

项目

AMD_AI_assistant [AMD_AI_Project - 副本]

.cache

config

data

docs

examples

logs

models

qianwen3

qianwen0.6

src

document_qa

qa

training

web

.cache

__init__.py

app.py

config.py

conversation_memory.pkl

__init__.py

.gitignore

LICENSE

README.md

requirements.txt

外部库

临时文件和控制台

train_manager.py

app.py

web\config.py

README.md

requirements.txt

不满足软件包要求 'torch-directml>=1.13.0'

安装要求 忽略要求

1

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

gradio>=4.0.0

torch>=2.0.0

torch-directml>=1.13.0

transformers>=4.35.0

accelerate>=0.24.0

sentencepiece>=0.1.99

protobuf>=3.20.0

pydantic>=2.0.0

python-dotenv>=1.0.0

numpy>=1.24.0

scipy>=1.10.0

requests>=2.31.0

tqdm>=4.65.0

pillow>=10.0.0

markdown>=3.5.0

2、安装虚拟环境

3、将干问的FP16模型下载到这个文件夹, 就是说点开这个文件夹就是文件, 模型等

1、需要python 3.10 其他版本可能不支持AMD (其实是因为某些包更新没跟上python更新)

AMD_AI_assistant > models > qianwen3 > qianwen0.6

13:13 (31 字符, 2 行 换行符) CRLF UTF-8 4 个空格 Python 3.10 (AMD_AI_Project)

项目

AMD_AI_assistant [AMD_AI_Project - 副本]

.cache

config

data

docs

examples

logs

models

qianwen3

qianwen0.6

src

document_qa

qa

training

web

.cache

__init__.py

app.py

config.py

conversation_memory.pkl

__init__.py

.gitignore

LICENSE

README.md

requirements.txt

外部库

临时文件和控制台

train_manager.py

app.py

web\config.py

README.md

requirements.txt

不满足软件包要求 'torch-directml>=1.13.0'

安装要求

忽略要求

1 9

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

"recommended_vram": 8,

"cache_dir": str(CACHE_DIR / "huggingface"),

"type": "huggingface"

}

"Qwen2.5-3B": {

"name": "Qwen/Qwen2.5-3B-Instruct",

"description": "能力强, 回答详细",

"size_gb": 6.0,

"recommended_vram": 12,

"cache_dir": str(CACHE_DIR / "huggingface"),

"type": "huggingface"

}

}

本地模型路径映射 (相对路径)

LOCAL_MODEL_PATHS = {

"Qwen3-0.6B": MODELS_DIR / "qianwen3" / "qianwen0.6",

"Qwen2.5-0.5B-文档版": MODELS_DIR / "trained" / "20260129_195952" / "final_model",

"Qwen2.5-0.5B-雪雪训练": MODELS_DIR / "trained" / "20260201_214732" / "checkpoint-400",

"Qwen2.5-0.5B-阿米娅训练": MODELS_DIR / "trained" / "20260202_170728" / "final_model"

}

===== 生成配置 =====

GENERATION_CONFIG = {

"default": {

"max_new_tokens": 512,

"temperature": 0.7,

"top_p": 0.8,

"repetition_penalty": 1.05,

"do_sample": True,

"pad_token_id": None, # 自动设置

"eos_token_id": None # 自动设置

}

"creative": {

"max_new_tokens": 768,

"temperature": 0.8,

"top_p": 0.9,

"repetition_penalty": 1.1,

"do_sample": True

}

}

4、也可以将自己训练的FP16模型（也就是正常模型）放在这里，上面训练模型也会自动存放在这里，程序会自动在这里检测模型，没有检测到，就是位置没放对。MODELS指的是模型目录，已经自动检测



🚀 AMD 7900XTX AI助手 - 增强版

模型切换 | 对话记忆 | DirectML加速

📁 项目目录: F:\py_work\AMD_AI_Project\AMD_AI_Project

⚙️ 模型控制

选择AI模型

Qwen3-0.6B

📊 模型信息

📊 模型信息:

- 名称: Qwen3-0.6B
- 路径: F:\py_work\AMD_AI_Project\AMD_AI_Project\models\qianwen3\qianwen0.6
- 状态: ☒ 可用
- 描述: 本地模型: Qwen3-0.6B
- 大小: 1.0 GB
- 推荐显存: 4 GB
- 类型: local

🚀 加载/切换模型

📋 加载状态

请选择模型并点击加载

💬 对话记忆

📜 记忆历史

💬 聊天界面

💬 对话

- 7、选择模型（检测到的都会出现在这）
- 8、加载模型 约10s(由运算能力和模型大小决定)
- 9、输入消息

这里我增加了深度思考功能，会以最后一次点击为准。效果持续约5轮对话，默认是深度思考

💬 输入消息

请输入您的问题...

🤔 深度思考

🤔 否深度思考

模型切换 | 对话记忆 | DirectML加速

项目目录: F:\py work\AMD AI Project\AMD AI Project


 模

选择AI模型







▼

模型信息

模型信

- 名称: Qwen3-0.6B
- 路径: F:\py_work\AMD_AI_Project\AMD_AI_Project\models\qianwen3\qianwen0.6
- 状态:  可用
- 描述: 本地模型: Qwen3-0.6B
- 大小: 1.0 GB
- 推荐显存: 4 GB
- 类型: local

 加载状态

 模型加载成功!
 当前模型: Qwen3-0.6B
 运行设备: DirectML
 词汇量: 151,643
 模型大小: 1.0 GB
 模型路径:
 F:\py_work\AMD_AI_Project\AMD_AI_Project\models\qianwen3\qianwen0.6

💡 可以开始聊天了!

对

这样就完成了生成任务。

 聊

 对

<think> 嗯，用户问的是“鱼香肉丝怎么做？”，我需要先回忆一下这个菜的基本做法。鱼香肉丝应该是一道以鱼和肉为主料的菜肴，通常会用酱油、醋、糖等调料来调味。首先，我得确认用户是否了解基本的食材，比如鱼片或者鸡肉，还有调料的比列。

可能用户是第一次尝试做菜，所以需要详细步骤。首先，准备食材，比如鱼片、肉丝、酱油、醋、糖、葱姜蒜等。然后，处理鱼片，可能需要先煮熟，再用葱姜去腥。接着，炒制的时候，要控制火候，避免糊锅。最后，淋上鱼香酱，让味道更浓郁。

还要考虑用户可能的疑问，比如如何调制鱼香酱，或者如何搭配其他配料。可能需要提醒用户提前准备好调料，或者建议使用不同的肉类代替鱼。另外，用户可能想知道是否有其他版本，比如用蔬菜或海鲜，但问题里没有提到，所以保持基础回答即可。

最后，检查步骤是否清晰，确保用户能够按照步骤操作，并且有正确的调料比例。可能还需要提醒用户注意火候，避免过早翻炒导致食材变色。这样用户就能正确制作出美味的鱼香肉丝了。

</think>

鱼香肉丝是一种经典家常菜，主要以鱼肉或猪肉为主料，搭配酱油、醋、糖、葱姜蒜等调料烹制而成。以下是详细的做法：

材料准备:

- 鱼片 (可选) : 约200克
- 肉丝 (如猪肉或鸡肉) : 约200克
- 酱油 · 话量

● 输入消息

请输入您的问题..

 深度思考

🤔 香深度思

🚀 AMD 7900XTX AI助手 - 增强版

模型切换 | 对话记忆 | DirectML加速

📁 项目目录: F:\py_work\AMD_AI_Project\AMD_AI_Project

⚙️ 模型控制

选择AI模型

Qwen3-0.6B

📊 模型信息

📊 模型信息:

- 名称: Qwen3-0.6B
- 路径: F:\py_work\AMD_AI_Project\AMD_AI_Project\models\qianwen3\qianwen0.6
- 状态: ☒ 可用
- 描述: 本地模型: Qwen3-0.6B
- 大小: 1.0 GB
- 推荐显存: 4 GB
- 类型: local

🚀 加载/切换模型

📷 加载状态

☒ 模型加载成功!

🤖 当前模型: Qwen3-0.6B

🎮 运行设备: DirectML

📊 词汇量: 151,643

📁 模型大小: 1.0 GB

📁 模型路径: F:\py_work\AMD_AI_Project\AMD_AI_Project\models\qianwen3\qianwen0.6

💡 可以开始聊天了!

💬 聊天界面

💬 对话

将鱼片或肉丝煮熟，捞出备用。

◦ 葱姜蒜切末备用。

2. 调制鱼香酱:

◦ 在碗中加入酱油、白醋、糖和少许盐，搅拌均匀。

3. 炒制肉丝:

◦ 热锅凉油后放入肉丝，翻炒至变色。

◦ 加入葱姜蒜末炒香，倒入调好的鱼香酱继续翻炒。



生成统计:

- 速度: 5.0 token/秒
- 长度: 512 tokens
- 时间: 103.38秒
- 上下文长度: 17 tokens
- 模型: Qwen3-0.6B
- 类型: local

速度非常感人只有5token/s，这里是以生成的汉字数量计数。理论来说应该会高的多，如果之后各大佬的组件更新，应该会提速。不过现在这个速度，用来日常聊天非常不错。

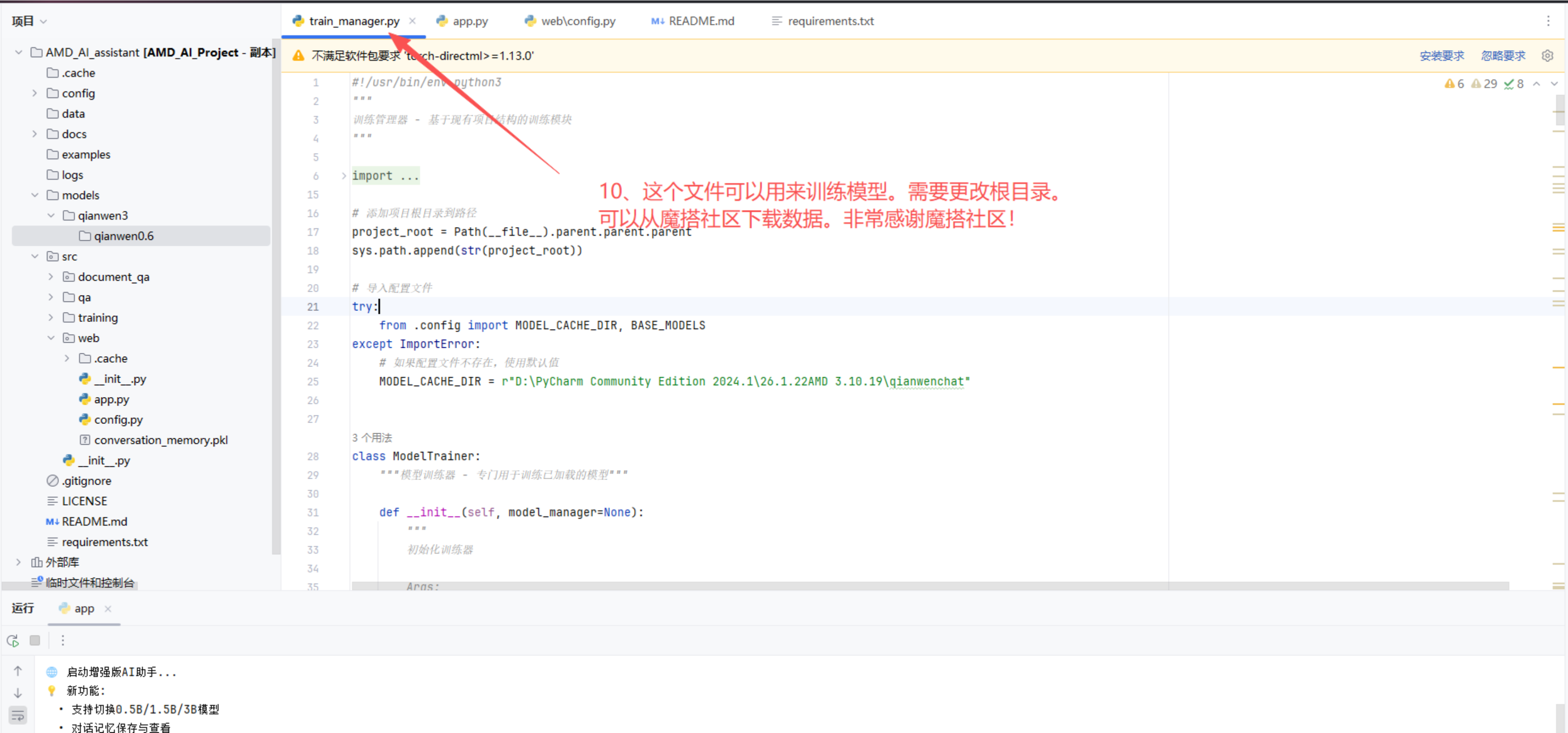


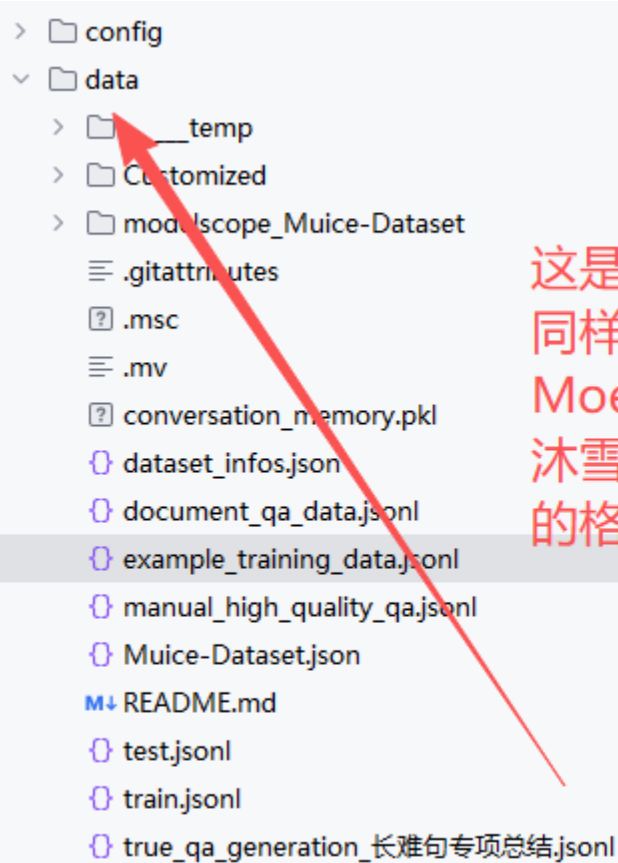
💬 输入消息

请输入您的问题...

🤔 深度思考

🤔 否深度思考





这是数据格式，
同样也支持魔搭社区
Moemu老师的
沐雪中文训练集
的格式

训练数据下载到这个根目录下的文件夹

```
5 {"instruction": "什么是神经网络?", "response": "[样本5] 神经网络是受生物神经网络启发的计算模型, 由相互连接的节点(神经元)组成, 用于模拟"}
6 {"instruction": "介绍一下Python语言的特点和应用场景", "response": "[样本6] Python是一种高级编程语言, 具有以下特点: 1. 简洁易读的语法"}
7 {"instruction": "介绍一下Python语言", "response": "[样本7] Python是一种高级编程语言, 由Guido van Rossum于1991年创建。它具有简洁易"}
8 {"instruction": "什么是机器学习?", "response": "[样本8] 机器学习是人工智能的一个分支, 它使计算机能够从数据中学习并做出预测或决策, 而无"}
9 {"instruction": "如何学习编程?", "response": "[样本9] 学习编程可以从基础开始: 1. 选择一门语言(如Python) 2. 学习基本语法 3. 做实际"}
10 {"instruction": "AMD显卡有什么优势?", "response": "[样本10] AMD显卡的优势包括: 1. 性价比高 2. 支持开源驱动 3. 在专业计算和AI方面表"}
11 {"instruction": "什么是神经网络?", "response": "[样本11] 神经网络是受生物神经网络启发的计算模型, 由相互连接的节点(神经元)组成, 用于模"}
12 {"instruction": "介绍一下Python语言的特点和应用场景", "response": "[样本12] Python是一种高级编程语言, 具有以下特点: 1. 简洁易读的语法"}
13 {"instruction": "介绍一下Python语言", "response": "[样本13] Python是一种高级编程语言, 由Guido van Rossum于1991年创建。它具有简洁"}
14 {"instruction": "什么是机器学习?", "response": "[样本14] 机器学习是人工智能的一个分支, 它使计算机能够从数据中学习并做出预测或决策, 而无"}
15 {"instruction": "如何学习编程?", "response": "[样本15] 学习编程可以从基础开始: 1. 选择一门语言(如Python) 2. 学习基本语法 3. 做实际"}
16 {"instruction": "AMD显卡有什么优势?", "response": "[样本16] AMD显卡的优势包括: 1. 性价比高 2. 支持开源驱动 3. 在专业计算和AI方面表"}
17 {"instruction": "什么是神经网络?", "response": "[样本17] 神经网络是受生物神经网络启发的计算模型, 由相互连接的节点(神经元)组成, 用于模"}
18 {"instruction": "介绍一下Python语言的特点和应用场景", "response": "[样本18] Python是一种高级编程语言, 具有以下特点: 1. 简洁易读的语法"}
19 {"instruction": "介绍一下Python语言", "response": "[样本19] Python是一种高级编程语言, 由Guido van Rossum于1991年创建。它具有简洁"}
20 {"instruction": "什么是机器学习?", "response": "[样本20] 机器学习是人工智能的一个分支, 它使计算机能够从数据中学习并做出预测或决策, 而无"}
21
```