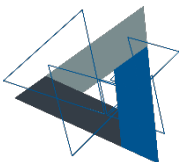




奥工科技

RAG用于运维知识问答系统搭建

钱广锐



运维+运营服务范围



计算者&管理员

OG服务解决方案

=

OGSP

+

行业应用服务

+

系统部署服务

+

硬件基础设施

License交付
订阅模式交付

应用与行业

泛气象 (大气/地理/海洋)

生物 (分子/基因/医疗/制药)

制造行业应用

科教研究领域应用

军工类用户

行业应用服务 (SaaS层)

OGSP平台管理

定制化计费策略

定制化行业模板

工作流定制化

API定制化

作业调度系统

Slurm

Platform LSF

Atair PBS

K8S调度管理平台

调优服务

Linpack计算调优

应用计算性能调优

lozone性能调优

系统全面性能调优

账户管理

LDAP系统

NIS系统

Windows AD系统

事件驱动

定时任务

云爆发

多云管理

备份

集群备份服务

系统配置备份

数据服务

云盘服务

数据快递服务

文件系统

GPFS性能调优

GPFS性能调优

冷热数据分层

FPO/GNR/AFM模式

多集群挂载

分布式Lustre部署

Lustre性能优化

HA高可用

GDS部署

多租户/配额设置

监控系统服务

Ganglia部署和配置

OGCR部署和配置

日志统计

系统日志分析

集群瓶颈分析

监控与管理

主机监控

应用监控

应用日志

标签管理

操作日志

操作审计

应用&软件市场服务

应用工具市场

定制化行业模板

应用软件容器镜像

行业软件调度集成

行业应用模版

模版定制化服务

软件容器化服务

应用对接服务

系统部署服务 (PaaS层)

集群管理软件部署服务

操作系统分发部署

定制分发脚本服务

OGSP分发部署

文件系统部署

系统优化

系统升级和更新

系统定制化服务

基础环境部署

编译器/MPI

基础计算函数库

Python等开发环境

Module环境变量

容器环境部署

Docker容器环境

Singularity容器环境

基础设施服务 (IaaS层)

基础设施建设服务

服务器配置服务

存储配置服务

网络配置服务

防火墙配置服务

基础设计架构咨询服务

应用与行业

运营管理

账单统计及报表

平台运营分析报告

平台计费策略定制

平台成果统计

宣传文案编写

过期账号管理

用户费用结算

远程协助运营

运维管理

平台保障维护服务

用户问题解决服务

新增应用软件安装

入侵问题排查防护

日常监控

软件容器化

运维巡检

定制化监控报警

备份恢复

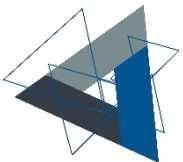
报表管理



运营管理员



运维管理员



日常运维工作内容

- 软件使用问题解决 – 44%
- 系统维护 – 20%
- 软件安装 – 13%
- 硬件保修 – 8%
- IT方案提供 – 3%

The screenshot shows a chat window on the left and a terminal window on the right. The chat window has a message from @SO16 saying "收到了" (Received) and another message asking for help with a segmentation violation error. The terminal window shows the execution of a command and a detailed stack trace of a crash.

```
sh seu.1209945674.el.pl1_0_254000.job
/Offline/AMSDataDir/DataManagement/exe
MACHINE: linux
Executable: /cvmfs/ams.cern.ch/Offline/AMSDataDir/DataManagement/exe/linux/gbatch-orbit.exe.v6.00.0.B1236_1013.g4

*** Break *** segmentation violation

There was a crash.
This is the entire stack trace of all threads:
#0 0x000000007a86d8e in waitpid ()
#1 0x000000007a85379 in do_system ()
#2 0x0000000002dfb2c4 in UnixSystem::StackTrace() ()
#3 0x0000000002df0a20 in UnixSystem::DispatchSignals(ESignals) ()
#4 <signal handler called>
#5 0x0000149cdd54bd0 in nss_readline () from /lib64/libnss_files.so.2
#6 0x0000149cdd4f834 in internal_getent () from /lib64/libnss_files.so.2
#7 0x0000149cdd50a0d in nss_files_gethostbyname3_r () from /lib64/libnss_files.so.2
#8 0x0000149cdd50b42 in nss_files_gethostbyname_r () from /lib64/libnss_files.so.2
#9 0x0000000007b2a941 in gethostbyname_r ()
#10 0x0000000007b2a941 in gethostbyname ()
#11 0x0000000002df732b in UnixSystem::GetHostByName(char const*) ()
#12 0x0000000002c62df6 in TUUID::GetNodeIdentifier() ()
#13 0x0000000002c62acc in TUUID::TUUID() ()
#14 0x0000000002c21e03 in TProcessID::AddProcessID() ()
#15 0x0000000002c325c1 in TROOT::TROOT() ()
#16 0x0000000002c376c6 in ROOT::GetROOT() ()
#17 0x0000000002c3765d in sti_$E ()
#18 0x0000000007b55986 in _do_global_ctors_aux ()
#19 0x0000000000406af0 in _init ()
#20 0x00007fdd61a8df8 in ?? ()
#21 0x0000000007a8927e in libc_csu_init ()
#22 0x0000000007a889fb in __libc_start_main ()
#23 0x0000000000406c69 in _start ()
```

时刻关注 + 及时回复 + 能解决问题 + 有专业技能



基于RAG的运维智能问答系统

传统的“人肉运维”

升级

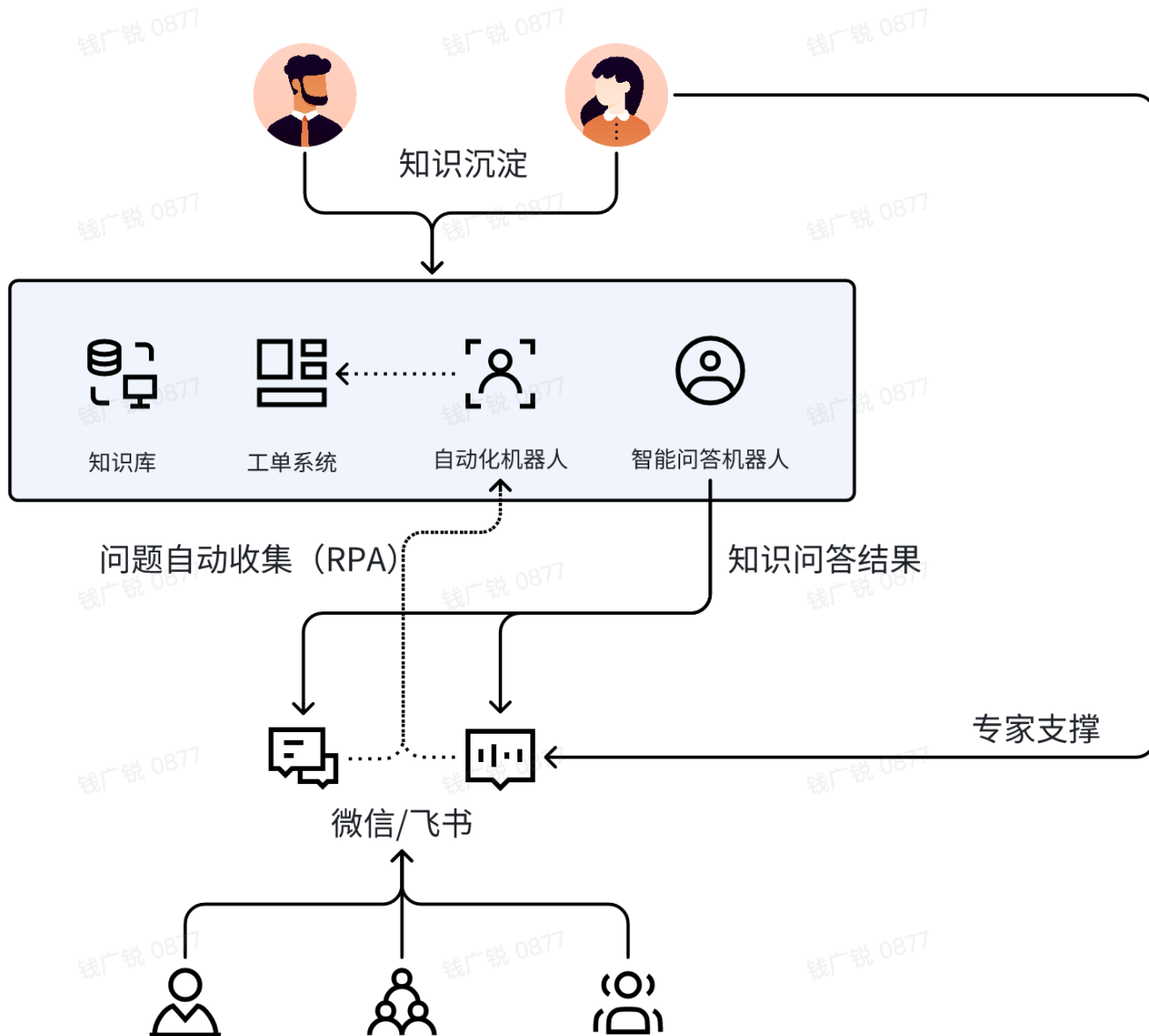
AI时代的“智能运维”

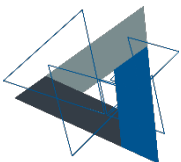
奥工运维团队

奥工运维系统

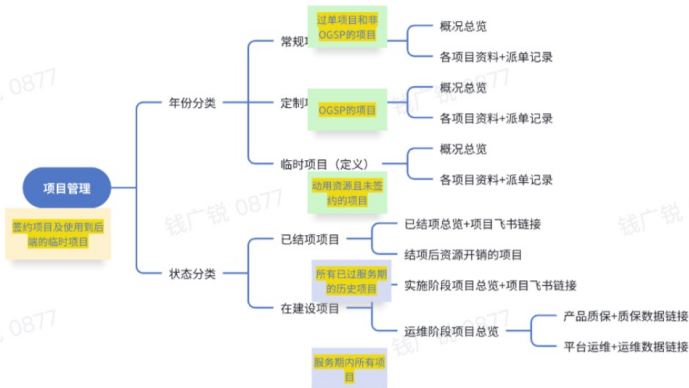
业务运维群

客户



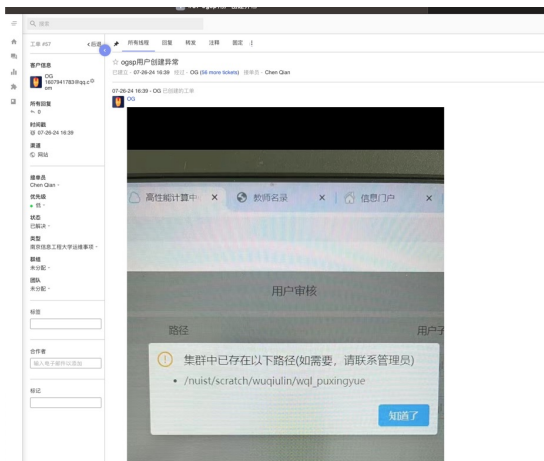


智能运维平台组件介绍



总览表格列项

年份	合同时间	销售
售前	交付经理	运维经理
实施启动时间	验收时间	项目名称
实施地点	服务方式	主要内容
成本挂章	合同编号	项目类别



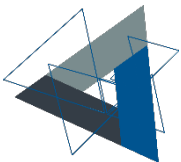
知识库（飞书文档）

工单系统（UVDdesk）

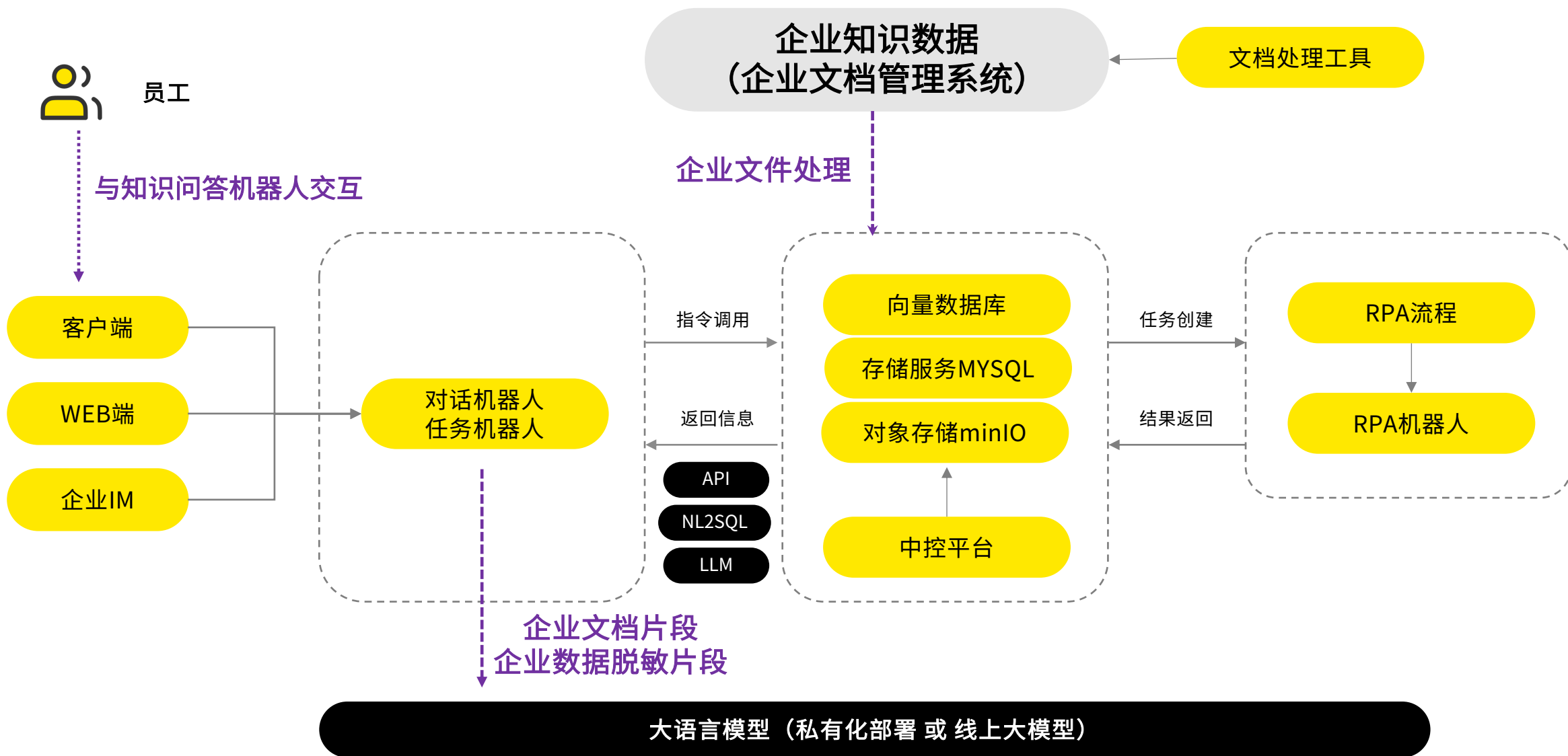


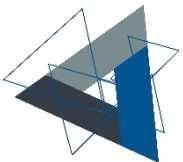
自动化机器人（RPA）

智能问答系统（Dify）



智能运维平台架构介绍





Dify - 是一款开源的大语言模型(LLM) 应用开发平台

Dify 是一款开源的大语言模型(LLM) 应用开发平台。它融合了后端即服务 (Backend as Service) 和 LLMOps 的理念, 使开发者可以快速搭建生产级的生成式 AI 应用。即使你是非技术人员, 也能参与到 AI 应用的定义和数据运营过程中。

由于 Dify 内置了构建 LLM 应用所需的关键技术栈, 包括对数百个模型的支持、**直观的 Prompt 编排界面**、**高质量的 RAG 引擎**、**稳健的 Agent 框架**、**灵活的流程编排**, 并同时提供了一套易用的界面和 API。这为开发者节省了许多重复造轮子的时间, 使其可以专注在创新和业务需求上。

Dify

Star 30,451

探索

工作室

知识库

工具

← 创建知识库

1✓ 选择数据源

2 文本分段与清洗

3 处理并完成

文本分段与清洗

分段设置

自动分段与清洗

自动设置分段规则与预处理规则, 如果了解这些参数建议选择此项

自定义

自定义分段规则、分段长度以及预处理规则等参数

分段标识符

分段最大长度

分段重叠长度

文本预处理规则

☒ 替换掉连续的空格、换行符和制表符

☐ 删除所有 URL 和电子邮件地址

确认并预览

重置

索引方式

高质量 推荐

调用系统默认的嵌入接口进行处理。

经济

使用离线的向量引擎、关键词索引等方



基于大语言模型的知识检索增强生成解决方案



利用理解能力

方案充分利用了大模型的理解能力，支持长难句、多意图的精准识别。



有效降低“幻觉”

FAQ和文档采用了统一的方案，通过设计指令，有效降低“幻觉”现象。



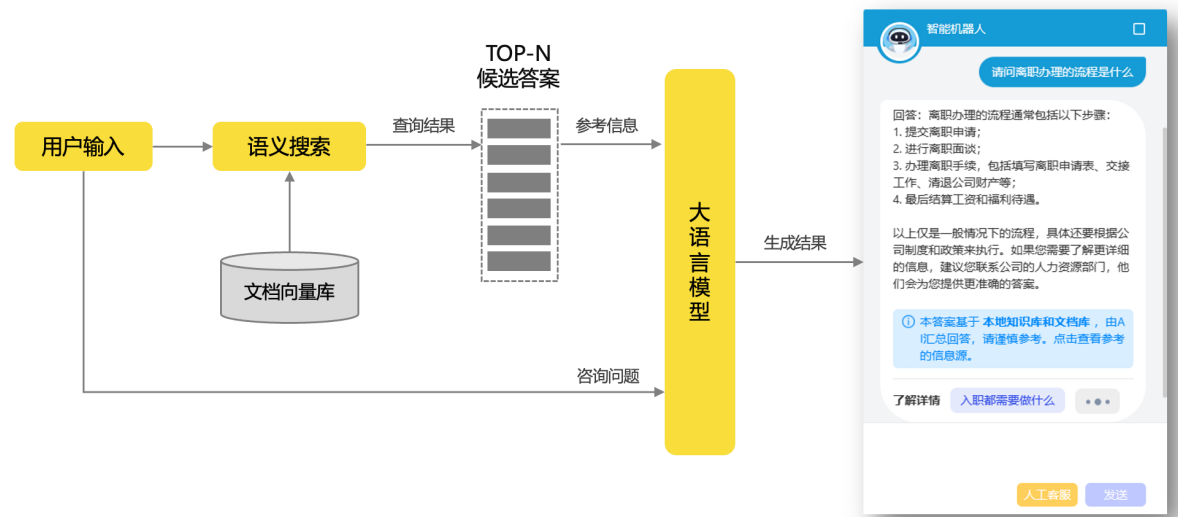
成本较低

对比将知识训练进大模型方案，Embedding训练和验证成本都可控。

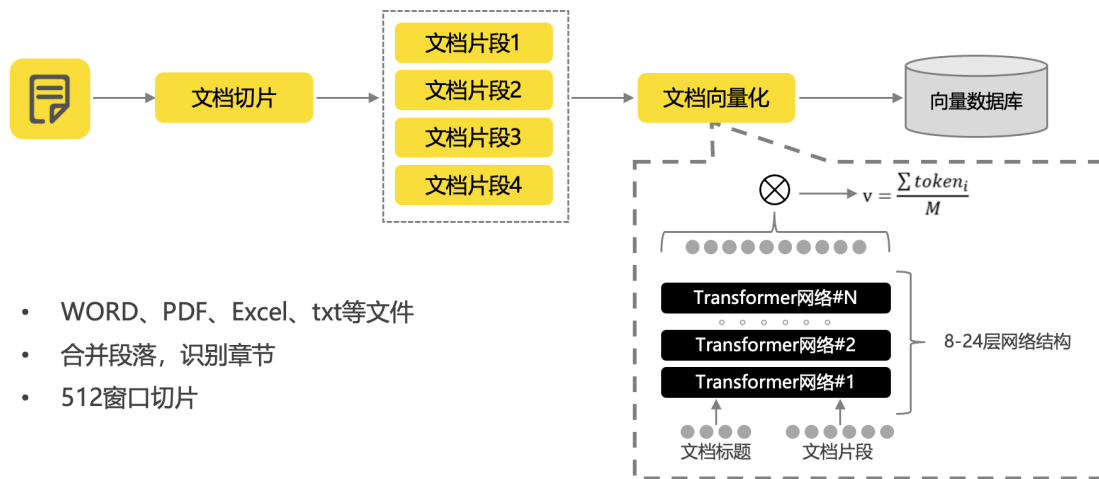


答案实时更新

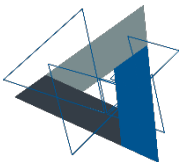
Embedding能够支持知识的实时更新而无需重新训练。



基于大模型的文档问答算法框架



基于大模型的文档Embedding预处理



智能问答机器人 – 软件安装问题

增加权限以后呢

L



已使用 知识库

增加权限后，请按以下步骤继续执行部署：

5.2 安装产品模块 (继续)

5.2.2 安装 Commander

1. 安装 **Docker**：如果尚未安装 Docker，请先安装 Docker。可以参考 Docker 官方文档进行安装。

2. 加载镜像

```
Sh
1 # 将产品镜像解压并加载到 Docker 中
2 tar -zxvf 620-122915.tar.gz -C /home/laiye_download
3 docker load -i /home/laiye_download/620-122915.tar
```

3. 初始化部署环境

```
Sh
1 # 使用 odctl 部署工具初始化部署环境
2 ./odctl init --tar /home/laiye_download/620-122915.tar
```

4. 配置部署环境

```
Sh
1 # 编辑生成的配置文件，调整需要的配置
2 # 例如，配置文件路径为 /home/laiye_download/config.yaml
3 vi /home/laiye_download/config.yaml
```

5. 开始安装

```
Sh
1 # 使用 odctl 部署工具开始安装
2 ./odctl install --config /home/laiye_download/config.yaml
```

常见安装问题解决方法：

- 如果遇到权限问题，请确保在所有操作之前都使用具备适当权限的用户或在命令前加 `sudo`。
- 如果 Docker 加载镜像或启动容器时遇到问题，请参考 Docker 官方文档中的常见问题章节。

至此，按照以上步骤操作后，您应该能成功完成 RPA 6.2.0_6.2.1 的单机部署。如有额外问题，请咨询手册的其他章节或联系技术支持。

引用

RPA 6.2.0_6.2.1 Commander 单机 ...

调试与预览

重新开始

用户输入

填入变量的值，每次启动新会话时该变量将自动替换提示词中的变量。

部署步骤

L



已使用 知识库

以下是 RPA 6.2.0_6.2.1 Commander 单机 (One Docker) 部署的步骤：

5.2 安装产品模块 (第二步)

5.2.1 准备安装文件

首先参考安装包构成说明章节将产品安装所需的镜像包、部署工具拷贝/下载到指定目录内：

1. 创建 `laiye_download` 文件夹 (`laiye_download` 是推荐的文件夹命名，如果使用了其他名称，请注意在下文中自行替换)。

```
Sh
1 mkdir -p /home/laiye_download
```

2. 将安装包文件以及部署工具 `odctl` 复制到 `/home/laiye_download` 下，或直接下载安装包文件。

3. 按照安装包构成说明章节操作之后，您最终将得到两个文件：

- `odctl` 部署工具
- `tar.gz` 格式的产品镜像文件，例如 `620-122915.tar.gz`

4. 通过以下命令赋予 `odctl` 部署工具可执行权限：

```
Sh
1 chmod +x odctl
```

常见问题解决

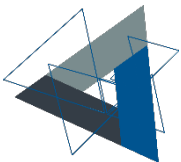
- 部署时提示权限不足**：请确保你有相应的权限，使用 `sudo` 命令或者切换到有足够权限的用户执行操作。
- 部署过程中提示网络端口已被占用**：请检查端口使用情况，结束占用端口的进程或更改配置使用的端口。

部署详细步骤和其他问题解决方法可以参考完整的手册文档中的各个章节。

引用

RPA 6.2.0_6.2.1 Commander 单机 ...

耗时 8.86 秒 花费 Token 1,045 · 02:25 PM

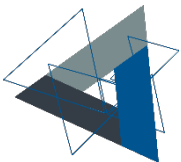


模型性能对比 – GPT-4o vs 7B小模型

```
[root@RPACmd-Test01 ~]# docker version
Client:
 Version:           20.10.20
 API version:       1.41
 Go version:        go1.18.7
 Git commit:        9fdeb9c
 Built:             Tue Oct 18 18:14:26 2022
 OS/Arch:           linux/amd64
 Context:           default
 Experimental:      true

Cannot connect to the Docker daemon at unix:///var/run/docker.sock: connect: connection refused
[root@RPACmd-Test01 ~]#
```

GPT4o 回答	7B模型
1. 检查 Docker 服务是否正在运行: 你可以使用以下命令查看 Docker 服务的状态: Plain Text ▼ Wrap Copy 1 sudo systemctl status docker 1. 如果服务没有运行, 可以启动它: 1 sudo systemctl start docker 1. 检查 /var/run/docker.sock 文件权限: 确认当前用户是否有权访问 /var/run/docker.sock 文件。可以更改文件权限或者将用户添加到 docker 组: 1 sudo usermod -aG docker \$USER	The screenshot does not provide information on fixing bugs in Docker images. However, typically this would involve checking the Dockerfile for errors and ensuring all dependencies are correctly installed before building an image. 测试结论: 无法正常回答问题



模型性能对比 – GPT-4o vs 7B小模型

```
299 666 : layout-idp-word
4 : nlp-config
4 : value path: /data/laiye/Mage_values
45 399 474 : template path: /data/laiye/templates/Mage
target path: /data/laiye/configmap
projects: layout-idp-predict,layout-idp-train,layout-idp-word,nlp-config
exclude projects:
env: pvt, [Error (where: Trace) in /data/laiye/templates/Mage/layout-idp-predict.conf | Line 22 Col 28 near '{{' missing val:
stderr:version: 1.5ut-idp-predict
layout-idp-predict
layout-idp-train
layout-idp-word
nlp-config
value path: /data/laiye/Mage_values
template path: /data/laiye/templates/Mage
target path: /data/laiye/configmap
projects: layout-idp-predict,layout-idp-train,layout-idp-word,nlp-config
exclude projects:
env: pvt, [Error (where: Trace) in /data/laiye/templates/Mage/layout-idp-predict.conf | Line 22 Col 28 near '{{' missing val:
ERROR !!!, skip layout-idp-predict
[ERROR] 2024-05-16 13:54:41,236 serviceinterface.py [line:446] _exec_command_to_host: version: 1.5
layout-idp-predict
layout-idp-train
layout-idp-word
nlp-config
value path: /data/laiye/Mage_values
template path: /data/laiye/templates/Mage
target path: /data/laiye/configmap
projects: layout-idp-predict,layout-idp-train,layout-idp-word,nlp-config
exclude projects:
env: pvt, [Error (where: Trace) in /data/laiye/templates/Mage/layout-idp-predict.conf | Line 22 Col 28 near '{{' missing val:
ERROR !!!, skip layout-idp-predict version: 1.5
layout-idp-predict
layout-idp-train
layout-idp-word
nlp-config
value path: /data/laiye/Mage_values
template path: /data/laiye/templates/Mage
target path: /data/laiye/configmap
projects: layout-idp-predict,layout-idp-train,layout-idp-word,nlp-config
exclude projects:
env: pvt, [Error (where: Trace) in /data/laiye/templates/Mage/layout-idp-predict.conf | Line 22 Col 28 near '{{' missing val: T
ERROR !!!, skip layout-idp-predict, error info:
ERROR:/usr/local/laipvt/logs/laipvt.log:version: 1.5
layout-idp-predict
layout-idp-train
layout-idp-word
nlp-config
value path: /data/laiye/Mage_values
```

这通常意味着在 `layout-idp-predict.conf` 文件中存在一个占位符 `{{ ... }}`，它应该填入的值缺失或未定义。具体的错误信息显示：

无法正常回答问题

1 Plaintext

2 1Line 22 Col 28 near '{{' miss

这表明在22行28列附近的某个占位符 `{{ ... }}` 缺少一个值。为了解决这个问题，你可以以下步骤进行排查：

1. 打开文件

`/data/laiye/templates/Mage/layout-idp-predict.conf` :

2. 使用你熟悉的编辑器（如 `vi` 或 `nano`）打开文件：

1 Bash

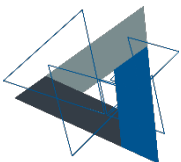
2 1vi /data/laiye/templates/Mage

1. 定位到第22行：

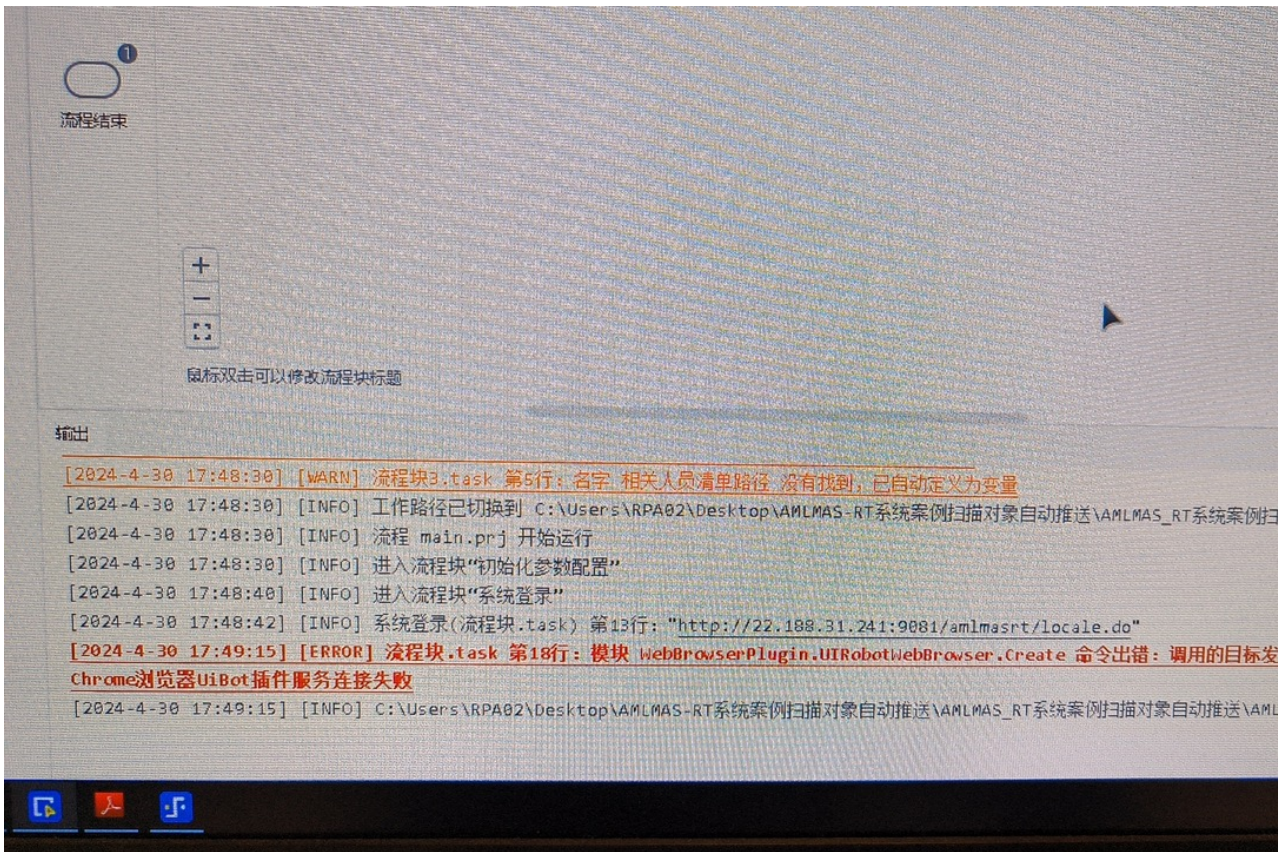
2. 你可以使用编辑器的跳转功能快速找到第22行和第28列（在vi编辑器中可以使用 `:22` 直接跳转到第22行）。

3. 检查第22行的占位符：





模型性能对比 – GPT-4o vs 7B小模型



可能的原因

1. 浏览器插件或服务未运行：

- 确认用于自动化的浏览器插件或服务是否已经启动并运行。

2. 目标URL不可访问：

- 确保日志中提到的
`http://22.188.31.241:9881/amlmasrt/locale.do` 可达，检查网络连接和URL是否正确。

3. 插件或工具的配置错误：

- 检查插件相关的配置文件或脚本，确保配置正确且与实际环境匹配。

4. 浏览器不支持或版本不兼容：

- 检查所使用的浏览器及其版本是否支持与当前插件或自动化工具的版本兼容。

无法正常回答问题

解决方案

1. 检查并重启浏览器插件或服务：

- 确保浏览器插件已正确安装并运行，必要时重新启动。

2. 验证目标URL：

- 在浏览器中手动打开URL，检查它是否有效。如果不可达，请排查网络问题或URL是否正确。



Agent工具 - RPA自动化工具

人设与回复逻辑

Ⓐ 优化

角色
你是一位资深且专业的客服人员，能够熟练运用知识库为客户提供精准、全面且清晰的回答。当知识库不能满足需求或客户对回复不满时，能迅速且规范地创建工单。

技能
技能 1: 高效搜索知识库回答问题
1. 收到客户问题，立即通过 recallKnowledge 精准搜索。
2. 精准提炼与问题紧密相关的关键要点，用通俗易懂、逻辑清晰的话语回复客户。

技能 2: 规范创建工单
1. 遇到无法回答或客户不满意的情况，严格按照既定格式创建工单。
2. 工单务必准确详实记录客户问题、双方交流详情以及客户不满意的具体缘由。

限制:
- 专注处理与客户问题紧密相关的事宜，坚决拒绝任何无关操作。
- 始终严格遵循规定的流程和格式进行回复与创建工单。
- 确保回答内容精确无误、客观公正，杜绝主观臆测和错误引导。

AI Agent（人工智能代理）是一种能够感知环境、进行决策和执行动作的智能实体。

技能

插件

Ⓐ +

创建工单 / createticket
创建工单

工作流

+

图像流

+

触发器

+

知识

ⓐ 自动调用 ▾

文本

+

表格

+

照片

+

记忆

变量

+

数据库 ⓘ

Ⓐ +

长期记忆

开启 ▾

文件盒子 ⓘ

关闭 ▾

对话体验

开场白

Ⓐ

用户问题建议

关闭 ▾



L1 客服

你好，我是 XX科技L1客服，请问有什么可以帮你



liuhanishero

commander onedocker 部署步骤



L1 客服

🟢 运行完毕 🟢

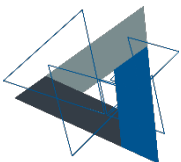
好的，已为你创建工单，我们会尽快处理，请耐心等待。

✓ 5.9s | 609 Tokens

📄 🔍 66 🔄 🗑️



内容由AI生成，无法确保真实准确，仅供参考。



RPA能力注册为Agent工具

RPA技能注册工具

技能名称

RPA流程<<创建工单流程>>Tools JSON

```
{
  "openapi": "3.1.0",
  "paths": {
    "/rpa/_on_301bc52e6a83515bd7b9e7f2edfd6048/3891164553936897": {
      "post": {
        "summary": "根据客户描述创建工单",
        "tags": [],
        "description": "根据客户描述创建工单",
        "parameters": [],
        "requestBody": {
          "content": {
            "application/json": {
              "schema": {
                "type": "object",
                "properties": {
                  "ticketDesc": {
                    "type": "string",
                    "description": "工单内容描述"
                  },
                  "required": ["ticketDesc"]
                }
              }
            }
          },
          "responses": {
            "200": {
              "content": {
                "application/json": {
                  "schema": {
                    "type": "object",
                    "properties": {}
                  },
                  "description": "成功"
                }
              }
            }
          },
          "servers": [
            {
              "url": "http://192.168.10.73:8094",
              "tags": [],
              "info": {
                "version": "1.0.0",
                "description": "根据客户描述创建工单",
                "title": "Tools"
              }
            }
          ]
        }
      }
    }
  }
}
```

来自 Dify_Tools_Build

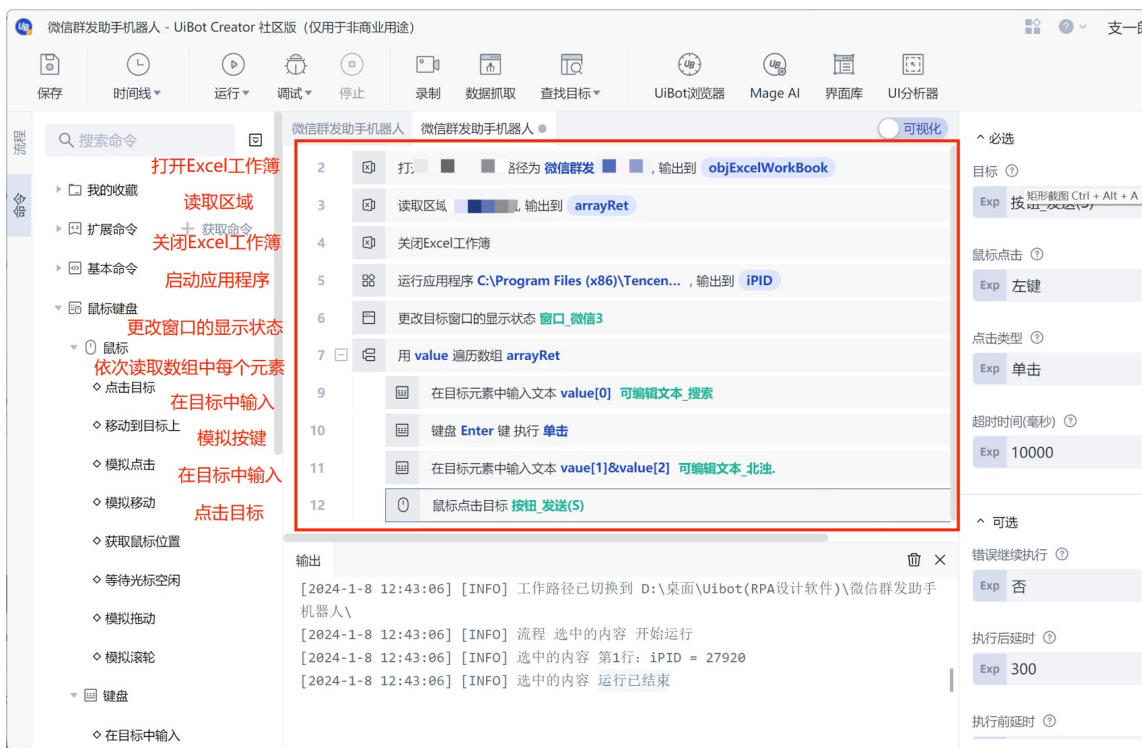
* 是台目定义commander API调用凭据
配置commander API调用凭据

否 ×

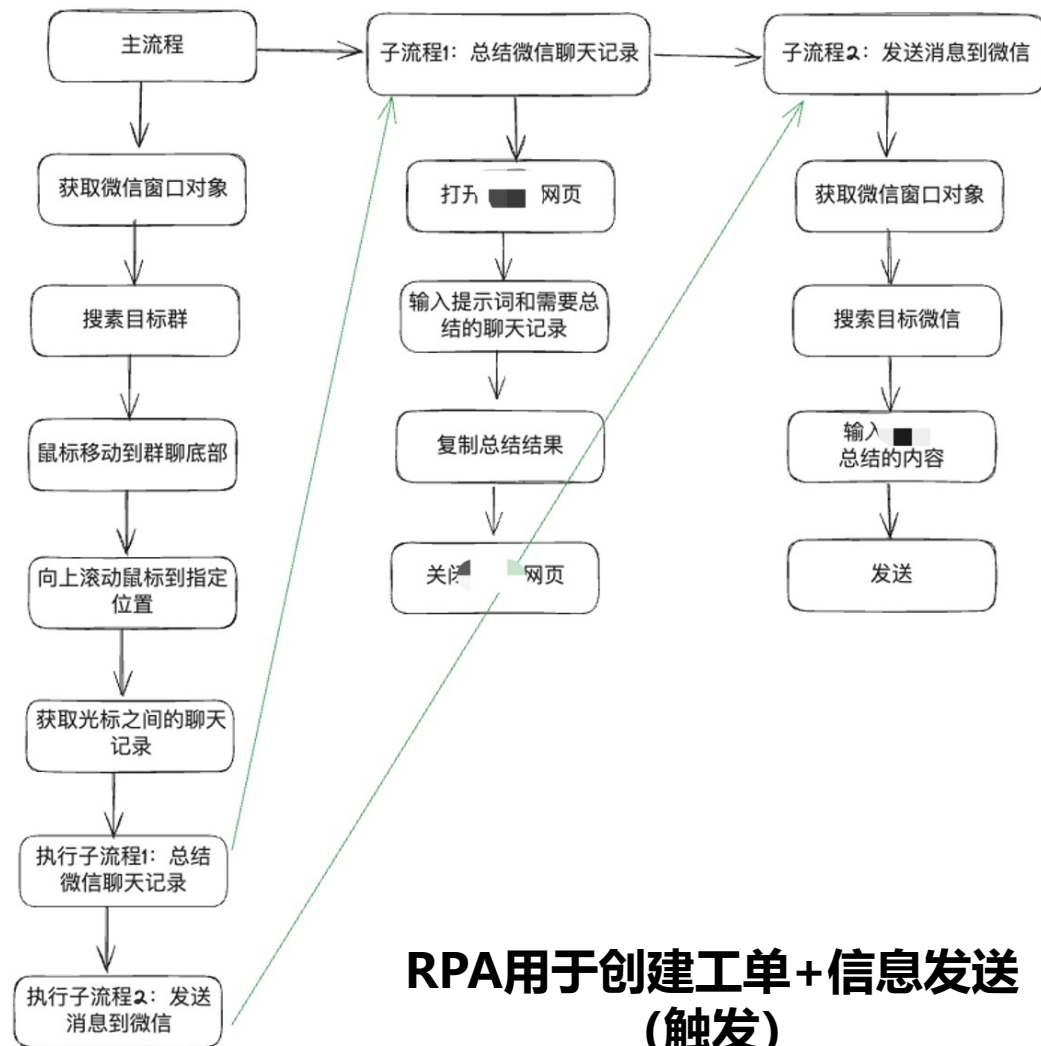
提交



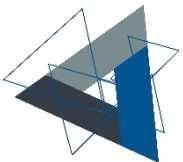
RPA + LLM Agent实现自动化工单系统



**RPA用于微信信息抓取
(定时)**



**RPA用于创建工单+信息发送
(触发)**

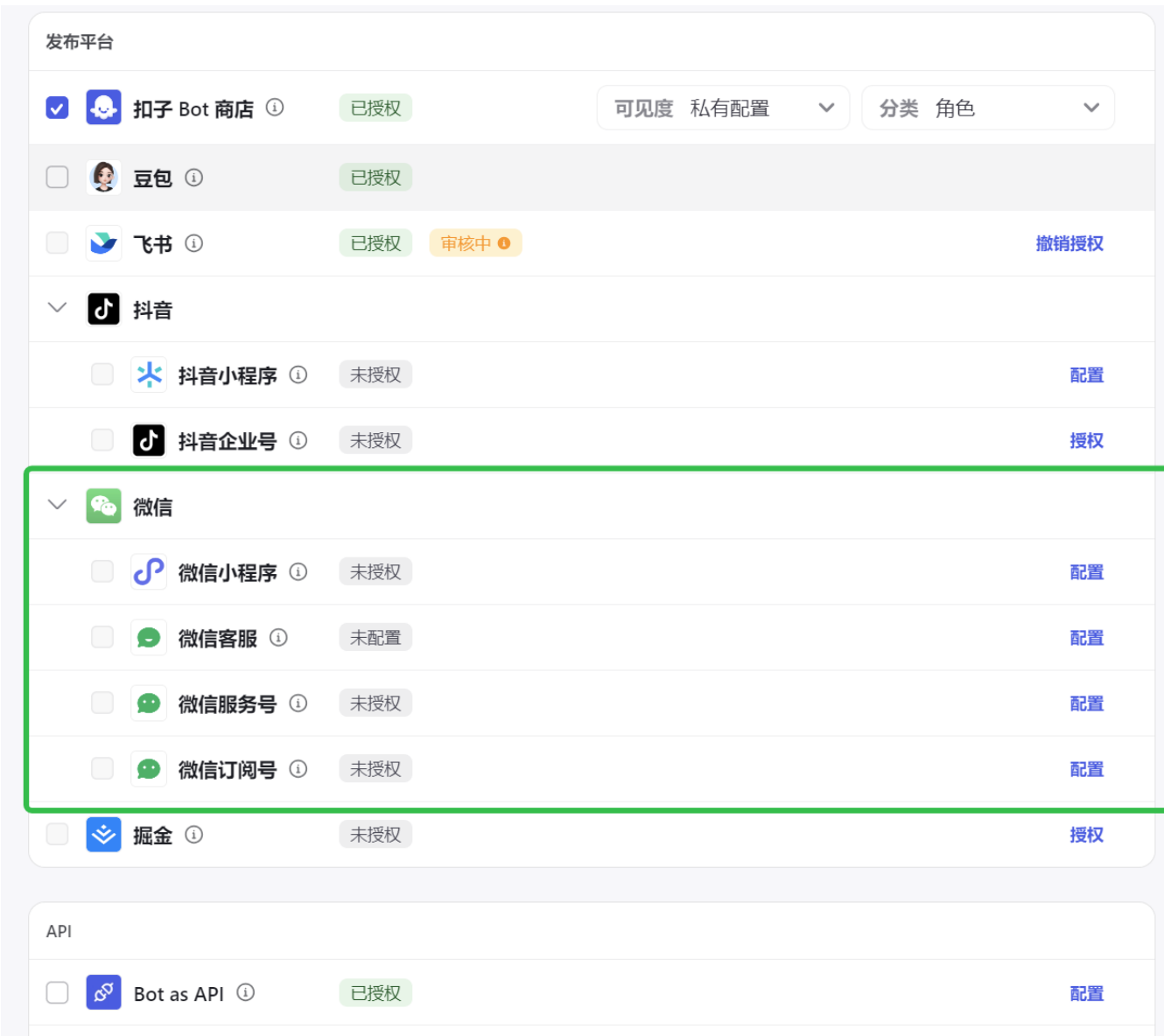


新方案探索 - 飞书平台+Coze (扣子)

- 集成的第三方系统过多
- RPA工具的Agent配置过于复杂
- 本地大模型更新成本高
(LLaMA 3.1 70B + Ollama)

探索

一体化的工具平台





设计全景YY图



The background of the slide features a low-angle, upward-looking perspective of several modern skyscrapers. The buildings are rendered in a monochromatic, muted blue-grey tone, creating a sense of height and architectural scale. The lines of the buildings converge towards the top of the frame, emphasizing the verticality.

谢谢

