



研讨 / 分享 / 感悟

仓颉社区中的灵感碰撞 社区先行者的干货分享

苍穹 (CangChain)

-- 大语言模型编程框架

主讲人：杨海龙

- 大语言模型
- 苍穹框架
- 开发进展
- 未来规划



大语言模型

什么是大语言模型？



大语言模型是一种基于深度学习的人工智能模型，通常是神经网络的一种变体，被设计用来处理和生成自然语言文本。这些模型之所以称为“大型”，是因为它们包含了数十亿到数百亿个参数，这使得它们能够处理大量的语言数据和理解复杂的语法、语义和上下文。

大语言模型的作用？



自然语言理解和生成、信息检索和推荐系统、文本分类和情感分析、自动化写作和内容生成、语音识别和生成、虚拟助手和对话系统、知识提取和问题回答、自动化翻译和跨语言沟通、医疗诊断和科学研究、教育和培训。

大语言模型有哪些缺点？



数据偏见、生成不当内容、数据隐私问题、资源消耗、过拟合、缺乏常识推理、能量效率低下、长期依赖问题、算法黑盒性、滥用风险

大语言模型有这么多的缺点，作为它的编程框架能做什么呢？

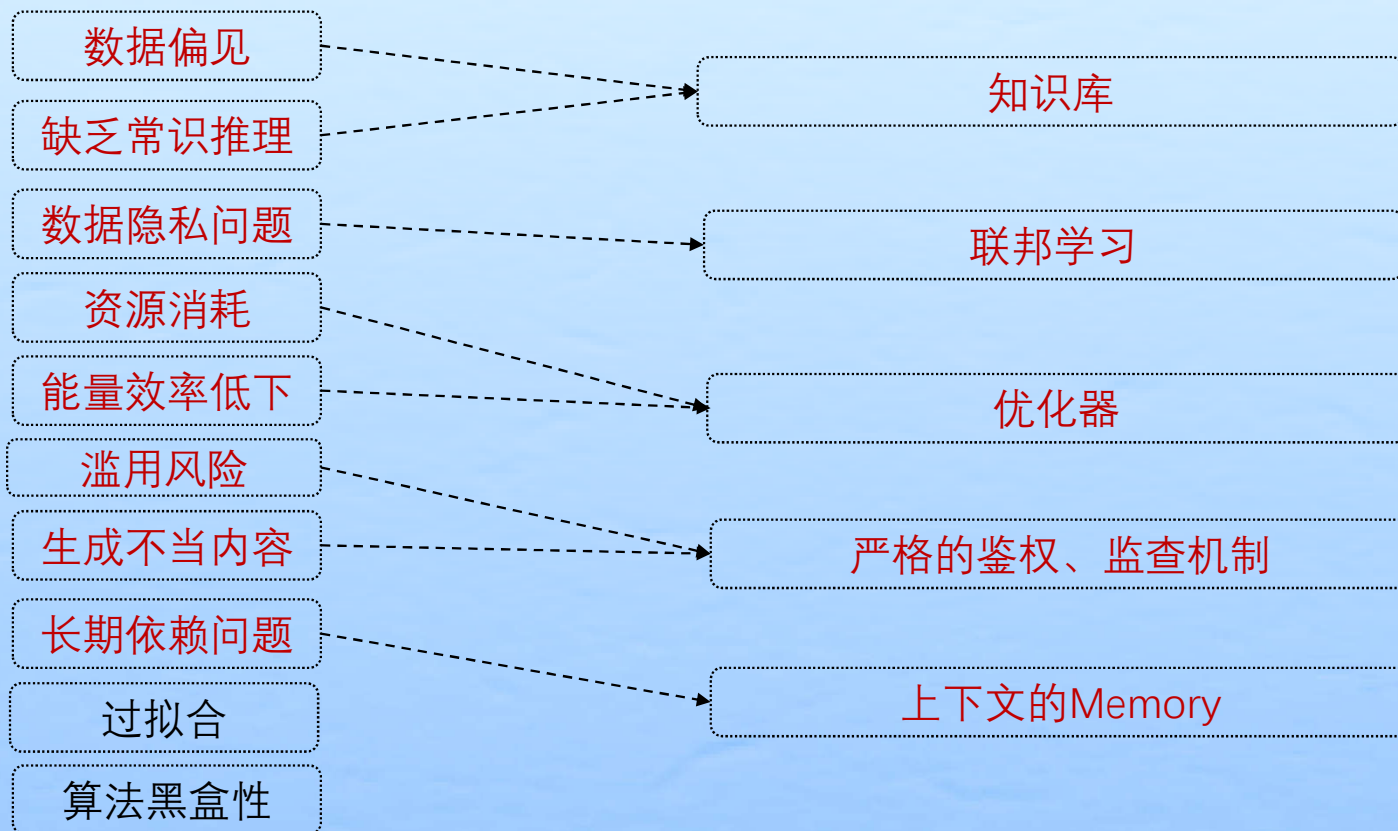
WORKSHOP



大语言模型与编程框架

大语言模型

编程框架



本地小模型
专家系统调用Tool
向量数据库
Knowledge Graph

优化冗余的query
优化提示词，减少token数
本地小模型截获query

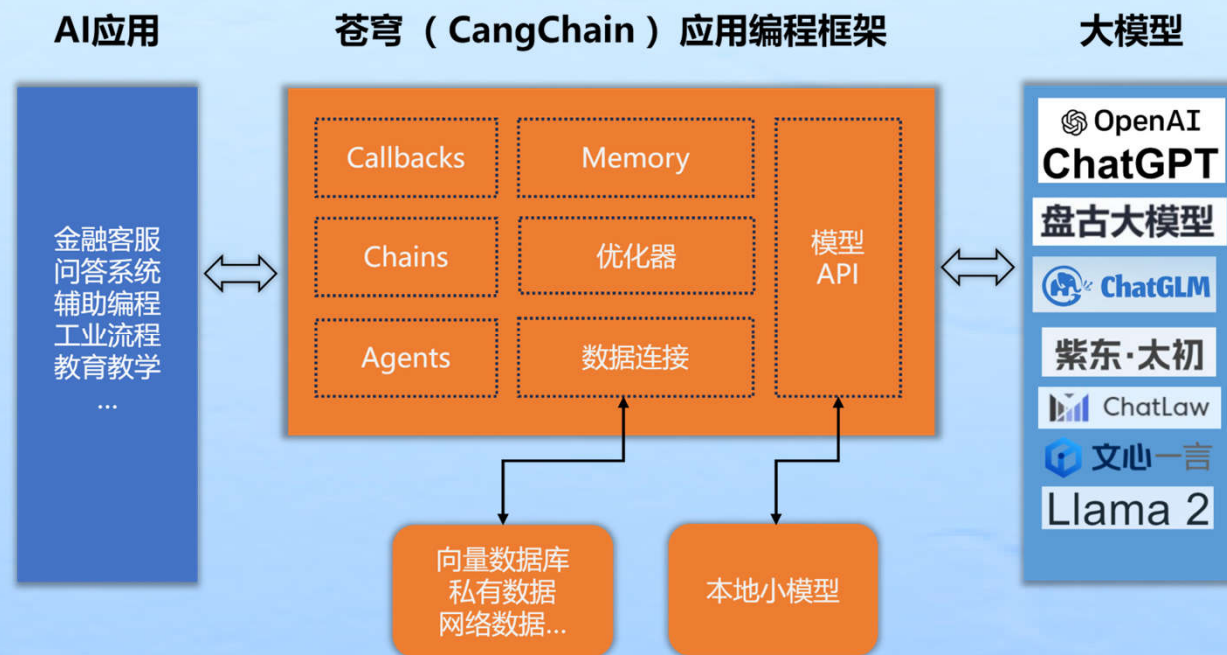
ChatLaw调用Tool
小模型、智能体多层检查

苍穹 (CangChain) 框架





苍穹用来形容广阔的天空、壮阔的景象，代表着壮阔、辽阔的意境。比如《诗经》中的“苍苍者天”，《庄子》中的“苍苍乎如在其上”的描述。苍穹常常被用来比喻高远的理想或抱负，也可以指代神话中的天空之神。







苍穹(CangChain)框架服务于软件厂商、模型厂商，帮助终端用户快速开发AI应用。



苍穹 (CangChain) 框架

<https://gitee.com/HW-PLLab/cangchain>

develop  cangchain / src 新建文件	
 zyb	删除文件 src/titoken/README.md 9d4f588 18小时前
...	
chain	feat: add SimpleQAChain as example
llmapi	code refactoring and add readme
localvd	add lib path
onnx	add lib path
schema	feat: add SimpleQAChain as example
titoken	删除文件 src/titoken/README.md
tool	google tool
vdclient	Add the basic structure of vdclient
Template.cj	init

titoken分词器	
 zyb	提交于 18小时前 已验证
 google tool	dongyu1009 提交于 1天前
 Add the basic structure of vdclient	xiaohanyuan 提交于 12天前
 feat: Chain & LLMChain & Memory	frankstein73 提交于 22天前
 feat: Chain & LLMChain	Dest1n1 提交于 23天前
 !init llmapi lib	gldong 编写于 2个月前

```
1 package chain
2 from cangchain import schema.BaseMemory
3 from std import collection.*
4 type Dict = HashMap<String, String>
5 public abstract class Chain {
6     let memory = None<BaseMemory>
7     public open func generate(inputs: Collection<Dict>): Collection<Dict>
8
9 > public open func preprocess(input:Dict): Dict { ...
15 }
16
17 > public open func postprocess(output: Dict): Dict { ...
20 }
21
22
23 > public func predict(input: Dict): Dict { ...
27 }
28
29 > public func predict(inputs: String): Dict { ...
33 }
34
35 > public func apply(inputs: Collection<Dict>): Collection<Dict> { ...
39 }
40
41 > public func apply(inputs: Collection<String>): Collection<Dict> { ...
45 }
46 }
```

苍穹 (CangChain) 框架

提示词优化

LangChain中ReAct是非常重要的技术
CangChain中ReAct也是核心技术。

```

from langchain.agents import load_tools
from langchain.agents import initialize_agent
from langchain.agents import AgentType
from langchain.llms import OpenAI
    
```

联邦学习

FL+CangChain

1. 利用CangChain来协调、帮助大模型的联邦学习(微调)。 **面向大模型**
假设一些用户(客户端)持有一定数量的问答数据，这些数据可能是用户自行标注的，具有专家经验的数据，且从未暴露给他人，他们希望获得更有任务针对性的大模型以便应用，但不想泄露自身数据隐私。
2. 结合联邦学习技术训练属于CangChain自身的小模型。 **面向CangChain本身的改进**
假设一些用户(客户端)持有一定数量的、与任务相关的数据，这些数据可能是用户自行标注的，具有专家经验的数据，且从未暴露给他人，CangChain与这些用户的交互还不够优秀，希望安全地利用这些信息来提高CangChain的易用性。
3. 使得CangChain支持联邦学习，开发基于仓颉的联邦学习框架。结合例如onnx等小模型库，支持包括但不限于大模型的联邦学习、拓展CangChain的应用、用大模型来帮助联邦学习..... **面向联邦学习**
假设一些用户(客户端)持有一定数量除文本问答数据以外的其它私有数据，他们不仅想与大模型交互，还想希望利用CangChain，通过联邦学习训练的其它模型，完成更多的任务和应用。

多语言支持

LangChain

- Python
- Javascript /TypeScript
- Java
- Go
- Rubby
- Dart

LangChain.js
LangChain Java

苍穹 CangChain

- CangJie
- Python
- Java
- go
- Javascript /TypeScript/ETS
- Wasm
- DSL

苍穹核心使用仓颉构建，目标是多语言支持但要整合在一个项目中。

- 利用仓颉跨语言调用的能力
- 基于类似gRPC做跨语言调用
- 框架中设计适配器或者虚拟层接收多语言接入调研中。。。

本地小模型

小模型的定义：
一个NLP的model，辅助CangChain，支撑联邦学习、提示词优化等。

```

graph TD
    LocalModel[Local Model] --> CangChain[苍穹 CangChain]
    CangChain <--> LargeLanguageModel[Large Language Model]
    
```

苍穹 (CangChain) 进展

调研任务

任务	状态	时间	责任人
Semantic kernel 的调研	完成	2023.7	葛煦旸、舒文韬
Titokens 调研	完成	2023.7	张玉斌
联邦学习的调研	完成	2023.7	张启贤
提示词优化、向量数据库的 client 和 sentence transformer 的调研	完成	2023.7	杨雨涵、袁肖瀚
Langchain 和 openAI 的 API 的一些调研	完成	2023.7	葛毅扬
多语言支持的调研	完成	2023.8	葛煦旸
Knowledge graph 的调研	进行中	2023.8	葛毅扬

设计任务

任务	状态	时间	责任人
Agents 的设计	完成	2023.7	葛煦旸、舒文韬
Chains 的设计	完成	2023.7	葛煦旸、舒文韬
Titokens 的设计	完成	2023.7	张玉斌
Chroma 向量数据库接口设计	完成	2023.7	杨雨涵、袁肖瀚

实现任务

任务	状态	时间	责任人
LLM 的 API 封装	完成	2023.8	董国良
Chain 的基类	完成	2023.8	葛煦旸
Agents 实现	进行中	2023.8	杨海龙、葛煦旸、舒文韬
Google 答案的 tool	完成	2023.8	董昱
Mathematica 的 tool	进行中	2023.8	董昱
向量数据客户端核心框架	进行中	2023.8	袁肖瀚
DuckDB 的 c 口封装	进行中	2023.8	葛毅扬
基于 OpenAI 的 webapi 的 sentence transformer	进行中	2023.8	杨雨涵
Titoken (BPE 编码) 库	进行中	2023.9	杨海龙、张玉斌
Word2vec 模型加载	进行中	2023.9	张玉斌
本地向量化库的 sentence transformer	未开始	2023.9	杨雨涵、袁肖瀚
Onnx 库	进行中	2023.9	杨海龙、张玉斌
关于 gRPC 的 agent 和 tool, 支持多语言	未开始	2023.10	杨海龙
向量的相似度计算库封装	进行中	2023.8	袁肖瀚

苍穹 (CangChain) 的未来规划

苍穹架构

大模型语言框架的调研和苍穹的功能、架构设计

23.7



苍穹MVP

Tokens库、向量数据接口库的设计和实现，输出MVP

23.8



23.9



23.10



23.12



2024



苍穹基础设施建设

LLM API库、Agents模块的设计和实现

苍穹样板应用

完成首款样板企业和应用的锁定

苍穹领域ISV

XX银行研发中心、金融领域ISV

WORKSHOP

Thanks!



添加“苍穹 (CangChain)
项目”微信群



WORKSHOP