Cosmos.ai agentic framework sharing

Tuesday, April 22, 2025 11:05

Tool/api
Sql
Pandas
Cosmos
Ccts cache

配一个public key（git)
Venv

Multi pod to deploy app
Dev env
Prod env: could also be visited
Pre-release: complex flow- could be copied or exported to prod env -download and upload and schema will be uploaded
The two env are isolated.
Emmbedding, preparation, prod

Multi instance

Revert; history

Enable Customized probes - k8s ability,
探针目前是http
Enable auth - enabled, url , could be visited by any user . Project level isolated, public, authorization - keymaker , header, client token, legal, 403 forbidden, onboarding to apply for client token (refer document)

Custom environment variables; deployment timeout

Orchestration: edm variable generation poc demo
Http llm function tool call

- Crew(employee, agent)
- task
- flow(an org with multi crew)
- Hierachical : more complex,
  - multi-task to be taken sequentially;
  - also agent, prefer default manager LLM, manager agent
- Sequential: 5 tasks, taken one by one
- Flow: connect multiple crew- how crew are sequentially organaized

Core: take tasks

Source rule/json
Target rule/json
To automatically generate jacket (jq expression?) by llm(hallucination)- validate - iterate, fail multi times, will hand to human
Generate_jq_task --> jq query validator(http client, to call http services to judge if the jq expression is correct)

Multiagent, with hierachical structure , so will be multiple times of llm call. So for complex tasks, llm will be requested for multiple times - kwargs max_iter: 2(最多call llm两次，不设置可能会call几十轮)

Release date 10min 30min

Release latency - 10min ~ 30min ； security；

Setup new pod
Cold start 2min

Every component could be performed individually. Debug using breakpoint
Hover : view output
没有命中milvus --> null
流量监测，回收
Exit debug mode
Custom component feature: prompt- write llm component prompt - write python code follow user guide, python schema.
Http request
Agentic ai

Flow - endpoint exposed to trigger - run flow - view code snippet, curl

Base image, iteration, select langflow base image: custom image- upgrade升级，每个版本改动会显示release note

Debug mode: could the prompt be seen? Like Langsmith, llm visibility. Build completed - > component --> online monitor: splunk log (持久化）& local log (不持久化). Prompt and other private data will not be logged/printed.

Agent configuration: same with crewai

- Agent: role, goal, backstory,
- Goal
- Role: JQ query validator

Tool: caculator, github API, HTTP Request API, Python REPL(Python REPL 就是Python 的交互式命令行界面，也就是说，当你在命令行输入 python 或 python3 并回车后进入的界面), --these are platform 内置tools

Manager llm: 4o
Ordinary llm: 3.5; open source llms

React agent, v1 version, -- > choose one llm, if it thinks it's not capable, than choose other llm. The task is not completed, will choose another llm to finish the task.

Framework upgrade, 兼容v1（some components in v1 are 内化 in v2)

Q: Raptor services; call codes; many codes are in base image; certifications and configurations are not in base image
Build Endpoint here,-- > connect

How two teams collaborate to use this framework agentai- user guide
Crewai framework investigation

Datadog traffic prod AIML cosmos.ai
Mcp client - mcp hub / mcp registery - visit recorded services
Poc mvp demo; mcp server will be integrated in the flow, will support more tools; custom component, validator, agent

Q: Prompt evaluation framework, prompt management. A set of pre defined QA, let llm to judge if the answer is appropriate.
A: no such feature at present. could tell the pm to schedule for the developers in cosmos team.

Q: llm, 4o? Or open source(gemini 1.0) gateway

Could use their sdk directly - gemini example code - India Miturn team, gemini gateway

Channel to consult

Agent mvp

调研crewai vs autogen，选择了crewai

展示每一步iterator，crewai 可以，定制化crewai不行

目前Langchain react，langraph也支撑自己一些东西

场景：analyst, investigator, 改一些条件/判断--> trigger

Evaluate

Leverage platform, monitor, k8s, capability.

1st version : Byoc -> cosmos