AI 大模型开发工程师
之基于微调模型的知识库项目改造

讲师：李希沅

目录

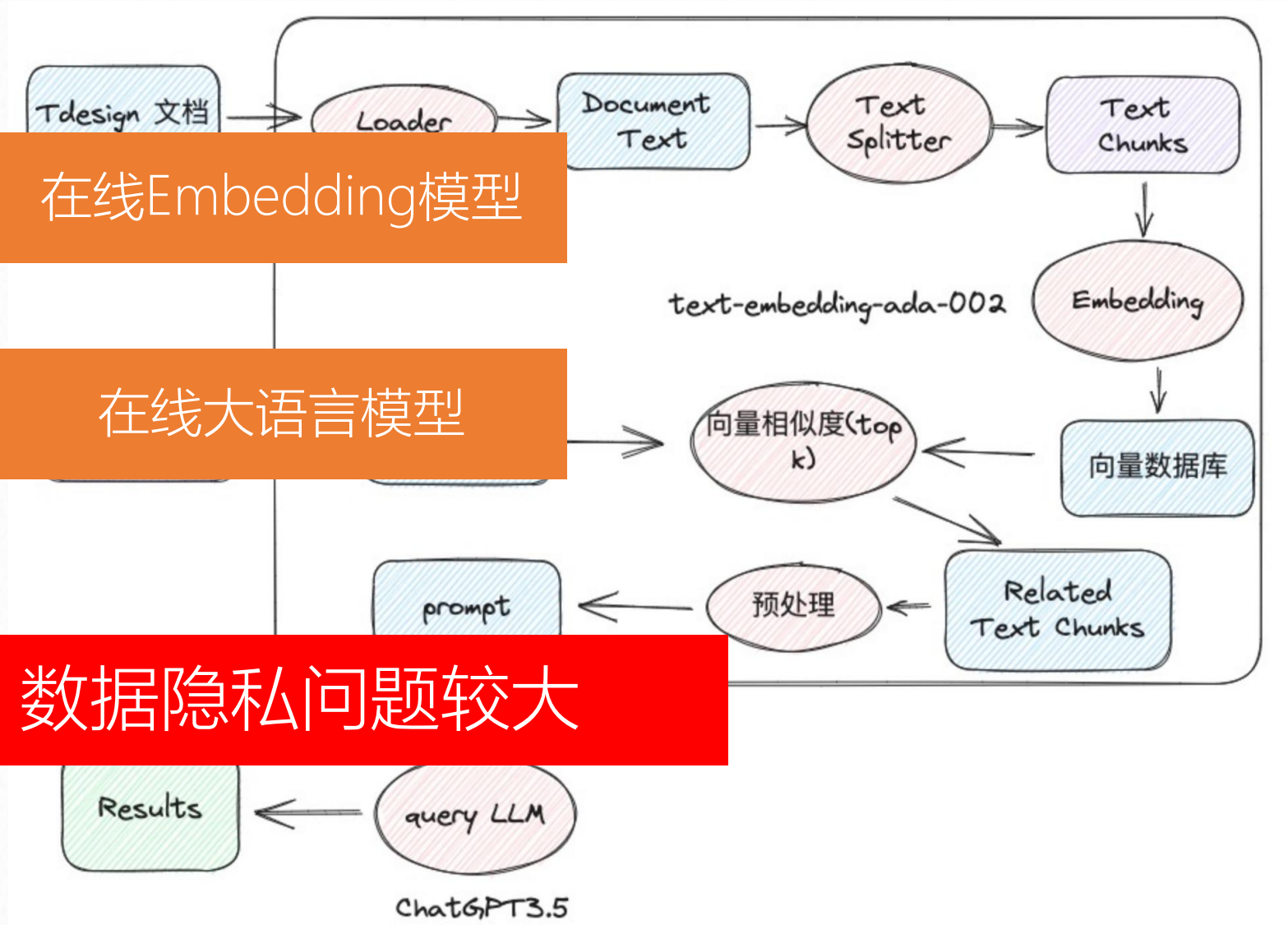
1 基于微调模型的本地知识库架构改造

01、CVP架构模式回顾

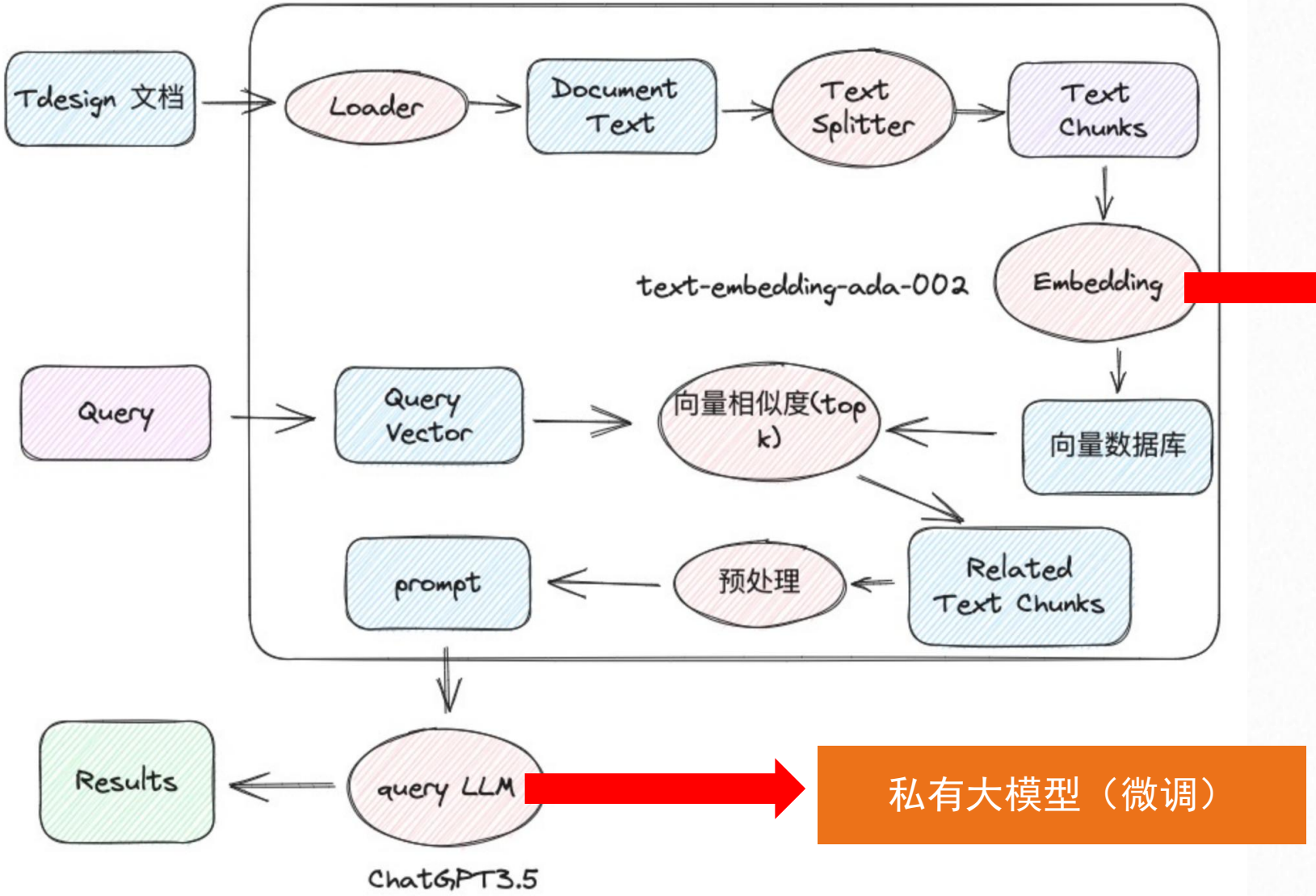
1、知识数据向量化

2、知识数据召回

3、查询返回结果



02、RAG（retrieval augmented generation）



私有Embedding模型

微调+向量库

私有大模型（微调）

03、模型的选型

开源的

可商业使用的

支持中文的

高性能的

低成本可部署的

ChatGLM3 是智谱AI和清华大学 KEG 实验室联合发布的新一代对话预训练模型。ChatGLM3-6B 是 ChatGLM3 系列中的开源模型，在保留了前两代模型对话流畅、部署门槛低等众多优秀特性的基础上，ChatGLM3-6B 引入了如下特性：

1. **更强大的基础模型：** ChatGLM3-6B 的基础模型 ChatGLM3-6B-Base 采用了更多样的训练数据、更充分的训练步数和更合理的训练策略。在语义、数学、推理、代码、知识等不同角度的数据集上测评显示，**ChatGLM3-6B-Base 具有在 10B 以下的基础模型中最强的性能。**
2. **更完整的功能支持：** ChatGLM3-6B 采用了全新设计的 Prompt 格式，除正常的多轮对话外。同时原生支持工具调用（Function Call）、代码执行（Code Interpreter）和 Agent 任务等复杂场景。
3. **更全面的开源序列：** 除了对话模型 ChatGLM3-6B 外，还开源了基础模型 ChatGLM3-6B-Base、长文本对话模型 ChatGLM3-6B-32K。以上所有权重对学术研究**完全开放**，在填写问卷进行登记后**亦允许免费商业使用。**

通义千问-72B（Qwen-72B）主要有以下特点：

1. **大规模高质量训练语料**：使用超过3万亿tokens的数据进行预训练，包含高质量中、英、多语言、代码、数学等数据，涵盖通用及专业领域的训练语料。通过大量对比实验对预训练语料分布进行了优化。
2. **强大的性能**：Qwen-72B在多个中英文下游评测任务上（涵盖常识推理、代码、数学、翻译等），效果显著超越现有的开源模型。具体评测结果请详见下文。
3. **覆盖更全面的词表**：相比目前以中英词表为主的开源模型，Qwen-72B使用了约15万大小的词表。该词表对多语言更加友好，方便用户在不扩展词表的情况下对部分语种进行能力增强和扩展。
4. **较长的上下文支持**：Qwen-72B支持32k的上下文长度。

- **运行BF16或FP16模型需要多卡至少144GB显存（例如2xA100-80G或5xV100-32G）；运行Int4模型至少需要48GB显存（例如1xA100-80G或2xV100-32G）。**

04、Embedding模型选择

MTEB排行榜：https://huggingface.co/spaces/mteb/leaderboard

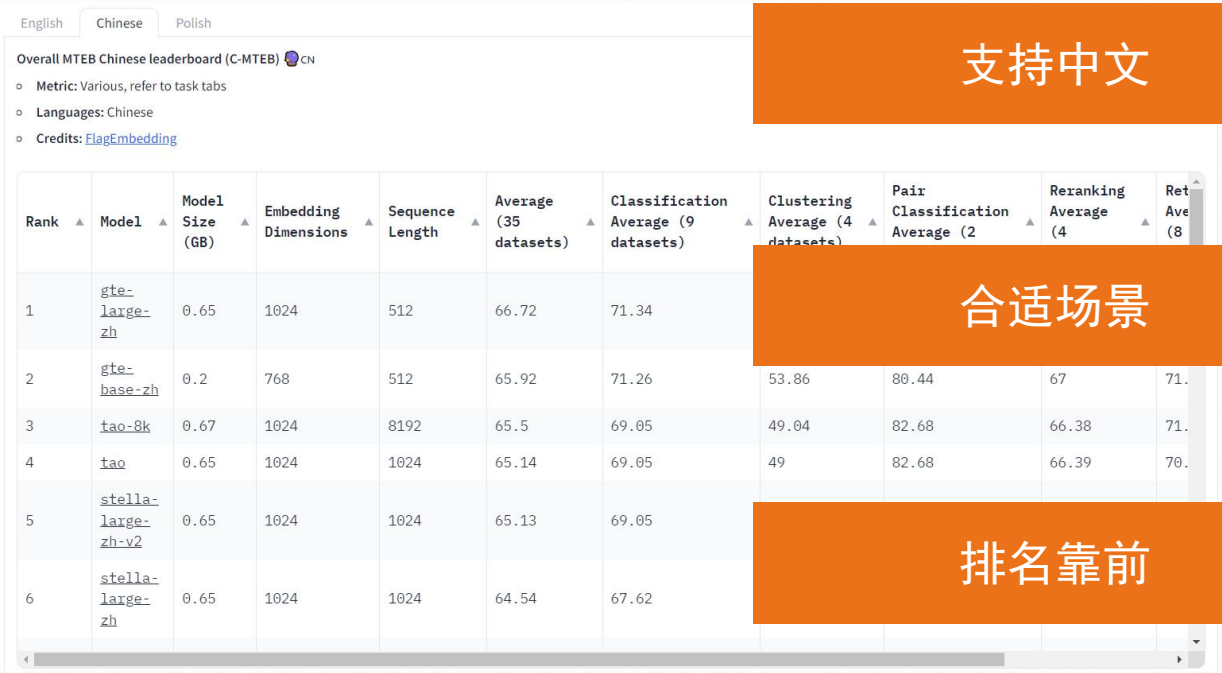
MTEB 是衡量文本嵌入模型在各种嵌入任务上性能的重要基准

开源的

支持中文

合适场景

排名靠前



Rank	Model	Model Size (GB)	Embedding Dimensions	Sequence Length	Average (35 datasets)	Classification Average (9 datasets)	Clustering Average (4 datasets)	Pair Classification Average (2	Reranking Average (4	Ret Ave (8
1	gte-large-zh	0.65	1024	512	66.72	71.34				
2	gte-base-zh	0.2	768	512	65.92	71.26	53.86	80.44	67	71.
3	tao-8k	0.67	1024	8192	65.5	69.05	49.04	82.68	66.38	71.
4	tao	0.65	1024	1024	65.14	69.05	49	82.68	66.39	70.
5	stella-large-zh-v2	0.65	1024	1024	65.13	69.05				
6	stella-large-zh	0.65	1024	1024	64.54	67.62				

- shibing624/text2vec-base-chinese模型，是用CoSENT方法训练，基于hfl/chinese-macbert-base在中文STS-B数据训练得到，并在中文STS-B测试集评估达到较好效果，运行examples/training_sup_text_matching_model.py代码可训练模型，模型文件已经上传HF model hub，中文通用语义匹配任务推荐使用

```
# 加载embedding
embedding_model_dict = {
    "thenlper-base": "thenlper/gte-base-zh",
    "ernie-base": "nghuyong/ernie-3.0-base-zh",
    "text2vec": "GanymedeNil/text2vec-large-chinese",
    "text2vec2": "uer/sbert-base-chinese-nli",
    "text2vec3": "shibing624/text2vec-base-chinese",
}
```

05、改造后的技术选型

类型	对应技术
LLM模型	ChatGLM3-6B
Embedding模型	text2vec-base-chinese或者更多
LLM应用开发框架	LangChain
向量数据库	FAISS/pinecone/Milvus
前端框架	streamlit/gradio
训练技术	高效微调
高效微调技术	LoRA

2 基于微调模型的本地知识库资源准备

01、资源评估

推理的GPU资源要求

简单测试样例的实际测试数据

量化等级	生成 8192 长度的最小显存
FP16	15.9 GB
INT8	11.1 GB
INT4	8.5 GB

1. NVIDIA Pascal架构的GPU，如TitanXp，GTX 10系列等。 这类GPU缺乏低精度的硬件加速能力，但却具备中等的单精度算力。由于价格便宜，适合用来练习训练小模型(如Cifar10)或调试模型代码。
2. NVIDIA Volta/Turing架构的GPU，如GTX 20系列, Tesla V100等。 这类GPU搭载专为低精度(int8/float16)计算加速的TensorCore, 但单精度算力相较于上代提升不大。我们建议在实例上启用深度学习框架的混合精度训练来加速模型计算。 相较于单精度训练，混合精度训练通常能够提供2倍以上的训练加速。
3. NVIDIA Ampere架构的GPU，如GTX 30系列，Tesla A40/A100等。 这类GPU搭载第三代TensorCore。相较于前一代，支持了TensorFloat32格式，可直接加速单精度训练 (PyTorch已默认开启)。但我们仍建议使用超高算力的float16半精度训练模型，可获得比上一代GPU更显著的性能提升。
4. 寒武纪 MLU 200系列加速卡。 暂不支持模型训练。使用该系列加速卡进行模型推理需要量化为int8进行计算。 并且需要安装适配寒武纪MLU的深度学习框架。
5. 华为 Ascend 系列加速卡。 支持模型训练及推理。但需安装MindSpore框架进行计算。

镜像	PyTorch 2.0.0 Python 3.8(ubuntu20.04) Cuda 11.8 更换
GPU	RTX 4090(24GB) * 1 升降配置
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz
内存	90GB
硬盘	系统盘: 30 GB

每块GPU应配备至少4~8核心的CPU

02、私有模型部署

获取工程

```
root@autodl-container-dd9f46bdad-dd54918f:~/glm# git clone https://github.com/THUDM/ChatGLM3
Cloning into 'ChatGLM3'...
remote: Enumerating objects: 469, done.
remote: Counting objects: 100% (234/234), done.
remote: Compressing objects: 100% (107/107), done.
remote: Total 469 (delta 154), reused 167 (delta 122), pack-reused 235
Receiving objects: 100% (469/469), 15.19 MiB | 12.18 MiB/s, done.
Resolving deltas: 100% (242/242), done.
root@autodl-container-dd9f46bdad-dd54918f:~/glm# cd ChatGLM3
root@autodl-container-dd9f46bdad-dd54918f:~/glm/ChatGLM3#
```

安装依赖

```
root@autodl-container-dd9f46bdad-dd54918f:~/glm/ChatGLM3# pip install -r requirements.txt
Looking in indexes: http://mirrors.aliyun.com/pypi/simple
Requirement already satisfied: protobuf in /root/miniconda3/lib/python3.8/site-packages (from -r requirement
Collecting transformers>=4.30.2
  Downloading http://mirrors.aliyun.com/pypi/packages/12/dd/f17b11a93a9ca27728e12512d167eb1281c151c4c6881d3
     | 2.3 MB 1.4 MB/s eta 0:00:04
```

私有模型测试

streamlit run web_demo2.py

03、个性化需求

个性化需求1：本地访问服务器端口

使用SSH将实例中的端口代理到本地，具体步骤为：

Step.1 在实例中启动您的服务（比如您的服务监听6006端口，下面以6006端口为例）

Step.2 在本地电脑的终端(cmd / powershell / terminal等)中执行代理命令：

ssh -CNg -L 6006:127.0.0.1:6006 root@123.125.240.150 -p 42151

其中root@123.125.240.150和42151分别是实例中SSH指令的访问地址与端口，请找到自己实例的ssh指令做相应替换。

6006:127.0.0.1:6006是指代理实例内6006端口到本地的6006端口。

（注：仅限于跟我一样用的AutoDL平台的场景）

个性化需求2: 修改streamlit的端口号为：6006

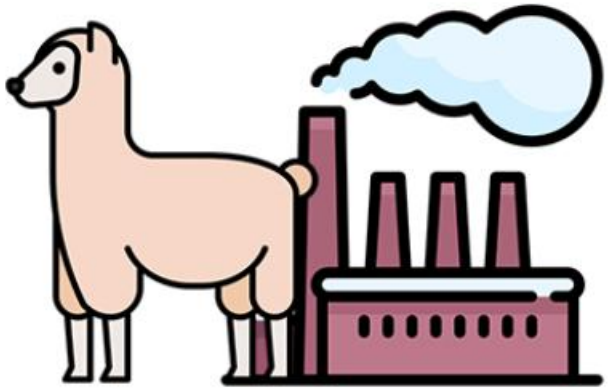
find / -name config.py

进入文件：/root/miniconda3/lib/python3.8/site-packages/streamlit/config.py

```
_create_option(
    "server.port",
    description="""
        The port where the server will listen for browser connections.""",
    default_val=6006,
    type_=int,
)
```

3 基于微调模型的本地知识库模型微调

01、LLama-Factory



LLaMA-Factory

Easy and Efficient LLM Fine-Tuning

简单易用的训练推理一体化 Web UI

git地址：https://github.com/hiyouga/LLaMA-Factory.git

训练、评估和推理一体化界面

几乎为 0 的命令行操作和零代码编辑

预置 25 种模型和 24 种多语言训练数据

中英文双语界面即时切换

即时的训练状态监控和简洁的模型断点管理

1. 下载工程

git clone https://github.com/hiyouga/LLaMA-Factory.git

conda create -n llama_factory python=3.10

conda activate llama_factory

2. 安装依赖

cd LLaMA-Factory

pip install -r requirements.txt

3. 启动服务

python src/train_web.py

#模型下载

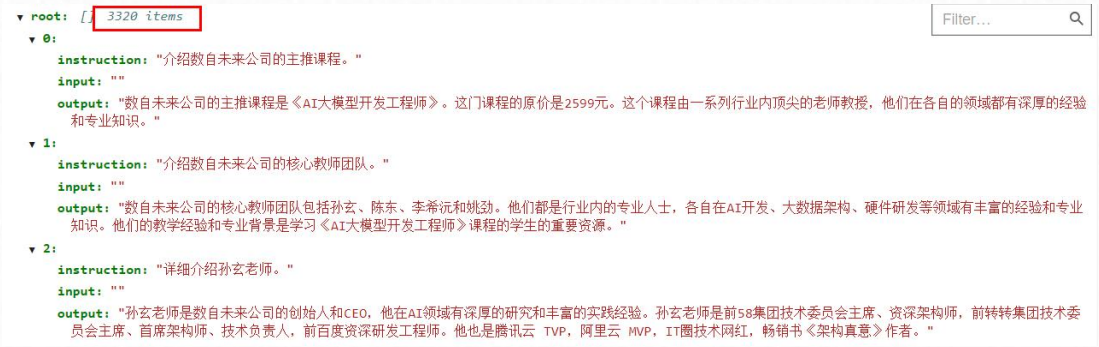
from modelscope import snapshot_download

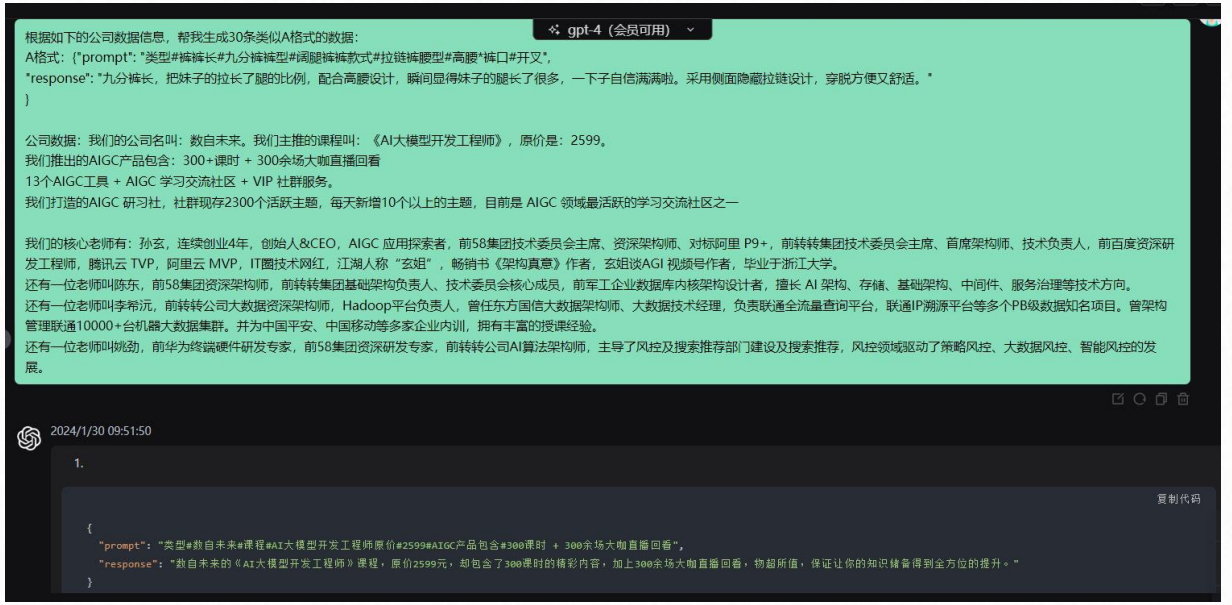
snapshot_download(model_id='ZhipuAI/chatglm3-6b',cache_dir="./")

```
root@autodl-container-cb2246bbbe-1e711273:~/antodl-tmp/zhipu/ZhipuAI/chatglm3-6b# ll
total 12195732
drwxr-xr-x 2 root root       4096 Jan 30 10:12 ./
drwxr-xr-x 3 root root         33 Jan 30 10:10 ../
-rw-r--r-- 1 root root         42 Jan 30 10:10 .mdl
-rw------- 1 root root       1422 Jan 30 10:12 .msc
-rw------- 1 root root       4133 Jan 30 10:10 MODEL_LICENSE
-rw------- 1 root root       4478 Jan 30 10:12 README.md
-rw------- 1 root root       1317 Jan 30 10:10 config.json
-rw------- 1 root root         37 Jan 30 10:10 configuration.json
-rw------- 1 root root       2332 Jan 30 10:10 configuration_chatglm.py
-rw------- 1 root root      55596 Jan 30 10:10 modeling_chatglm.py
-rw------- 1 root root 1827781090 Jan 30 10:10 pytorch_model-00001-of-00007.bin
-rw------- 1 root root 1968299480 Jan 30 10:11 pytorch_model-00002-of-00007.bin
-rw------- 1 root root 1927415036 Jan 30 10:11 pytorch_model-00003-of-00007.bin
-rw------- 1 root root 1815225998 Jan 30 10:11 pytorch_model-00004-of-00007.bin
-rw------- 1 root root 1968299544 Jan 30 10:12 pytorch_model-00005-of-00007.bin
-rw------- 1 root root 1927415036 Jan 30 10:12 pytorch_model-00006-of-00007.bin
-rw------- 1 root root 1052808542 Jan 30 10:12 pytorch_model-00007-of-00007.bin
-rw------- 1 root root      20437 Jan 30 10:12 pytorch_model.bin.index.json
-rw------- 1 root root      14692 Jan 30 10:12 quantization.py
-rw------- 1 root root      11279 Jan 30 10:12 tokenization_chatglm.py
-rw------- 1 root root    1018370 Jan 30 10:12 tokenizer.model
-rw------- 1 root root        244 Jan 30 10:12 tokenizer_config.json
```

02、数据准备


gpt-4（会员可用）
2024/1/30 10:34:25
根据如下的公司数据信息，帮我生成30条类似A格式的数据：
A格式：{
"instruction": "三原色是什么？",
"input": "",
"output": "三原色通常指的是红色、绿色和蓝色（RGB）。它们是通过加色混合原理创建色彩的三种基础颜色。"
}
公司数据：我们的公司名叫：数自未来。我们主推的课程叫：《AI大模型开发工程师》，原价是：2599。
我们推出的AIGC产品包含：300+课时 + 300余场大咖直播回看
13个AIGC工具 + AIGC 学习交流社区 + VIP 社群服务。
我们打造的AIGC 研习社，社群现存2300个活跃主题，每天新增10个以上的主题，目前是 AIGC 领域最活跃的学习交流社区之一
2024/1/30 10:34:25
1. { "instruction": "数自未来是做什么的？", "input": "", "output": "数自未来是一家提供AI教育服务的公司，旗下有《AI大模型开发工程师》等课程。" }
2. { "instruction": "数自未来主推的是哪个课程？", "input": "", "output": "数自未来主推的课程是《AI大模型开发工程师》。" }
3. { "instruction": "《AI大模型开发工程师》课程的原价是多少？", "input": "", "output": "《AI大模型开发工程师》课程的原价是2599。" }
4. { "instruction": "AIGC产品包含哪些内容？", "input": "", "output": "AIGC产品包含300+课时，300余场大咖直播回看，13个AIGC工具，AIGC学习交流社区，以及VIP社群服务。" }
5. { "instruction": "数自未来的AIGC研习社有多少个活跃主题？", "input": "", "output": "数自未来的AIGC研习社有2300个活跃主题。" }
6. { "instruction": "AIGC研习社每天新增多少个主题？", "input": "", "output": "AIGC研习社每天新增10个以上的主题。" }
7. { "instruction": "AIGC研习社在哪个领域是最活跃的学习交流社区之一？", "input": "", "output": "AIGC研习社在AIGC领域是最活跃的学习交流社区之一。" }
8. { "instruction": "AIGC产品提供了多少个工具？", "input": "", "output": "AIGC产品提供了13个工具。" }
9. { "instruction": "《AI大模型开发工程师》这个课程提供多少课时？", "input": "", "output": "《AI大模型开发工程师》这个课程提供了300+课时。" }


root: [] 3320 items
Filter...
0:
instruction: "介绍数自未来公司的主推课程。"
input: ""
output: "数自未来公司的主推课程是《AI大模型开发工程师》。这门课程的原价是2599元。这个课程由一系列行业内顶尖的老师教授，他们在各自的领域都有深厚的经验和专业知识。"
1:
instruction: "介绍数自未来公司的核心教师团队。"
input: ""
output: "数自未来公司的核心教师团队包括孙玄、陈东、李希沅和姚劲。他们都是行业内的专业人士，各自在AI开发、大数据架构、硬件研发等领域有丰富的经验和专业知识。他们的教学经验和专业背景是学习《AI大模型开发工程师》课程的学生的重要资源。"
2:
instruction: "详细介绍孙玄老师。"
input: ""
output: "孙玄老师是数自未来公司的创始人和CEO，他在AI领域有深厚的研究和丰富的实践经验。孙玄老师是前58集团技术委员会主席、资深架构师，前转转集团技术委员会主席、首席架构师、技术负责人，前百度资深研发工程师。他也是腾讯云 TVP，阿里云 MVP，IT圈技术网红，畅销书《架构真意》作者。"

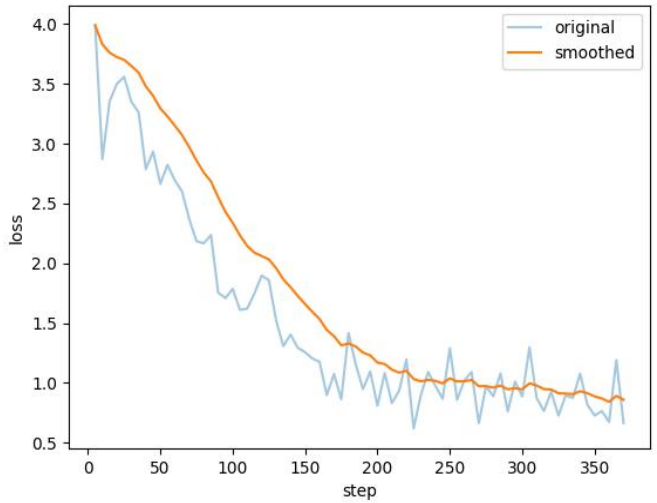

gpt-4（会员可用）
根据如下的公司数据信息，帮我生成30条类似A格式的数据：
A格式：{"prompt": "类型#裤裤长#九分裤裤型#阔腿裤裤款式#拉链裤腰型#高腰*裤口#开叉",
"response": "九分裤长，把妹子的拉长了腿的比例，配合高腰设计，瞬间显得妹子的腿长了很多，一下子自信满满啦。采用侧面隐藏拉链设计，穿脱方便又舒适。"
}
公司数据：我们的公司名叫：数自未来。我们主推的课程叫：《AI大模型开发工程师》，原价是：2599。
我们推出的AIGC产品包含：300+课时 + 300余场大咖直播回看
13个AIGC工具 + AIGC 学习交流社区 + VIP 社群服务。
我们打造的AIGC 研习社，社群现存2300个活跃主题，每天新增10个以上的主题，目前是 AIGC 领域最活跃的学习交流社区之一
我们的核心老师有：孙玄，连续创业4年，创始人&CEO，AIGC 应用探索者，前58集团技术委员会主席、资深架构师、对标阿里 P9+，前转转集团技术委员会主席、首席架构师、技术负责人，前百度资深研发工程师，腾讯云 TVP，阿里云 MVP，IT圈技术网红，江湖人称“玄姐”，畅销书《架构真意》作者，玄姐谈AGI 视频号作者，毕业于浙江大学。
还有一位老师叫陈东，前58集团资深架构师，前转转集团基础架构负责人、技术委员会核心成员，前军工企业数据库内核架构设计者，擅长 AI 架构、存储、基础架构、中间件、服务治理等技术方向。
还有一位老师叫李希沅，前转转公司大数据资深架构师，Hadoop平台负责人，曾任东方国信大数据架构师、大数据技术经理，负责联通全流量查询平台，联通IP溯源平台等多个PB级数据知名项目。曾架构管理联通10000+台机器大数据集群。并为中国平安、中国移动等多家企业内训，拥有丰富的授课经验。
还有一位老师叫姚劲，前华为终端硬件研发专家，前58集团资深研发专家，前转转公司AI算法架构师，主导了风控及搜索推荐部门建设及搜索推荐，风控领域驱动了策略风控、大数据风控、智能风控的发展。
2024/1/30 09:51:50
1.
复制代码
{
"prompt": "类型#数自未来#课程#AI大模型开发工程师原价#2599#AIGC产品包含#300课时 + 300余场大咖直播回看",
"response": "数自未来的《AI大模型开发工程师》课程，原价2599元，却包含了300课时的精彩内容，加上300余场大咖直播回看，物超所值，保证让你的知识储备得到全方位的提升。"
}

03、数据准备

{
"instruction": "介绍数自未来公司的主推课程。",
"input": "",
"output": "数自未来公司的主推课程是《AI大模型开发工程师》。这门课程的原价是2599元。这个课程由一系列行业内顶尖的老师教授，他们在各自的领域都有深厚的经验和专业知识。"
}
,
{
"instruction": "介绍数自未来公司的核心教师团队。",
"input": "",
"output": "数自未来公司的核心教师团队包括孙玄、陈东、李希沅和姚劲。他们都是行业内的专业人士，各自在AI开发、大数据架构、硬件研发等领域有丰富的经验和专业知识。他们的教学经验和专业背景是学习《AI大模型开发工程师》课程的学生的重要资源。"
}
,
{
"instruction": "详细介绍孙玄老师。",
"input": "",
"output": "孙玄老师是数自未来公司的创始人和CEO，他在AI领域有深厚的研究和丰富的实践经验。孙玄老师是前58集团技术委员会主席、资深架构师，前转转集团技术委员会主席、首席架构师、技术负责人，前百度资深研发工程师。他也是腾讯云 TVP，阿里云 MVP，IT圈技术网红，畅销书《架构真意》作者。"
}

04、微调演示和验证


高级设置
Train
Evaluate & Predict
Chat
Export
训练阶段
目前采用的训练方式。
Supervised Fine-Tuning
数据路径
数据文件夹的路径。
data
数据集
alpaca_gpt4_zh
预览数据集
截断长度
输入序列分词后的最大长度。
1024
学习率
AdamW 优化器的初始学习率。
5e-5
训练轮数
需要执行的训练总轮数。
1.5
最大样本数
每个数据集最多使用的样本数。
20000
计算类型
是否启用 FP16 或 BF16 混合精度训练。
fp16
bf16
fp32
批处理大小
每块 GPU 上处理的样本数量。
3
梯度累积
梯度累积的步数。
4
学习率调节器
采用的学习率调节器名称。
cosine
最大梯度范数
用于梯度裁剪的范数。
1.0
验证集比例
验证集占全部样本的百分比。
0.1


original
smoothed
loss
step
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
50
100
150
200
250
300
350


Chatbot
介绍数自未来公司的主推课程。
数自未来公司的主推课程是《AI大模型开发工程师》。这门课程的原价是2599元。这个课程由一系列行业内顶尖的老师教授，他们在各自的领域都有深厚的经验和专业知识。
详细介绍孙玄老师
孙玄老师是数自未来公司的创始人，也是《AI大模型开发工程师》这门课程的主讲人。他拥有丰富的行业经验和技术专业知识，在AI大模型开发领域有深厚的经验和深厚的造诣。孙玄老师是AI大模型开发的先驱者，他的课程深入浅出，通俗易懂，让学员能够快速掌握AI大模型开发的相关知识。

05、微调演示和验证

[1]:
```
from transformers import AutoTokenizer, AutoModelForCausalLM
```

[2]:
```
tokenizer = AutoTokenizer.from_pretrained("/root/autodl-tmp/chatglm_result_lora", trust_remote_code=True)
```

[4]:
```
import torch
model = AutoModelForCausalLM.from_pretrained("/root/autodl-tmp/chatglm_result_lora",
                                             trust_remote_code=True,
                                             temperature=0,
                                             do_sample=True,
                                             low_cpu_mem_usage=True,
                                             torch_dtype=torch.half,
                                             device_map="auto")
```

```
Loading checkpoint shards: 100% 15/15 [00:04<00:00, 4.87it/s]
WARNING:root:Some parameters are on the meta device device because they were offloaded to the cpu.
```

[5]:
```
model.eval()
print(model.chat(tokenizer, "介绍数自未来公司的主推课程", history=[])[0])
```

```
/root/miniconda3/lib/python3.8/site-packages/transformers/generation/utils.py:1518: UserWarning: You have modified the pretrained model configuration to control generation. This is a deprecated strategy to control generation and will be removed soon, in a future version. Please use and modify the model generation configuration (see https://huggingface.co/docs/transformers/generation_strategies#default-text-generation-configuration )
  warnings.warn(
```

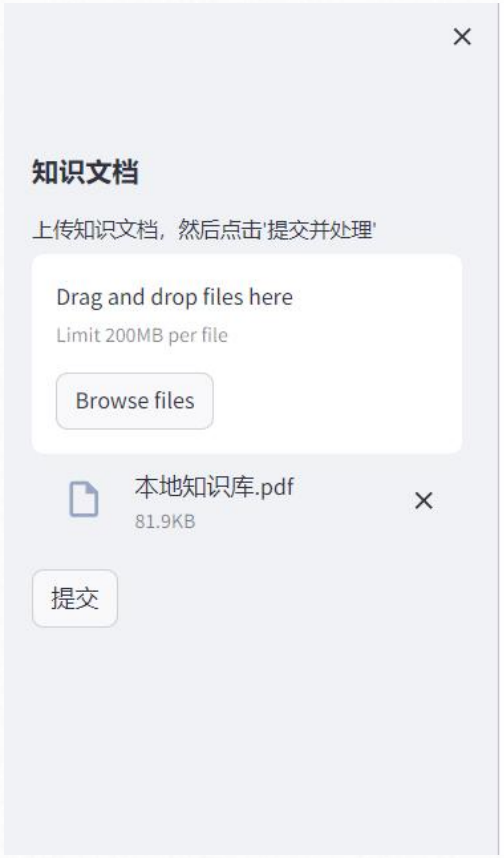
数自未来公司的主推课程是《AI大模型开发工程师》。这门课程是数自未来公司和智谱AI共同推出的，智谱AI是人工智能大模型开发工程师的开创者。

《AI大模型开发工程师》课程的内容包括：

1. AI大模型开发基础：大模型开发工程师需要具备扎实的计算机基础，学习并熟练使用深度学习、自然语言处理、分布式系统等技术。
2. 大模型开发实践：该课程将介绍大模型开发工程师在实践中需要掌握的技能和工具，包括大模型的训练、调优、部署等。
3. 大模型应用实践：课程将介绍大模型在行业中的应用场景，帮助学生理解大模型开发工程师的实际工作内容。

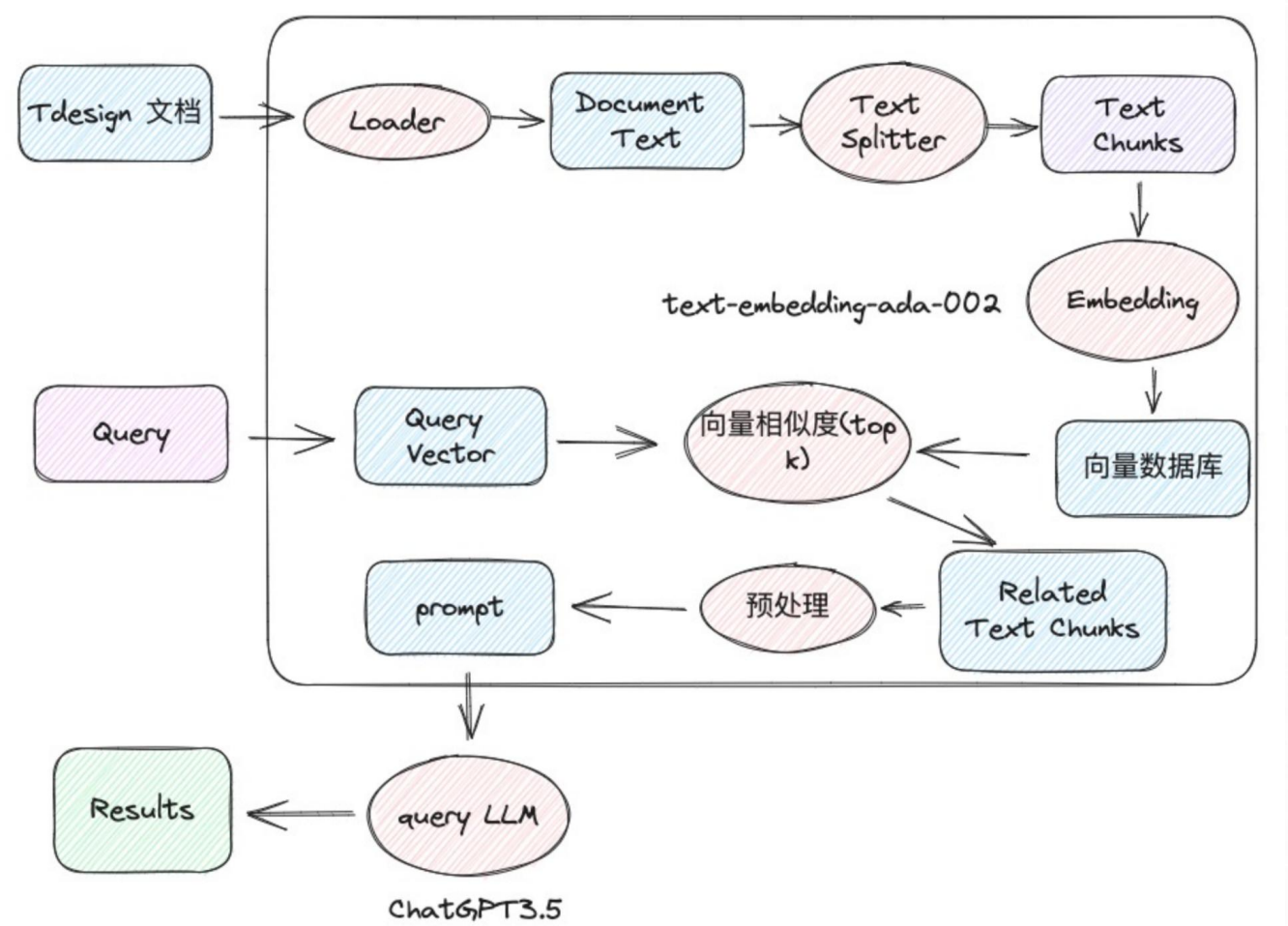
06、项目升级

项目代码升级



4 基于私有模型的本地知识库项目总结

01、项目总结



基于CVP架构

基于私有模型

基于微调

基于私有模型+微调+向量数据库

代码没怎么设计

前端页面不好看

使用的模型较小

微调模型质量一般

02、基于CVP的架构

OpenAI大模型

清华智谱大模型

谷歌Gemini大模型

不需要部署

能使用到最顶级的大模型

使用门槛低

未来会有第三方背书

03、私有化部署

本地私有化定价

智谱AI大模型私有化部署服务是基于大模型算法模型的云计算服务，为客户提供高效的数据处理、模型训练和部署服务。我们为您提供智谱大模型文件和相关的工具包，客户可自行训练和部署推理服务，同时智谱会提供部署应用相关的技术支持和咨询，以及模型的更新。通过私有化部署方案，实现数据的完全掌控和模型的安全运行。

咨询本地私有化定价

模型	套餐包含		单价（年付）
ChatGLM-130B	推理实例license	不限量/年	3960万元/年
	推理&微调工具包	1年	
	咨询服务	15人天/年	
ChatGLM-66B	推理实例license	不限量/年	1680万元/年
	推理&微调工具包	1年	
	咨询服务	15人天/年	
ChatGLM-32B	推理实例license	不限量/年	680万元/年
	推理&微调工具包	1年	
	咨询服务	10人天/年	
ChatGLM-12B	推理实例license	不限量/年	180万元/年
	推理&微调工具包	1年	
	咨询服务	6人天/年	

自己私有化部署

贵

对技术要求高

04、零代码/低代码

GPTs

GLMs

灵境平台

谢谢观看