265, 0.060987324,
-0.002575687, -0.076468945, -0.006767327, -0.053074267, 0.009725925, -0.05268764,
0.032771066, 0.038590457, -0.0039342837, -0.045914114, 0.02513468, 0.0021120834,
-0.03730243, 0.02977119, 0.06439825, -0.008863884, -0.020030886, -0.02747345,
-0.0073932316, -0.0048822267, -0.0067345817, -0.019864578, 0.015068145, 0.024992965,
-0.013167202, 0.027090143, 0.004015794, -0.08040249, -0.03474101, 0.0045948057,
-0.030831112, -0.012045554, 0.02483818, -0.015777161, -0.009927092, 0.018526495,
0.0118054245, 0.0211727, -0.037752703, -0.001975264, -0.0012090072, 0.027399225,
0.004032278, 0.022544393, 0.029868845, 0.00067225663, -0.004111884, -0.0016398347,
-0.0004749759, 0.041708183, -0.017336763, -0.0136280665, 0.019985555, -0.09729808,
-0.08953787, 0.04107954, -0.011029402, -0.018815925, 0.011755434, 0.022931322,
0.014722142, -0.021672897, -0.023385474, 0.021395497, 0.024507977, 0.04235797,
0.015848517, -0.02279181, 0.006232009, -0.031116417, -0.044243816, 0.048988286,
0.0065010227, 0.011547768, 0.0549056, -0.04781384, 0.0029705032, -0.026103448,
0.024926864, 0.012209596, -0.003091049, 0.02546409, -0.038644683, -0.007874508,
-0.0073739677, 0.036731694, 0.024388144, -0.044597827, -0.021421095, -0.0059583816,
0.05000934, 0.009413394, 0.040630322, -0.03523095, -0.004540993, 0.0022993265,
-0.023829076, 0.02254886, -0.017369289, -0.070587724, 0.0075032255, 0.031555068,
0.008978053, 0.009435259, -0.017410524, 0.042360272, 0.03270071, 0.010293682,
-0.024027297, -0.008167552, 0.03461271, 0.013336475, -0.009405411, -0.0094858855,
0.0004914884, 0.0026585604, 0.0040770057, -0.024476796, -0.0046170247, -0.015485166,
-0.0057349876, 0.01156721, -0.0007533831, 0.003609351, -0.01711615, -0.0011313154,
-0.019069092, -0.0044528176, -0.023705354, 0.015393332, 0.029787816, -0.0032745432,
-0.038109522, 0.019249914, 0.0008609008, -0.004534888, 0.031711407, -0.020981707,
-0.0397812, 0.04040954, 0.020614095, 0.019030828, 0.02284263, -0.035594705, -0.16899337,
0.018541625, 0.014923138, -0.007641145, 0.0075152004, -0.024446165, -0.0076794797,
-0.018340291, 0.00046605608, -0.03280913, -0.01234831, -0.021243177, -0.031833846,
-0.00918224, -0.019617198, -0.03971306, 0.01629078, -0.035811763, 0.027423305,
-0.027777009, -0.015663758, -0.024587302, 0.05940994, -0.002367637, 0.017544618,
0.013158937, -0.027401974, 0.024363685, -0.041683983, -0.07853095, 0.04206934,
0.020653142, -0.03020147, 0.032506283, 0.028300753, -0.019385654, 0.030426295,
-0.048927456, -0.0049383417, 0.02057371, -0.007289443, 0.02098331, -0.024779443,
0.030403, -0.032178733, 0.0074722483, -0.026696924, -0.0057991054, 0.0046112672,
```

0.010001329, 0.0052110488, 0.003479432, 0.0062448625, 0.00091557216, -0.030177973,
0.01767895, -0.0018189346, 0.005169421, 0.022286616, -0.0017267654, -0.015250424,
-0.049802978, 0.025996951, 0.015015538, 0.011650537, -0.015945734, 0.013876556,
0.031699087, 0.027250074, 0.0066137933, 0.019414512, 0.027010594, 0.0034491217,
-0.04149347, 0.03261539, 0.05666244, -0.035440873, -0.038834725, -0.00057862076,
-0.094313756, -0.018587569, -0.013786597, 0.013335928, -0.026358103, 0.018386213,
-0.034061395, -0.029979499, 0.03101407, 0.007732108, 0.2302883, 0.053795822, -0.04571347,
-0.024319816, 0.08307778, -0.038508948, 0.0057074195, -0.005989547, -0.03456581,
-0.028869793, -0.04034803, -0.016712103, -0.0066257915, -0.00882439, -0.0077595403,
0.009169582, -0.028189782, -0.0033542505, 0.056126468, -0.023846501, -0.0066834236,
-0.0039994065, 0.025167191, 0.0070630596, 0.0010949638, -0.022675293, 0.01223915,
0.07312742, 0.00886766, -0.006981496, -0.028843962, -0.016596032, -0.014287267,
0.011711063, -0.023769658, -0.007999069, 0.018134821, 0.013600745, 0.06240502,
-0.027478902, -0.039845858, -0.0061672656, 0.046465103, 0.033261463, 0.08106834,
0.015202136, 0.03412636, -0.062328212, -0.02208452, -0.008780762, -0.02163691,
-0.010332136, -0.053890415, 0.030947028, -0.08494406, -0.029017588, -0.03273794,
-0.021868378, 0.005420783, 0.014095832, 0.03149366, 0.016462496, 0.0173673, -0.043623086,
-0.0020759876, -0.03034031, 0.04324677, -0.07422934, 0.017031752, 0.006030616,
-0.014976466, -0.016536182, -0.014204306, 0.012176292, 0.035498124, 0.048893087,
0.033731975, 0.05893236, -0.04438582, -0.019384483, 0.042567324, -0.035785895,
-0.027839916, 0.010269249, -0.0029354466, 0.034064908, 0.042051334, -0.02377663,
0.00018843584, -0.035804797, 0.010596093, 0.040386043, -0.07433216, -0.012972507,
0.0020062684, 0.00087044924, -0.025378624, -0.037642613, -0.015890175, -0.0076076183,
-0.02487156, -0.029907221, -0.013884005, 0.0052179443, -0.0051465426, -0.029555114,
0.020426996, 0.040726706, -0.027832404, 0.042246837, -0.0033184467, 0.001172159,
-0.0557444, 0.03280031, 0.03765769, 0.042062365, 0.013983726, 0.0004090046, 0.0014503683,
0.04151597, -0.027207311, -0.03811464, 0.04106502, 0.039898492, -0.018392889,
-0.00559719, 0.0068726214, 0.0076489043, 0.0006230939, 0.033178918, 0.01079643,
-0.041384447, -0.02852786, 0.016783055, 0.055578187, 0.0037498036, 0.03197655,
0.062441908, 0.002617525, 0.05285312, 0.030939603, -0.038866654, -0.0042523113,
-0.045382734, -0.009775364, -0.034698278, 0.048598345, -0.06818018, -0.02444149,
-0.00949042, 0.04478301, -0.021221437, 0.014549152, -0.03238616, 0.012591191,
0.0032901457, 0.009852417, -0.014907928, -0.044620603, -0.011357798, -0.026407277,
-0.005224016, 0.039710406, 0.10613283, 0.008799815, 0.019459253, -0.027430058,
0.040994428, 0.06255134, -0.0154246325, 0.020769319, 0.010436879, 0.009245113,
0.039452195, 0.0220522, -0.036322378, -0.022910323, -0.026929844, -0.06490222,
0.010154379, 0.036104105, 0.01758559, 0.016539378, 0.045427833, 0.013154021, 0.009437805,
-0.018297847, -0.037246346, 0.023390587, -0.020376557, 0.01932413, 0.095888935,
0.00019674169, -0.007847683, 0.07149473, 0.032855626, -0.012178423, 0.017537348,
-0.0023458574, -0.019856326, 0.06914988, -0.053961936, 0.0052872654, 0.012509775,
0.02971595, -0.02288805, 0.027495164, -0.000319835, -0.012909792, -0.0010698951,
-0.03811992, -0.02886807, -0.03195666, -0.00096493005, 0.008205409, -0.011321164,
-0.03252674, 0.0045426735, -0.00061344256, -0.043989014, 0.0035914248, -0.01197999,
0.045499917, -0.043762688, -0.0055888724, 0.06170304, -0.030959107, -0.010835247,
0.009822792, -0.02960603, 0.0030709354, 0.01912414, -0.03826324, -0.009199489,
0.016952321, 0.019091725, 0.027671369, 0.035184704, 0.005566593, -0.004224325,
-0.023847291, 0.020012423, 0.028360244, 0.059198152, -0.00022528647, 0.005460027,
0.011523086, 0.016808476, 0.027759623, 0.02160105, -0.004696154, 0.0938206, 0.02182346,
-0.012023529, -0.02564297, -0.0050284294, 0.018947987, 0.027122408, -0.047256127,
-0.0041446886, -0.0007468413, 0.010792308, 0.012999448, 0.0009446954, 0.0041484386,
-0.009155726, 0.038582344, 0.017999995, -0.01361257, -0.029035058, -0.015319298,
0.018348051, -0.037154764, -0.021188546, -0.030136174, 0.01371045, -0.0252252,
0.036648184, -0.03341299, -0.015977893, -0.009471925, 0.07168737, -0.06356442,
-0.00022004705, -0.024585567, 0.026660563, 0.014813339, -0.011186062, 0.02635014,
-0.01858569, -0.011048534, -0.0037236298, -0.013637644, 0.019399958, -0.020726122,
-0.005873677, -0.05457326, -0.0010754383, -0.032321937, 0.024502382, -0.017577903,
0.05186453, -0.041495655, -0.034449518, -0.022555443, -0.003808817, -0.0064183106,
-0.029723026, -0.010237768, 0.024252448, 0.0057710805, -0.02287218, 0.046242468,

-0.033879757, -0.016698122, -0.027835608, -0.031813916, -0.023639053, -0.06083357,
0.0026263874, 0.023786323, -0.0026732201, 0.014249788, 0.015631843, 0.0615537,
-0.008399952, -0.073089555, -0.013478393, -0.06011397, 0.020935731, -0.038242858,
0.010878561, -0.008781969, -0.013526375, 0.023306286, -0.0076850546, -0.004033106,
-0.01432605, 0.024405079, -0.022795958, 0.020023946, 0.062027246, 0.041369814,
-0.034915254, 0.05650035, -0.033667177, 0.0017321882, 0.028220907, 0.02724711,
-0.011040877, 0.009466217, 0.06299186, -0.032366026, -0.031843692, 0.013810198,
0.036252715, -0.012403185, -0.03576739, -0.03492998, 0.036709167, -0.02419729,
0.065110564, -0.019976975, 0.010940757, -0.010370927, 0.0048032184, 0.00061093795,
0.042379238, -0.02673291, -0.0049099107, 0.0005162485, 0.083228916, 0.004646965,
-0.03581079, -0.033165492, 0.016541002, -0.04952429, -0.021216413, 0.019397082,
0.009300314, -0.008058479, -0.002323441, -0.008978433, -0.010337525, -0.005220179,
-0.031576138, 0.008024545, -0.016106777, 0.024416858, -0.012258581, -0.027702363,
0.018079732, -0.058961343, -0.022970779, -0.14390658, 0.0037366857, -0.0032212595,
-0.01208507, 0.0030887993, -0.009112169, -0.023670638, -0.004746355, -0.003891055,
-0.04284588, 0.01887782, 0.04493371, -0.04388891, -0.023415668, 0.025585275, 0.03896533,
0.013118991, 0.009227519, 0.003197749, 0.050215654, 0.0037740078, -0.003807509,
-0.020907104, 0.0070870104, 0.02154187, -0.034253534, 0.0051916186, 0.013102805,
-0.05013027, -0.011375772, -0.034398254, -0.012215769, 0.008936165, 0.013042285,
-0.0076083266, 0.020103235, 0.0015896345, 0.0013469604, 0.018086607, 0.0064340974,
-0.009976818, 0.0202195, -0.04537388, -0.0037930352, 0.00905707, 0.056803208,
-0.05901939, -0.0025249731, -0.026721176, 0.029771479, -0.009664647, 0.03187292,
0.048367433, 0.044916652, -0.030181149, 0.022398885, 0.017789925, 0.0104143135,
0.04470509, 0.043254152, 0.0026467436, -0.015192392, 0.00457335, -0.023504077,
-0.013324363, 0.018283904, -0.09116972, 0.023821378, -0.013131093, 0.0167326,
-0.032134093, 0.024058864, 0.01773737, 0.0029694063, 0.015400375, 0.021107076,
0.040144313, -0.018088255, -0.018179972, -0.0061101853, 0.0333514, -0.051985517,
0.0018389969, -0.042181797, 0.010265013, 0.030729504, -0.0021242746, -0.030638387,
-0.017514488, -0.043879565, -0.07925115, -0.009313998, -0.05037993, 0.0062779468,
-0.021197595, 0.009778766, -0.026689053, 0.008420625, -0.013590769, -0.025906648,
-0.019427065, -0.0040084007, -0.0051668566, -0.03791699, -0.030223792, 0.021360977,
0.05214987, -0.001195285, 0.056878682, 0.021255476, 0.030073736, -0.009004809,
-0.00044008184, 0.018010443, 0.05210606, -0.009380206, 0.014023647, -0.027397914,
0.0024216578, 0.028537884, 0.0041281483, 0.004196239, -0.042434014, -0.01951087,
0.010246227, 0.019259237, 0.044823032, -0.015578959, 0.055282183, -0.029959694,
-0.002444019, -0.030419223, -0.0015842384, 0.019480113, 0.016508475, 0.002307729,
0.027967995, -0.00045019708, 0.027565105, -0.027752383, 0.015513533, 0.0042188945,
-0.020022955, -0.020090383, -0.02833768, -0.008205656, 0.018416148, 0.013030916,
0.0043110526, 0.019754246, 0.02399727, -0.04217088, -0.029771129, -0.01836493,
0.030383382, -0.011573766, 0.0028558928, -0.031085022, 0.013471276, 0.015004199,
0.018147502, 0.009186545, -0.028823266, -0.01669298, -0.07682935, 0.0026605711,
-0.011341015, -0.052770227, 0.008115369, 0.044612564, 0.0425546, 0.032530054,
-0.056424312, -0.009478192, 0.04076847, -0.027977053, 0.026075717, 0.03094319,
0.0077947774, -0.010361255, 0.0379331, 0.022605725, 0.04959574, 0.01899966, 0.007897847,
0.023821533, -0.026519801, -0.0055518188, 0.028449029, 0.012021892, 0.0055719055,
0.0065661324, -0.021502847, -0.039742764, 0.045740023, -0.006839873, 0.0048939977,
0.00086617976, -0.021977283, 0.006932612, -0.011130043, -0.018758686, -0.013963125,
-0.008874906, -0.00250807, -0.03496783, 0.0045763687, -0.0135702705, -0.04224915,
0.057138246, -0.0035801234, -0.039638594, -0.0042460803, 0.003417834, -0.038236164,
-0.014725476, 0.010474428, -0.007192687, 0.012550738, 0.049059138, -0.02168995,
0.014943881, -0.02944271, 0.015512776, -0.006800395, 0.016161487, -0.05883302,
0.028153684, -0.008107896, -0.009384435, -0.013420283, 0.05274957, -0.023136541,
0.013694571, -0.018405624, -0.012455675, -0.007547505, 0.02333343, 0.013213835,
-0.044884287, 0.024978612, 0.007462058, 0.017579155, -0.04433104], index=0,
object='embedding'), Embedding(embedding=[-0.039806955, 0.018106725, -0.02714153,
-0.033833414, -0.010921911, -0.08787385, -0.048164885, 0.031297795, 0.032188654,
-0.004530212, -0.021804372, 0.024625015, -0.007042427, 0.009888381, 0.009500467,
-0.025422748, 0.011871752, -0.00732404, 0.018392378, -0.004384966, 0.0001918134,

-0.04004225, 0.010167631, 0.00806233, 0.03110558, 0.04625598, -0.012658598, -0.028965505,
-0.021548064, 0.056681987, -0.018512, 0.0221357, -0.03510766, -0.0047702184,
-0.016245274, -0.015167115, 0.023144336, -0.0524422, -0.031156112, -0.027200451,
0.009988742, -0.020565098, 0.014143164, -0.015100003, 0.033231463, -0.044389643,
-0.004136181, -0.025562724, -0.058502603, -0.014722696, -0.011400933, 0.014331452,
0.07658601, -0.014814307, 0.0031236135, -0.0033029092, 0.019509245, 0.017938983,
-0.024790436, -0.015248284, -0.011639638, 0.00501442, -0.0055467803, 0.008408418,
-0.024580697, 0.0754493, 0.019086119, -0.022098487, -0.008814615, -0.021437159,
-0.030012729, 0.0044206856, -0.036046855, -0.012681524, -0.024640366, -0.019347185,
-0.0052223173, 0.015137555, -0.048428133, 0.046845492, 0.065149285, 0.002647668,
0.0007207915, 0.02435114, 0.044494085, 0.0182996, -0.029606061, 0.037432224,
-0.033808608, 0.0022643944, 0.009645186, -0.022846015, 0.024837328, -0.047011603,
-0.035402983, -0.0046847723, -0.039559603, 0.0074818777, 0.02330356, 0.025716277,
-0.014304545, 0.023733906, -0.006246653, 0.008509405, -0.011293758, -0.025966475,
-0.01006334, 0.008181945, -0.01438242, 0.009540608, 0.06885038, 0.04916865, 0.0018908386,
0.003854289, -0.03604726, -0.0146035105, -0.022960719, 0.019865872, -0.0058088247,
0.010324922, -0.012516908, 0.015785128, 0.051367365, -0.025630152, -0.017501123,
-0.01962036, -0.00045100553, 0.044624988, -0.021584759, 0.03565252, -0.026582485,
0.027201204, -0.059125, 0.01317713, -0.012250775, -0.025474267, 0.027487954, 0.016126018,
0.022486532, -0.024704559, 0.037192255, 0.017422775, -0.019903207, -0.036879085,
0.011453935, -0.00235618, 0.013886989, -0.023742465, 0.014823697, -0.0013640152,
0.007340349, 0.025479008, 0.029555285, -0.017932586, -0.03817522, 0.0007982303,
-0.012220362, 0.00021680194, -0.036326665, -0.0062908116, 0.032608576, 0.012039895,
-0.020109933, 0.002184031, 0.014053472, -0.03662654, 0.0044956626, 0.07524842,
-0.0152201755, 0.01951614, -0.015556351, -0.013448956, 0.048647966, 0.0122912,
-0.016437223, 0.021207852, 0.030076804, -0.016976235, 0.0035818152, -0.0107744625,
-0.061285947, -0.021923166, -0.017293802, -0.08759088, -0.039984487, -0.00043348063,
0.010560561, -0.021260453, 0.010690529, -0.00076214195, -0.0052662776, -0.07078946,
-0.020219805, -0.0030916699, 0.0038979545, -0.0032112007, 0.003736981, 0.06447358,
0.054365616, 0.020362832, -0.011558257, 0.027246028, 0.030825194, -0.03900765,
-0.038409732, -0.018651864, -0.036674645, -0.04837994, 0.041603886, -0.03700395,
-0.003150507, -0.05434538, 0.013208883, -0.0087698875, -0.010625307, -0.018957065,
0.029379977, -0.019123154, 0.0066628377, 0.00064306625, -0.0045157797, -0.012306675,
-0.01965497, -0.03788315, 0.021389931, -0.0076195886, -0.02418394, 0.07190492,
-0.058766495, -0.04346301, -0.011424005, 0.08432645, 0.010501825, -0.023052415,
-0.010896898, -0.041609973, 0.034540582, -0.025880303, -0.0044130436, 0.022009451,
-0.0456012, -0.026753843, -0.00026467428, 0.03633925, 0.021008782, 0.021594657,
-0.033761226, -0.003918439, -0.0036088256, -0.030125177, 0.06014599, -0.0052329483,
-0.01676937, 0.027992217, 0.0020218696, 0.01681878, 0.031245762, 0.0074888594,
0.039434277, 0.0072179767, 0.023201365, -0.01086617, -0.011059847, 0.036365334,
0.012356821, -0.004479378, -0.029226903, 0.028809851, 0.004290806, 0.016059155,
-0.042729758, -0.015265739, -0.023315625, -0.0048190304, 0.036037758, -0.0037047488,
0.0058000875, 0.004225492, 0.030909065, -0.020124217, 0.012732193, -0.049268555,
-0.010288543, 0.032568026, 0.004937674, -0.019989511, -0.00376297, -0.021907778,
0.0032743274, 0.021680184, 0.0031568797, -0.030931003, 0.036583208, 0.008307239,
-0.035271116, 0.021804618, -0.03550608, -0.17422535, -0.0055584437, 0.016568558,
-0.042162247, 0.009234726, -0.021887384, 0.0044160085, -0.032132085, -0.02526897,
-0.03257029, -0.061538894, -0.032033436, -0.05803612, 0.019335447, -0.048609257,
-0.0123426365, 0.0047235885, -0.053109802, 0.0016916203, -0.037515607, -0.003102214,
-0.0014287904, 0.07933123, -0.03516418, -0.02318397, -0.016031312, -0.017235648,
-0.018351972, -0.04145283, -0.031296615, 0.027566928, -0.04918952, -0.02337211,
0.023168243, -0.03689156, -0.007357466, 0.015730577, -0.02637404, -0.0011485332,
0.013727407, -0.032218985, 0.008697318, -0.053799678, 0.020576116, -0.011187619,
-0.0036567596, -0.009576436, -0.009067609, -0.0059511457, 0.047911406, 0.0032387092,
0.032790948, 0.0044828868, 0.045609824, -0.044275556, 0.032871027, 0.019262772,
0.02776002, 0.025117772, 0.007967346, -0.011527599, -0.035411373, 0.02457575,
-0.025654942, -0.0071522677, 0.029986482, 0.023516832, 0.030535799, 0.045038912,
0.011724758, 0.0022815238, 0.019235903, -0.0058248453, -0.007677952, 0.044396043,

0.044792693, -0.027284242, -0.0034480433, -0.0031397224, -0.09265861, 0.02985109,
-0.03535556, 0.008404779, -0.03251267, 0.00832229, -0.01800912, -0.015365692,
-0.006667201, 0.005070796, 0.2429714, 0.034296818, -0.061447185, 0.011751609, 0.03563779,
-0.045496084, 0.033172783, 0.016725417, -0.013325212, -0.04710844, -0.01763304,
0.014165556, 0.020341875, -0.022885386, 0.01367139, 0.062560104, -0.011401764,
-0.01932027, 0.06305335, -0.010825633, -0.0067342655, 0.0017462067, 0.021622764,
0.040241074, -0.060613107, -0.04294679, -0.0018433991, 0.031839903, 0.007688521,
0.014064593, -0.021199869, -0.04950631, 0.03862481, 0.059081998, 0.019803561,
-0.018866148, 0.00047172944, 0.03317995, 0.041107073, -0.03278018, -0.05122741,
-0.010277512, 0.021450896, 0.0075447876, 0.0232653, -0.016950065, 0.021917164,
-0.014481437, -0.0060826093, -0.005029667, -0.01805828, 0.03047291, -0.04469266,
0.023791978, -0.029562002, -0.03152938, -0.015890347, -0.035565127, -0.037068065,
0.05428713, 0.053404607, 0.00011869666, 0.015649194, -0.026654938, -0.040659923,
0.04302314, 0.025084848, -0.056900717, 0.01611657, 0.061198022, -0.010592068,
-0.0117429765, -0.04247498, 0.010234422, 0.046964835, 0.049254086, 0.024806071,
0.043843374, -0.022250427, -0.0021817428, 0.0040312316, -0.021270124, -0.04342302,
0.009678571, 0.041352488, 0.055536613, 0.022192832, -0.007861454, 0.025131604,
-0.025723273, 0.0131876785, 0.026801726, -0.06978861, 0.00529091, -0.026728887,
-0.004692603, -0.029864728, -0.020519625, -0.0012926307, -0.031000795, 0.0042536897,
-0.046934154, -0.008083166, 0.044085227, 0.03775721, -0.029693477, -0.011002456,
0.05681608, -0.04142204, -0.01024627, 0.011809906, -0.03645116, -0.01609644, 0.020747473,
0.05280843, 0.021996047, 0.027929071, -0.013967343, 0.013427497, 0.024700826,
-0.025466755, -0.03191814, 0.018574893, 0.015539701, 0.0111877015, 0.0011065699,
0.020807378, -0.0022950503, -0.04080207, 0.022003127, 0.009040768, -0.029722787,
-0.016087228, 0.0021534252, 0.05132611, -0.0043423655, 0.03197384, 0.014612478,
-0.0075006653, 0.03795542, 0.020516044, -0.000753212, 0.0140319625, -0.036455624,
-0.0062314887, -0.020856056, 0.028727965, -0.040752828, -0.03098454, -0.03632487,
0.057478394, 0.02651742, -0.015345521, -0.03778333, -0.01893185, 0.043988228,
0.008374845, 0.03578321, -0.0336425, -0.0014386997, -0.01712994, -0.0057694726,
0.06346051, 0.079468384, -0.0037914685, 0.044671357, -0.021962574, 0.048231363,
0.026963266, 0.011023811, 0.046371337, -0.012278799, -0.013266753, 0.024954408,
0.03089167, -0.02914379, -0.0147609105, 0.034018278, -0.03385961, 0.02879893,
0.046497975, 0.019660087, 0.02409784, 0.020308346, -0.0067614387, 0.046343904,
-0.0005784594, -0.021660347, 0.054527197, -0.022563111, 0.028653856, 0.080075465,
-0.010942759, 0.008333316, 0.07266343, -0.0063063973, -0.015116326, -0.001291006,
0.007033947, -0.055138774, -0.013635839, -0.058863103, 0.018013697, -0.003587027,
0.03016848, 0.004340696, -0.02231035, -0.0021904027, -0.018317793, -0.034449887,
-0.044973347, -0.042687934, -0.012501527, 0.008154034, 0.008486107, -0.027704634,
-0.016208507, -0.017212354, 0.0154763, -0.057456914, 0.058836646, -0.010927607,
0.028967118, -0.023743823, 0.018120123, 0.008504358, -0.02019317, 0.027913263,
-0.033391576, -0.05849971, 0.0018396962, 0.0028468056, 0.012038477, 0.019820383,
0.033899166, -0.0034281653, 0.010240003, 0.007385701, -0.022350889, -0.030605549,
0.017475165, -0.010327821, 0.027654938, 0.05355119, -0.014949391, 0.0056421943,
-0.006985141, 0.06490143, -0.0019786155, 0.06540899, 0.00030945623, 0.06651776,
0.031434633, 0.029223382, 0.005312344, 0.02692162, 0.0080550695, 0.002547737,
-0.046926796, -0.0055953916, -0.027216688, 0.0192472, 0.019051598, 0.027252058,
-0.002625347, 0.011415791, 0.011703875, -0.042730447, -0.017118221, -0.04256073,
-0.01781555, 0.025160633, -0.009056251, -0.0028699443, 6.267121e-05, 0.0029575485,
-0.03786012, 0.032860763, -0.03924916, 0.018752553, -0.0035671783, 0.037145987,
-0.07339691, -0.026845668, -0.024349527, -0.010001643, -0.008324767, 0.02444552,
-0.0069263037, -0.039972678, -0.011900961, 0.0012417778, 0.017389448, 0.020070942,
0.02678553, 0.015119207, -0.052110363, 0.021610133, -0.015621386, 0.017013198,
-0.016733378, 0.050118178, -0.036132984, -0.028919403, -0.029918974, 0.000518692,
0.013747376, -0.024373489, -0.0054474077, 0.019377526, 0.024695145, -0.004618123,
0.041052736, 0.026499253, 0.04656181, -0.029189281, 0.00806193, -0.011962786,
-0.048480418, 0.0014013203, 0.003932788, 0.016817138, 0.011343118, -0.022587297,
-0.0015305758, 0.000102333805, 0.02695947, 0.011637096, -0.049498096, 0.018794134,
-0.0237115, 0.029978462, -0.015226518, -0.012754388, 0.050404325, -0.035898246,

-0.03161909, -0.0103445435, 0.049694147, -0.006435172, -0.02847327, 0.08106319,
0.0809235, -0.009108299, 0.028060187, -0.038263198, 0.0008455384, 0.030499715,
-0.006457368, 0.004814054, 0.016831778, 0.010108291, -0.03908352, -0.043085005,
0.018404072, 0.0045124697, 0.0065787556, -0.05353358, -0.01287885, 0.010193927,
-0.027740126, 0.06827365, -0.034389623, 0.028805405, -0.0023430856, 0.009260239,
0.005530282, -0.025005303, -0.043685574, 0.012968424, 0.012462829, 0.04922187,
0.011033892, -0.020876283, -0.015727568, -0.006340249, -0.04240463, -0.0387011,
0.031318422, -0.010669236, -0.038257945, 0.025644235, -0.036942434, -0.040654384,
0.026557662, -0.038425166, 0.010972449, -0.0026865592, 0.04297902, 0.013841087,
-0.03508849, 0.022146288, -0.059955824, -0.011984481, -0.14203429, 0.033375036,
0.025635537, -0.014800292, 0.021314079, -0.004364666, -0.04288085, 0.03326454,
-0.017278133, -0.015381165, 0.039260007, 0.030663054, -0.009398082, 0.0016157854,
0.0058923406, 0.038551237, 0.029295184, 0.026982203, 0.0046961685, 0.020004513,
-0.0002823051, 0.029769791, -0.06499312, 0.0027809944, 0.033780456, -0.054227453,
0.015405423, 0.014887773, -0.06223483, 0.019264648, -0.02459092, -0.028076837,
-0.0007747023, 0.0050467625, 0.0054090386, 0.044855643, 0.014965291, 0.0039615836,
0.0364642, -0.0151543515, -0.010991144, 0.014066373, -0.020417051, -0.015295324,
0.0032331492, 0.029506348, -0.05176175, -0.012887155, -0.049193565, -0.025174212,
-0.0057166666, -0.00039162458, 0.00026217767, 0.07197446, -0.008801158, 0.023327848,
-0.013825299, 0.019054411, 0.026037995, 0.029275186, 0.0014251247, 0.0050785993,
0.014754311, -0.02803208, -0.020658525, 0.009343213, -0.038446713, -0.0021920993,
0.030721895, 0.0060682576, -0.0029885068, 0.0290252, -0.006677525, -0.007283009,
0.009740268, 0.021990651, -0.024688762, -0.04346818, -0.029574063, -0.007917782,
0.05176251, -0.03194417, -0.010803628, -0.03710345, 0.01654293, 0.035606004,
0.00021680047, -0.01902328, 0.017944654, -0.014634677, -0.08091542, -0.0005132855,
-0.0646727, -0.044102196, -0.05027943, 0.023779303, 0.0013730207, -0.00033895578,
-0.021345852, -0.01240465, -0.025230281, 0.02300386, 0.0059149317, -0.006499429,
-0.04890679, 0.049107686, 0.041578628, -0.021378888, 0.023876995, -0.012183695,
0.013455828, -0.005322552, -0.015810125, 0.025522828, -0.010038473, -0.006050725,
0.019967675, -0.003245995, -0.025224932, 0.0439206, -0.0012436587, 0.031723216,
-0.02448605, -0.015330694, 0.026584158, -0.0051061427, 0.066581406, -0.0038770763,
0.020848919, -0.028565202, -0.046525814, -0.058668178, 0.034280267, -0.0088918265,
-0.026706906, 0.018157527, 0.0033295872, 0.005130095, 0.019549852, -0.008804092,
0.0069478713, 0.009930967, -0.032567203, -0.042641874, -0.025361327, -0.02444165,
0.060229223, -0.00029296556, -0.032582972, 0.008229873, 0.00873001, -0.00815309,
-0.016171323, -0.029477486, 0.048059177, 1.3053809e-06, 0.0039567174, -0.040332206,
0.00093539315, 0.055955756, 0.008039591, -0.098208874, -0.024455974, 0.012274973,
-0.026007501, -0.0077049667, 0.0053886627, -0.019571463, -0.010535064, 0.03599966,
0.034453247, 0.042971645, -0.047814555, -0.006743227, 0.033782616, -0.03835373,
0.02963017, 0.057553414, 0.012134137, -0.01209421, 0.008088897, 0.029915666, 0.024700345,
0.03985199, 0.018612375, -0.007387687, -0.028683772, 0.015096653, 0.020524723,
-0.0025782587, 0.00621819, -0.018208023, -0.044911884, -0.006225681, 0.011868896,
-0.00073360204, -0.00060059247, 0.025633158, 0.010927056, 0.028156856, 0.013699513,
-0.01723093, -0.007418714, 0.001602008, -0.006724474, -0.042085167, -0.044802777,
-0.0005950661, -0.024677, 0.0152367465, -0.035699733, -0.021288736, -0.015362072,
0.04180836, -0.033493835, -0.00041006325, -0.030007878, -0.0003501069, 0.025993193,
0.030151015, -0.014684397, 0.036694344, -0.049206484, 0.05902542, -0.0063680233,
0.027811198, -0.054977603, 0.04245663, -0.044224508, 0.0074903113, 0.019687254,
-0.045596678, -0.01462281, -0.0051411195, 0.026682809, -0.0054900898, 0.016229948,
0.025660323, 0.031181056, -0.052291203, 0.03282995, -0.014796475, -0.0059239697,
0.012493835], index=1, object='embedding')], model='bge-m3', object='list',
usage=Usage(prompt_tokens=14, total_tokens=14))

```
print(len(response.data))
print(len(response.data[0].embedding))
print(len(response.data[1].embedding))
```

```
2
1024
1024
```

至此，`Ollama` 启动并使用 `Embedding` 模型的相关使用方法就已经介绍完毕，并没有很复杂的流程，同时其参数也比较容易理解。

2.2.4 动态 Redis 缓存检索实现完整流程

- Step 1 ： 基础设置和依赖导入

```
import numpy as np
from typing import List
import redis
import requests
import hashlib
```

- Step 2：实现基础的 Embedding 功能

这里我们实现一个文本 `Embedding` 类，通过 `Ollama` 的 `/api/embed` 的 `REST API` 接口，将输入文本转换为向量。代码如下：

```
class OllamaEmbedding:
    def __init__(self, base_url: str = "http://localhost:11434", model: str = "bge-m3"):
        """初始化 Ollama Embedding

        Args:
            base_url: Ollama 服务地址
            model: 使用的模型名称
        """
        self.base_url = base_url.rstrip('/')
        self.model = model
        self.api_url = f"{self.base_url}/api/embed"

    def embed_query(self, text: str) -> List[float]:
        """单个文本转换为向量

        Args:
            text: 输入文本

        Returns:
            List[float]: 向量表示
        """
        payload = {
            "model": self.model,
            "input": text
        }

        try:
            response = requests.post(self.api_url, json=payload)
```

```
            response.raise_for_status()  # 检查请求是否成功
            result = response.json()
            return result["embeddings"][0]  # 返回第一个（也是唯一的）embedding
        except Exception as e:
            print(f"Error generating embedding: {e}")
            return []

    def embed_documents(self, texts: List[str]) -> List[List[float]]:
        """多个文本转换为向量

        Args:
            texts: 输入文本列表

        Returns:
            List[List[float]]: 向量表示列表
        """
        payload = {
            "model": self.model,
            "input": texts
        }

        try:
            response = requests.post(self.api_url, json=payload)
            response.raise_for_status()
            result = response.json()
            return result["embeddings"]   # 批量的话这里直接返回列表
        except Exception as e:
            print(f"Error generating embeddings: {e}")
            return []
```

接下来进行测试，其中需要重点关注的是：要根据自己的实际情况，修改 `base_url` 和 `model` 参数。

```
# 初始化 embedding 模型
embedding = OllamaEmbedding(
    base_url="http://192.168.110.131:11434",  # 根据自己的实际情况调整 endpoint
    model="bge-m3"         # 根据自己的实际情况调整模型名称
)

# 测试单个文本
test_text = "请问如何理解大模型技术？"
single_vector = embedding.embed_query(test_text)
print("\n单个文本测试:")
print(f"输入文本: {test_text}")
print(f"向量维度: {len(single_vector)}")
print(f"向量示例: {single_vector[:5]}...")  # 只显示前5个数

# 测试多个文本
test_texts = [
    "什么是大模型？",
    "什么是深度学习？",
    "计算机是如何模拟人类思维的？",
    "你好，请你详细的介绍一下你自己"
]
vectors = embedding.embed_documents(test_texts)
print("\n多个文本测试:")
print(f"输入文本数量: {len(test_texts)}")
print(f"生成向量数量: {len(vectors)}")
```

```
print(f"每个向量维度: {len(vectors[0])}")
print(f"第一个向量示例: {vectors[0][:5]}...")  # 只显示第一个向量的前5个数
```

```
单个文本测试:
输入文本: 请问如何理解大模型技术?
向量维度: 1024
向量示例: [-0.041555773, -0.04320696, -0.04493023, -0.010034573, 0.004685277]...

多个文本测试:
输入文本数量: 4
生成向量数量: 4
每个向量维度: 1024
第一个向量示例: [-0.029385682, -0.038512684, -0.048128907, -0.019845983, 0.0013321947]...
```

- Step 3：实现 Redis 连接和基本操作

```
class RedisSemanticCache:
    def __init__(self, embedding_model, prefix: str = "cache:",
                 host='192.168.110.131', password=None,port=6379, db=0):
        """初始化 Redis 缓存

        Args:
            embedding_model: 嵌入模型
            prefix: 键前缀
            host: Redis 主机地址
            port: Redis 端口
            db: Redis 数据库号
        """
        self.redis_client = redis.Redis(
            host=host,
            port=port,
            db=db,
            password=password
        )
        self.embedding = embedding_model
        self.prefix = prefix

    def _create_key(self, text: str) -> str:
        """使用 MD5 创建确定性的 key

        Args:
            text: 输入文本

        Returns:
            str: 格式为 "prefix:md5hash" 的 key
        """
        # 将文本转换为 MD5 哈希
        # 1. hash() 函数可以将任意长度的文本转换为一个固定长度的整数
        # 2. 相同的文本会生成相同的哈希值，不同的文本很大概率会生成不同的哈希值
        # 3. 使用哈希值作为 key 比直接使用原文本更节省空间
        # 4. Redis key 的查找会更快
        # 5. 避免原始文本中的特殊字符可能导致的问题
        # 6. 防止文本过长超出 Redis key 的长度限制
        md5_hash = hashlib.md5(text.encode('utf-8')).hexdigest()
        return f"{self.prefix}{md5_hash}"  # 前缀（prefix）可以用于命名空间隔离
```

```python
    def add_text(self, text: str, metadata: dict = None):
        # 生成embedding
        vector = self.embedding.embed_query(text)
        key = self._create_key(text)

        # 存储数据
        data = {
            'text': text,  # 存储原始文本
            'vector': str(vector),  # 存储向量
            'metadata': str(metadata or {})  # 存储元数据
        }

        # hset: 用于将多个字段-值对存储到哈希表 key 中
        self.redis_client.hset(key, mapping=data)

    def get_text(self, key: str) -> dict:
        # hgetall: 获取哈希表中所有字段-值对
        return self.redis_client.hgetall(key)

    def list_all_keys(self):
        """列出所有缓存的 keys"""
        pattern = f"{self.prefix}*"
        return [key.decode() for key in self.redis_client.keys(pattern)]

    def get_by_text(self, text: str):
        """通过原始文本查询"""
        key = self._create_key(text)
        result = self.redis_client.hgetall(key)
        if result:
            return {k.decode(): v.decode() for k, v in result.items()}
        return None

    def get_all_texts(self):
        """获取所有存储的文本数据"""
        all_data = []
        for key in self.list_all_keys():
            data = self.redis_client.hgetall(key)
            if data:
                all_data.append({k.decode(): v.decode() for k, v in data.items()})
        return all_data

    def clear_all_cache(self):
        """清除所有缓存数据"""
        pattern = f"{self.prefix}*"
        keys = self.redis_client.keys(pattern)
        if keys:
            self.redis_client.delete(*keys)
            print(f"已清除 {len(keys)} 条缓存记录")
        else:
            print("缓存为空")
```

接下来进行 `Embedding` 服务和 `Redis` 客户端的实例化，代码如下：

```python
# 初始化 embedding 模型
embedding = OllamaEmbedding(
    base_url="http://192.168.110.131:11434",  # 根据自己的实际情况调整 endpoint
    model="bge-m3"          # 根据自己的实际情况调整模型名称
```

```
)

# 初始化 Redis 缓存
redis_cache = RedisSemanticCache(
    host='192.168.110.131',  # 根据自己的实际情况调整 host
    password='g1601522830',  # 根据自己的实际情况调整 password
    port=6379,  # 根据自己的实际情况调整 port
    db=0,  # 根据自己的实际情况调整 db
    embedding_model=embedding
)
```

进行功能测试：

```
# 测试添加文本
test_text = "大模型是什么？"
redis_cache.add_text(test_text, metadata={"source": "test"})
```

```
# 测试其他功能

# 查看所有存储的数据
print("所有缓存的 keys:")
print(redis_cache.list_all_keys())

print("\n通过文本查询:")
result = redis_cache.get_by_text("哈哈哈，我是居居。")
print(result)

print("\n所有存储的数据:")
all_data = redis_cache.get_all_texts()
for item in all_data:
    print("\n---")
    print(f"文本: {item['text']}")
    print(f"元数据: {item['metadata']}")
    print(f"向量前5个值: {eval(item['vector'])[:5]}")  # 只显示向量的前5个值
```

- Step 4：实现相似度搜索

```
def cosine_similarity(vec1: List[float], vec2: List[float]) -> float:
    vec1 = np.array(vec1)
    vec2 = np.array(vec2)
    return np.dot(vec1, vec2) / (np.linalg.norm(vec1) * np.linalg.norm(vec2))

class SemanticCache(RedisSemanticCache):
    def similarity_search(self, query: str, threshold: float = 0.8) -> List[dict]:
        # 获取查询文本的embedding
        query_vector = self.embedding.embed_query(query)

        # 获取所有缓存的数据
        results = []
        for key in self.redis_client.keys(f"{self.prefix}*"):
            cached_data = self.get_text(key)
            if not cached_data:
                continue

            # 计算相似度
            cached_vector = eval(cached_data[b'vector'])  # 将字符串转回列表
```

```
            similarity = cosine_similarity(query_vector, cached_vector)

            if similarity >= threshold:
                results.append({
                    'text': cached_data[b'text'].decode(),
                    'similarity': similarity,
                    'metadata': eval(cached_data[b'metadata'].decode())
                })

        return sorted(results, key=lambda x: x['similarity'], reverse=True)
```

进行语义相似度功能测试。代码如下：

```
# 测试
semantic_cache = SemanticCache(
    host='192.168.110.131',  # 根据自己的实际情况调整 host
    password='g1601522830',  # 根据自己的实际情况调整 password
    port=6379,  # 根据自己的实际情况调整 port
    db=0,  # 根据自己的实际情况调整 db
    embedding_model=embedding
)


# 测试相似度搜索
query = "什么是大模型？"
results = semantic_cache.similarity_search(query, threshold=0.95)
print("\n搜索结果:")
for result in results:
    print(f"文本: {result['text']}")
    print(f"相似度: {result['similarity']:.3f}")
    print(f"元数据: {result['metadata']}\n")
```

```
搜索结果:
文本: 大模型是什么？
相似度: 0.990
元数据: {'source': 'test'}
```

```
all_data = redis_cache.get_all_texts()
for item in all_data:
    print("\n---")
    print(f"文本: {item['text']}")
    print(f"元数据: {item['metadata']}")
    print(f"向量前5个值: {eval(item['vector'])[:5]}")  # 只显示向量的前5个值
```

通过上述代码，我们可以看到，当查询文本为`"大模型是什么？"`时，搜索结果为`"请问什么是大模型？"`，并且相似度为`0.95`，所以我们现在实现了对用户输入问题的基于语义的缓存搜索。

现在回归到实际的业务场景中，`Prompt Cache`的应用逻辑应该是这样的：

1. 当用户输入问题时，先根据用户输入的问题，去缓存中搜索，如果缓存命中（即相似度大于某个阈值），则直接返回缓存中的答案。
2. 如果缓存没有命中，则根据用户输入的问题，执行大模型服务调用（API）逻辑，并在得到问答的结果后，将结果存入缓存中。

3. 以此类推，当用户再次输入相同的问题时，由于缓存中已经存在了该问题的答案，所以可以直接从缓存中返回答案，从而提高回答的效率。

现在我们就可以根据上述的逻辑，来进行优化，思路如下：

- 实现 OllamaChat 类 构建大模型服务调用（API）逻辑

```
class OllamaChat:
    def __init__(self,
                 base_url: str = "http://192.168.110.131:11434",
                 model: str = "deepseek-r1:1.5b",
                 redis_cache = None,
                 embedding = None):
        self.base_url = base_url.rstrip('/')
        self.model = model
        self.chat_url = f"{self.base_url}/api/chat"
        self.redis_cache = redis_cache
        self.embedding = embedding

    def get_answer_with_cache(self, question: str, similarity_threshold: float = 0.90) ->
dict:
        """获取答案（优先从缓存中获取）

        Args:
            question: 用户问题
            similarity_threshold: 相似度阈值，超过这个值就使用缓存的答案
        """
        # 获取问题的向量
        query_vector = self.embedding.embed_query(question)

        # 搜索最相似的问题
        max_similarity = 0
        cached_answer = None

        # 遍历所有缓存
        for key in self.redis_cache.list_all_keys():
            data = self.redis_cache.redis_client.hgetall(key)
            if not data:
                continue

            # 获取缓存的向量并计算相似度
            cached_vector = eval(data[b'vector'].decode())
            similarity = self._calculate_similarity(query_vector, cached_vector)

            # 更新最大相似度和对应的答案
            if similarity > max_similarity:
                max_similarity = similarity
                if similarity >= similarity_threshold:
                    metadata = eval(data[b'metadata'].decode())
                    cached_answer = {
                        'answer': metadata['answer'],
                        'source': 'cache',
                        'similarity': similarity,
                        'original_question': data[b'text'].decode()
                    }

        # 如果找到足够相似的缓存答案，直接返回
        if cached_answer:
```

```
            return cached_answer

        # 如果没有找到相似的缓存，调用模型获取答案
        answer = self._get_llm_answer(question)

        # 将新问答对存入缓存
        self.redis_cache.add_text(
            text=question,
            metadata={
                'answer': answer,
                'source': 'llm'
            }
        )

        return {
            'answer': answer,
            'source': 'llm',
            'similarity': 0,
            'original_question': None
        }

    def _calculate_similarity(self, vec1, vec2) -> float:
        """计算两个向量的余弦相似度"""
        vec1 = np.array(vec1)
        vec2 = np.array(vec2)
        return np.dot(vec1, vec2) / (np.linalg.norm(vec1) * np.linalg.norm(vec2))


    def _get_llm_answer(self, question: str) -> str:
        """从 LLM 获取答案"""
        payload = {
            "model": self.model,
            "messages": [
                {
                    "role": "user",
                    "content": question
                }
            ],
            "stream": False,
            "options": {
                "temperature": 1.3,
            }
        }

        try:
            response = requests.post(self.chat_url, json=payload)
            response.raise_for_status()
            result = response.json()
            return result['message']['content']
        except Exception as e:
            print(f"Error getting answer: {e}")
            return ""
```

实例化 `Embedding` 模型和 `Redis` 客户端实例。代码如下：

```
# 初始化 embedding 模型
embedding = OllamaEmbedding(
```

```
    base_url="http://192.168.110.131:11434",  # 根据自己的实际情况调整 endpoint
    model="bge-m3"         # 根据自己的实际情况调整模型名称
)

# 初始化 Redis 缓存
redis_cache = RedisSemanticCache(
    host='192.168.110.131',  # 根据自己的实际情况调整 host
    password='g1601522830',  # 根据自己的实际情况调整 password
    port=6379,  # 根据自己的实际情况调整 port
    db=0,  # 根据自己的实际情况调整 db
    embedding_model=embedding
)
```

先清空一下`Redis`的缓存：

```
# 使用示例
redis_cache.clear_all_cache()
```

```
已清除 6 条缓存记录
```

```
print("\n所有存储的数据:")
all_data = redis_cache.get_all_texts()
for item in all_data:
    print("\n---")
    print(f"文本: {item['text']}")
    print(f"元数据: {item['metadata']}")
    print(f"向量前5个值: {eval(item['vector'])[:5]}")  # 只显示向量的前5个值
```

```
所有存储的数据:
```

这里的返回结果就变成了空，说明现在`Redis`中是没有任何数据的。接下来我们实例化对话模型进行测试。代码如下：

```
# 初始化聊天模型
chat_model = OllamaChat(
    redis_cache=redis_cache,
    embedding=embedding
)

# 测试问答
question = "请介绍一下人工智能哈"
result = chat_model.get_answer_with_cache(question)
print(f"问题: {question}")
print(f"来源: {result['source']}")
print(f"答案: {result['answer']}")
```

```
问题: 请介绍一下人工智能哈
来源: llm
答案: <think>
嗯，用户问我关于“人工智能哈”是什么。首先，“人工智能”（Artificial Intelligence, AI）这个概念大家应该很熟悉吧？不过具体叫“AI Hashboard”或者“人工智能哈”听起来有些陌生。
```

我需要弄清楚这句话可能是指什么。也许是在某个特定平台或项目中的名称。我查了一下，发现这是一个Python库，名为alpinehashboard，用于开发和测试AI模型。这名字里包含了人工智能，所以用户可能是想知道这个工具的作用或者使用方法。

接下来，我要解释一下这个库的主要功能。它的主要用途是帮助开发者快速搭建和运行AI模型的环境。它提供了一个可视化的界面，方便不同层次的开发者操作。比如，AI开发工程师可以在IDE中直接编写代码，并用AI算法进行训练和测试，这样节省了时间并提高效率。

另外，用户可能还关心如何安装和使用这个库，我应该提醒他们如何从官方文档或者示例代码入手。如果他们在本地环境中运行，可以使用pip的命令；如果有网络可用的话，则可以用本地服务器下载并解压仓库中的包。

最后，我会总结一下AI Hashboard的核心优势：简洁高效、快速迭代、资源便捷。这些因素使其成为开发者和研究人员的首选工具。
</think>

您的问题涉及到“人工智能哈”这个表述，可能是一个特定平台或工具的名称，而非传统的“人工智能”的缩写（如AAI）。以下是对“人工智能哈”这一表述进行了解释：

1. **人工智能（Artificial Intelligence, AI）**
“人工智能”通常指的是一种由软件实现的、能够模拟人 intelligence 的高级系统。AI 系统能够理解、学习、执行并执行各种行动，例如聊天、翻译、推理、识别图像、规划路径等。

2. **人工智能哈**
“人工智能哈”可能是某种特定工具或平台（如“人工智能哈”是一个简称为 “hashboard”的平台，用于开发和训练 AI 模型）的名称。如果你是在寻找具体的代码实现或其他技术信息，请提供更多信息以便进一步解答。

例如：
- 如果是指一个 Python 库用于快速搭建和运行 AI 模型的工具，“alpinehashboard”就是这样的库。
- 如果是某个项目或平台，它专注于展示人工智能应用的实际案例和资源。

如果您能提供更多背景信息，我将为您提供更详细的解答。

```python
# 初始化聊天模型
chat_model = OllamaChat(
    redis_cache=redis_cache,
    embedding=embedding
)

# 测试问答
question = "请介绍一下人工智能"
result = chat_model.get_answer_with_cache(question)
print(f"问题: {question}")
print(f"来源: {result['source']}")
print(f"答案: {result['answer']}")
```

```
问题: 请介绍一下人工智能
来源: cache
答案: <think>
嗯，用户问我关于“人工智能哈”是什么。首先，“人工智能”（Artificial Intelligence, AI）这个概念大家应该很熟悉吧？不过具体叫“AI Hashboard”或者“人工智能哈”听起来有些陌生。
```

我需要弄清楚这句话可能是指什么。也许是在某个特定平台或项目中的名称。我查了一下，发现这是一个Python库，名为alpinehashboard，用于开发和测试AI模型。这名字里包含了人工智能，所以用户可能是想知道这个工具的作用或者使用方法。

接下来，我要解释一下这个库的主要功能。它的主要用途是帮助开发者快速搭建和运行AI模型的环境。它提供了一个可视化的界面，方便不同层次的开发者操作。比如，AI开发工程师可以在IDE中直接编写代码，并用AI算法进行训练和测试，这样节省了时间并提高效率。

另外，用户可能还关心如何安装和使用这个库，我应该提醒他们如何从官方文档或者示例代码入手。如果他们在本地环境中运行，可以使用pip的命令；如果有网络可用的话，则可以用本地服务器下载并解压仓库中的包。

最后，我会总结一下AI Hashboard的核心优势：简洁高效、快速迭代、资源便捷。这些因素使其成为开发者和研究人员的首选工具。
</think>

您的问题涉及到“人工智能哈”这个表述，可能是一个特定平台或工具的名称，而非传统的“人工智能”的缩写（如AAI）。以下是对“人工智能哈”这一表述进行了解释：

1. **人工智能（Artificial Intelligence, AI）**
“人工智能”通常指的是一种由软件实现的、能够模拟人 intelligence 的高级系统。AI 系统能够理解、学习、执行并执行各种行动，例如聊天、翻译、推理、识别图像、规划路径等。

2. **人工智能哈**
“人工智能哈”可能是某种特定工具或平台（如“人工智能哈”是一个简称为 “hashboard”的平台，用于开发和训练 AI 模型）的名称。如果你是在寻找具体的代码实现或其他技术信息，请提供更多信息以便进一步解答。

例如：
- 如果是指一个 Python 库用于快速搭建和运行 AI 模型的工具，“alpinehashboard”就是这样的库。
- 如果是某个项目或平台，它专注于展示人工智能应用的实际案例和资源。

如果您能提供更多背景信息，我将为您提供更详细的解答。

```python
# 初始化聊天模型
chat_model = OllamaChat(
    redis_cache=redis_cache,
    embedding=embedding
)

# 测试问答
question = "如何理解人工智能？"
result = chat_model.get_answer_with_cache(question)
print(f"问题: {question}")
print(f"来源: {result['source']}")
print(f"答案: {result['answer']}")
```

```
问题: 如何理解人工智能？
来源: llm
答案: <think>

</think>

人工智能（AI）指的是能够通过学习和合成人类智能的系统。它通过复杂的算法和模型，如机器学习、深度学习等，从数据中自动提取规律，并做出决策或预测。人工智能在多个领域得到了广泛应用，如自然语言处理、计算机视觉、推荐系统等领域，显著提高了生产力和社会效率。
```

```python
# 初始化聊天模型
chat_model = OllamaChat(
    redis_cache=redis_cache,
    embedding=embedding
)

# 测试问答
question = "什么是人工智能？"
result = chat_model.get_answer_with_cache(question)
print(f"问题: {question}")
print(f"来源: {result['source']}")
print(f"答案: {result['answer']}")
```

```
问题: 什么是人工智能？
来源: cache
答案: <think>

</think>

人工智能（AI）指的是能够通过学习和合成人类智能的系统。它通过复杂的算法和模型，如机器学习、深度学习等，从数据中自动提取规律，并做出决策或预测。人工智能在多个领域得到了广泛应用，如自然语言处理、计算机视觉、推荐系统等领域，显著提高了生产力和社会效率。
```

```python
# 测试问答
question = "哈哈哈"
result = chat_model.get_answer_with_cache(question)
print(f"问题: {question}")
print(f"答案: {result['answer']}")
print(f"来源: {result['source']}")
```

```
问题: 哈哈哈
答案: <think>

</think>

哈哈，谢谢你的逗笑！ ^w^ 有什么想聊的吗？
来源: llm
```

```python
# 测试问答
question = "哈哈哈"
result = chat_model.get_answer_with_cache(question)
print(f"问题: {question}")
print(f"答案: {result['answer']}")
print(f"来源: {result['source']}")
```

```
问题: 哈哈哈
答案: <think>

</think>

哈哈，谢谢你的逗笑！ ^w^ 有什么想聊的吗？
来源: cache
```

```python
print("\n所有存储的数据:")
all_data = redis_cache.get_all_texts()
for item in all_data:
    print("\n---")
    print(f"文本: {item['text']}")
    print(f"元数据: {item['metadata']}")
    print(f"向量前5个值: {eval(item['vector'])[:5]}")  # 只显示向量的前5个值
```

```
所有存储的数据:

---
文本: 哈哈哈
元数据: {'answer': '<think>\n\n</think>\n\n哈哈，谢谢你的逗笑！ ^w^ 有什么想聊的吗？', 'source': 'llm'}
向量前5个值: [-0.016039409, 0.016994113, -0.024663014, -0.0070588915, -0.015972273]

---
文本: 请介绍一下人工智能哈
元数据: {'answer': '<think>\n嗯，用户问我关于“人工智能哈”是什么。首先，“人工智能”（Artificial Intelligence, AI）这个概念大家应该很熟悉吧？不过具体叫“AI Hashboard”或者“人工智能哈”听起来有些陌生。\n\n我需要弄清楚这句话可能是指什么。也许是在某个特定平台或项目中的名称。我查了一下，发现这是一个Python库，名为alpinehashboard，用于开发和测试AI模型。这名字里包含了人工智能，所以用户可能是想知道这个工具的作用或者使用方法。\n\n接下来，我要解释一下这个库的主要功能。它的主要用途是帮助开发者快速搭建和运行AI模型的环境。它提供了一个可视化的界面，方便不同层次的开发者操作。比如，AI开发工程师可以在IDE中直接编写代码，并用AI算法进行训练和测试，这样节省了时间并提高效率。\n\n另外，用户可能还关心如何安装和使用这个库，我应该提醒他们如何从官方文档或者示例代码入手。如果他们在本地环境中运行，可以使用pip的命令；如果有网络可用的话，则可以用本地服务器下载并解压仓库中的包。\n\n最后，我会总结一下AI Hashboard的核心优势：简洁高效、快速迭代、资源便捷。这些因素使其成为开发者和研究人员的首选工具。\n</think>\n\n您的问题涉及到“人工智能哈”这个表述，可能是一个特定平台或工具的名称，而非传统的“人工智能”的缩写（如AAI）。以下是对“人工智能哈”这一表述进行了解释：\n\n### 1. **人工智能（Artificial Intelligence, AI）**\n“人工智能”通常指的是一种由软件实现的、能够模拟人 intelligence 的高级系统。AI 系统能够理解、学习、执行并执行各种行动，例如聊天、翻译、推理、识别图像、规划路径等。\n\n### 2. **人工智能哈**\n“人工智能哈”可能是某种特定工具或平台（如“人工智能哈”是一个简称为 “hashboard”的平台，用于开发和训练 AI 模型）的名称。如果你是在寻找具体的代码实现或其他技术信息，请提供更多信息以便进一步解答。\n\n例如：\n- 如果是指一个 Python 库用于快速搭建和运行 AI 模型的工具，“alpinehashboard”就是这样的库。\n- 如果是某个项目或平台，它专注于展示人工智能应用的实际案例和资源。\n\n如果您能提供更多背景信息，我将为您提供更详细的解答。', 'source': 'llm'}
向量前5个值: [-0.014650457, 0.0015913498, -0.043915708, 0.00037890443, -0.02943015]

---
文本: 如何理解人工智能？
元数据: {'answer': '<think>\n\n</think>\n\n人工智能（AI）指的是能够通过学习和合成人类智能的系统。它通过复杂的算法和模型，如机器学习、深度学习等，从数据中自动提取规律，并做出决策或预测。人工智能在多个领域得到了广泛应用，如自然语言处理、计算机视觉、推荐系统等领域，显著提高了生产力和社会效率。', 'source': 'llm'}
向量前5个值: [-0.016013233, -0.012850036, -0.017854566, -0.0044424483, -0.016207218]
```

现在，我们就完成了基于语义的`Redis`的动态缓存检索的完整流程。

回归到`AssistGen`的项目场景中，则需要考虑因素要更多一些，这主要包括：

1. 添加用户ID到缓存前缀,实现用户隔离;
2. 添加元数据存储(访问时间、访问次数等);
3. 实现自动清理机制;
4. 定期检查缓存大小,当超过最大限制时,删除最久未访问的数据

`Redis` 缓存功能服务接口核心文件位于：`/llm_backend/app/services/redis_semantic_cache.py` 文件。