附录B 开源资源汇总

> 精选学习资源、开源工具、数据集，助力快速上手大模型

B.1 开源大模型

B.1.1 通用基座模型

模型	参数规模	组织	许可	特点
LLaMA-3	8B, 70B	Meta	开源（有限商用）	性能优秀，社区活跃
Mistral	7B	Mistral AI	Apache 2.0	超越LLaMA-2-13B
Qwen	1.8B-72B	阿里巴巴	自定义	中文能力强
ChatGLM3	6B	智谱AI	自定义	中文对话
Baichuan2	7B, 13B	百川智能	自定义	中文
Yi	6B, 34B	零一万物	Apache 2.0	中英双语
Gemma	2B, 7B	Google	Gemma许可	轻量高效
Phi-3	3.8B	Microsoft	MIT	小而强大

B.1.2 代码模型

模型	参数规模	组织	用途
CodeLlama	7B-34B	Meta	代码生成、补全
StarCoder	15B	BigCode	多语言代码
DeepSeek-Coder	1.3B-33B	DeepSeek	代码生成
CodeQwen	7B	阿里巴巴	代码+中文

B.1.3 多模态模型

模型	类型	组织	能力
LLaVA	视觉-语言	UW/MSFT	图像理解
MiniGPT-4	视觉-语言	KAUST	图文对话
Qwen-VL	视觉-语言	阿里巴巴	图像、视频
CogVLM	视觉-语言	智谱AI	高分辨率图像

B.2 推理和部署工具

B.2.1 推理引擎

```
# vLLM（推荐）
from vllm import LLM, SamplingParams

llm = LLM(model="meta-llama/Llama-2-7b-hf")
sampling_params = SamplingParams(temperature=0.8, top_p=0.95)

outputs = llm.generate(prompts, sampling_params)
```

主流推理引擎对比：

引擎	特点	适用场景	性能
vLLM	PagedAttention, 高吞吐	**生产环境首选**	⭐⭐⭐⭐⭐
TensorRT-LLM	NVIDIA优化	NVIDIA GPU	⭐⭐⭐⭐⭐
Text Generation Inference	HuggingFace官方	快速部署	⭐⭐⭐⭐
llama.cpp	CPU推理	边缘设备、Mac	⭐⭐⭐
Ollama	本地运行	个人使用	⭐⭐⭐⭐

安装和使用：

```
# vLLM
pip install vllm

# TensorRT-LLM
pip install tensorrt_llm

# llama.cpp
git clone https://github.com/ggerganov/llama.cpp
cd llama.cpp && make

# Ollama（MacOS/Linux）
curl -fsSL https://ollama.com/install.sh | sh
ollama run llama2
```

B.2.2 微调框架

```
# LLaMA-Factory（推荐）
# 全功能一站式微调平台

git clone https://github.com/hiyouga/LLaMA-Factory
cd LLaMA-Factory
pip install -e .
```

```
# Web UI启动
llamafactory-cli webui
```

主流微调工具：

工具	特点	难度	推荐指数
LLaMA-Factory	全功能Web UI	⭐	⭐⭐⭐⭐⭐
Axolotl	配置文件驱动	⭐⭐	⭐⭐⭐⭐
FastChat	训练+部署	⭐⭐	⭐⭐⭐⭐
PEFT	HuggingFace官方	⭐⭐⭐	⭐⭐⭐⭐
DeepSpeed	大规模训练	⭐⭐⭐⭐	⭐⭐⭐⭐⭐

B.2.3 量化工具

```
# GPTQ量化
from auto_gptq import AutoGPTQForCausalLM, BaseQuantizeConfig

quantize_config = BaseQuantizeConfig(
    bits=4,
    group_size=128,
    desc_act=False,
)

model = AutoGPTQForCausalLM.from_pretrained(
    "meta-llama/Llama-2-7b-hf",
    quantize_config=quantize_config
)

# AWQ量化
from awq import AutoAWQForCausalLM

model = AutoAWQForCausalLM.from_pretrained("llama-2-7b")
model.quantize(tokenizer, quant_config={"bits": 4})
```

量化工具对比：

工具	方法	精度损失	速度
GPTQ	后训练量化	低	快
AWQ	激活感知	更低	更快
GGUF/GGML	llama.cpp格式	中等	快（CPU）
BitsAndBytes	动态量化	低	中

B.3 应用开发框架

B.3.1 LangChain

```python
from langchain.llms import OpenAI
from langchain.prompts import PromptTemplate
from langchain.chains import LLMChain

# 创建链
llm = OpenAI(temperature=0.7)
prompt = PromptTemplate(
    input_variables=["product"],
    template="为{product}写一句广告语："
)
chain = LLMChain(llm=llm, prompt=prompt)

# 运行
result = chain.run("智能手表")
```

LangChain核心组件：

- **Models**：LLM、Chat Models、Embeddings
- **Prompts**：模板、Few-shot、选择器
- **Chains**：组合多个组件
- **Agents**：使用工具的智能体
- **Memory**：对话历史管理
- **Retrievers**：文档检索

B.3.2 LlamaIndex

```python
from llama_index import VectorStoreIndex, SimpleDirectoryReader

# 加载文档
documents = SimpleDirectoryReader('data').load_data()

# 创建索引
index = VectorStoreIndex.from_documents(documents)

# 查询
query_engine = index.as_query_engine()
response = query_engine.query("什么是Transformer？")
```

LlamaIndex vs LangChain：

维度	LangChain	LlamaIndex
定位	通用LLM应用框架	专注RAG和索引
优势	生态丰富、工具多	RAG性能好、易用
适用	复杂应用、Agent	知识库问答

维度	LangChain	LlamaIndex
学习曲线	陡	平缓

B.3.3 其他框架

框架	特点	适用场景
Haystack	NLP pipeline	搜索、问答
Semantic Kernel	Microsoft开发	企业应用
AutoGen	多Agent框架	复杂任务协作
Chainlit	UI框架	快速构建对话界面

B.4 数据集资源

B.4.1 预训练数据集

数据集	规模	语言	获取方式
Common Crawl	PB级	多语言	公开下载
The Pile	825GB	英文	公开
RedPajama	1.2T tokens	多语言	开源
WuDaoCorpora	3TB	中文	申请
CLUECorpus	100GB	中文	开源

B.4.2 指令微调数据集

数据集	规模	语言	类型
Alpaca	52K	英文	Self-Instruct
ShareGPT	90K	多语言	真实对话
BELLE	2M	中文	指令
COIG	多个子集	中文	指令集合
OpenOrca	4.2M	英文	高质量指令

B.4.3 评估数据集

英文：

- **MMLU**：多任务语言理解（57个任务）
- **HellaSwag**：常识推理
- **TruthfulQA**：事实准确性
- **HumanEval**：代码生成（Python）

中文：

- **C-Eval**：中文综合评估（52个学科）
- **CMMLU**：中文多任务理解
- **AGIEval**：中国考试题目

B.5 向量数据库

B.5.1 对比表

数据库	类型	性能	易用性	推荐场景
Chroma	内嵌式	中	⭐⭐⭐⭐⭐	原型开发
Faiss	内存	⭐⭐⭐⭐⭐	⭐⭐⭐	单机、高性能
Milvus	分布式	⭐⭐⭐⭐⭐	⭐⭐⭐	大规模生产
Pinecone	云服务	⭐⭐⭐⭐	⭐⭐⭐⭐⭐	托管服务
Qdrant	Rust	⭐⭐⭐⭐	⭐⭐⭐⭐	高性能、过滤
Weaviate	分布式	⭐⭐⭐	⭐⭐⭐⭐	丰富功能

B.5.2 快速上手

```
# Chroma
import chromadb

client = chromadb.Client()
collection = client.create_collection("my_collection")

collection.add(
    documents=["这是第一个文档", "这是第二个文档"],
    ids=["id1", "id2"]
)

results = collection.query(
    query_texts=["查询文本"],
    n_results=2
)

# Faiss
import faiss
import numpy as np

d = 128  # 维度
index = faiss.IndexFlatL2(d)

# 添加向量
vectors = np.random.random((1000, d)).astype('float32')
index.add(vectors)
```

```
# 搜索
query = np.random.random((1, d)).astype('float32')
D, I = index.search(query, k=5)
```

B.6 模型评估工具

B.6.1 LM Evaluation Harness

```
# 安装
pip install lm-eval

# 评估模型
lm_eval --model hf \
    --model_args pretrained=meta-llama/Llama-2-7b-hf \
    --tasks hellaswag,arc_easy,arc_challenge \
    --device cuda:0 \
    --batch_size 8
```

B.6.2 OpenCompass

```
# 中文模型评估平台
git clone https://github.com/open-compass/opencompass
cd opencompass

# 运行评估
python run.py --models hf_llama_7b --datasets ceval_gen
```

B.6.3 其他工具

工具	用途	特点
Alpaca-Eval	指令遵循	GPT-4作为评判
MT-Bench	多轮对话	8类任务
Arena	人类评估	大规模对战

B.7 学习资源

B.7.1 在线课程

课程	平台	难度	推荐指数
CS224N NLP	Stanford	中	⭐⭐⭐⭐⭐
Fast.ai Practical Deep Learning	fast.ai	低-中	⭐⭐⭐⭐⭐
DeepLearning.AI LLM课程	Coursera	低	⭐⭐⭐⭐

课程	平台	难度	推荐指数
李宏毅机器学习	YouTube	中	⭐⭐⭐⭐⭐

B.7.2 书籍推荐

基础：

1. 《深度学习》（花书）- Ian Goodfellow
2. 《动手学深度学习》- 李沐

NLP/LLM：

1. 《Speech and Language Processing》- Jurafsky & Martin
2. 《Natural Language Processing with Transformers》- HuggingFace

代码实践：

1. 《Build a Large Language Model (From Scratch)》- Sebastian Raschka

B.7.3 博客和社区

必关注博客：

- **Lil'Log**：Lilian Weng（OpenAI）
- **Jay Alammar**：可视化解释Transformer
- **HuggingFace Blog**：最新技术动态
- **Andrej Karpathy Blog**：深度技术讲解

社区：

- **HuggingFace Forums**
- **r/MachineLearning** (Reddit)
- **AI研习社**（中文）
- **知乎 - AI话题**

B.7.4 实战项目

初级项目：

1. 情感分类（微调BERT）
2. 文本摘要（T5）
3. 简单聊天机器人

中级项目：

1. RAG问答系统
2. 代码补全工具
3. 指令微调自己的模型

高级项目：

1. 多Agent协作系统
2. 垂直领域大模型
3. 从头预训练小模型

B.8 开发工具

B.8.1 必备库

```
# 核心库
pip install transformers      # HuggingFace
pip install torch             # PyTorch
pip install accelerate        # 训练加速
pip install bitsandbytes      # 量化

# 应用开发
pip install langchain         # LLM应用框架
pip install llama-index       # RAG框架
pip install chromadb          # 向量数据库

# 训练微调
pip install peft              # PEFT (LoRA等)
pip install deepspeed         # 大规模训练
pip install flash-attn        # Flash Attention

# 推理部署
pip install vllm              # 高效推理
pip install llama-cpp-python # CPU推理
```

B.8.2 GPU云平台

平台	特点	价格	适用
AutoDL	国内、便宜	¥/小时	个人/小团队
Google Colab	免费GPU	免费/付费	学习/实验
Vast.ai	便宜、灵活	$/小时	训练/推理
RunPod	性价比高	$/小时	推理
Lambda Labs	专为深度学习	$/小时	训练

B.9 持续学习资源

B.9.1 论文追踪

- **arXiv**：每日最新论文
- **Papers with Code**：论文+代码+排行榜
- **Hugging Face Daily Papers**：社区精选

B.9.2 技术博客

- **OpenAI Blog**
- **Google AI Blog**

- **Meta AI Blog**
- **Anthropic Blog**

B.9.3 Newsletter

- **The Batch**（DeepLearning.AI）
- **Import AI**（Jack Clark）
- **AI Alignment Newsletter**

B.10 本附录小结

本附录汇总了大模型学习和开发的全部资源：

✅ **开源模型**：LLaMA、Mistral、Qwen等 ✅ **开发工具**：推理引擎、微调框架、应用框架 ✅ **数据集**：预训练、微调、评估数据 ✅ **学习资源**：课程、书籍、博客、社区

使用建议：

1. **入门**：从Colab + 小模型开始
2. **进阶**：本地部署 + 微调实验
3. **深入**：阅读论文 + 复现算法
4. **实战**：完整项目 + 开源贡献

资源更新：

- 本附录持续更新
- 关注HuggingFace和GitHub Trending
- 加入相关技术社区

下一附录： 附录C将提供术语表和中英文对照。