

企业级智能体的落地实践 (DAP)

张振宇

数字化进阶：从数据沉淀到智能决策

企业数字化的核心诉求

企业在数字化转型中积累了海量数据，但这些数据大多处于沉睡状态。企业急需将这些数据转化为实时、精准、可量化的决策支持，以提升业务效率和竞争力。

01

传统BI的局限性

传统BI工具只能呈现历史数据结果，缺乏动态预测和实时建议的能力。企业需要更智能的工具，将数据转化为可执行的洞察，实现从数据到知识再到决策的闭环。

02

智能体的崛起

智能体通过大模型技术，能够主动分析数据，提供实时预测和建议，帮助企业实现智能化决策，提升业务效率和竞争力。

03

传统智能化落地痛点

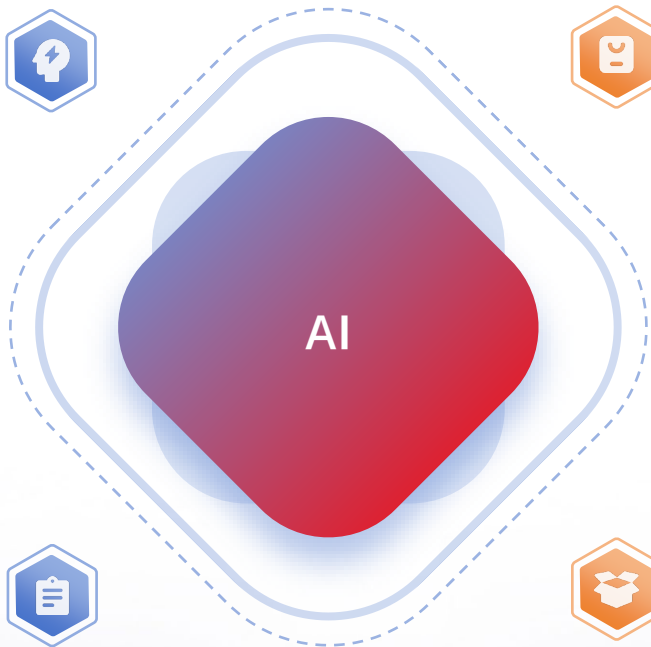
逻辑复杂
模型调用、工具链对接、流程编排等



定制化程度高
业务逻辑不可复制



迭代成本高
新模型适配、新数据中台对接



测试验证复杂

模型输出的不确定性、概率性，需持续回归测试



安全合规标准严

企业与监管机关的强审计



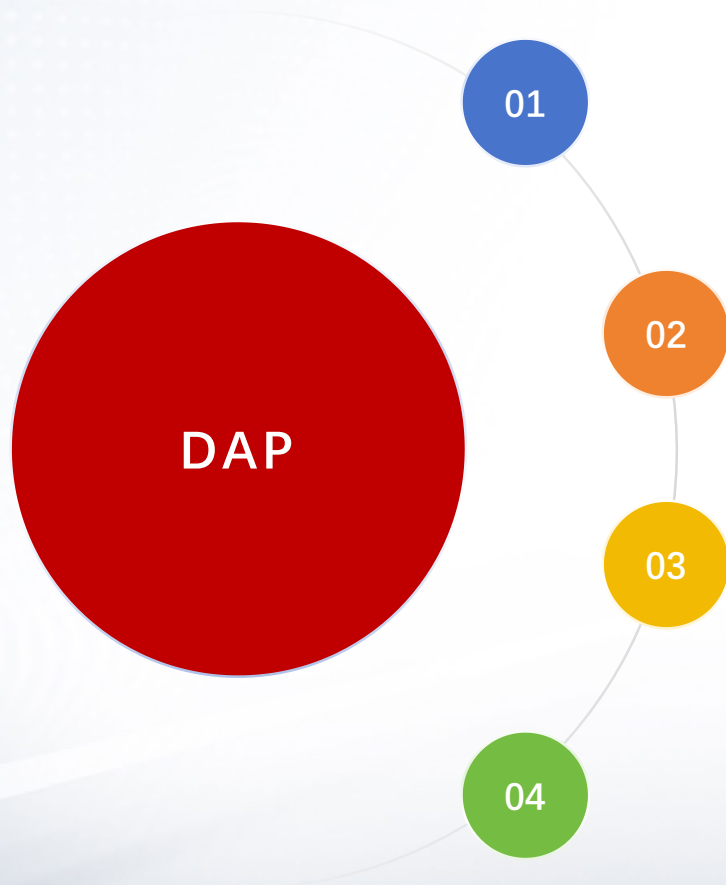
人才需求多

高质量智能体需要跨学科型人才



DAP的产品定位

帮企业跨越智能体落地“最后一公里”，以低门槛、高可靠、强适配特性，让 AI 技术快速转化为实际业务价值。



低门槛开发

- 可视化拖拽编排 workflow，无需编程基础搞定复杂业务逻辑
- 自动化数据预处理（多格式导入、自动清洗 / 向量化 / QA 分割），节省 80% 训练时间
- 模板生态 + 开箱即用功能，快速搭建 AI 客服、知识库问答等应用

高灵活适配

- 多模型兼容：支持 Deepseek、Qwen、文心一言等主流 LLM，适配自定义向量模型
- 全场景集成：API 无缝对接微信、飞书、公众号等平台，无需重构底层架构
- 部署模式：支持私有化部署（保障数据安全）。

强可靠输出

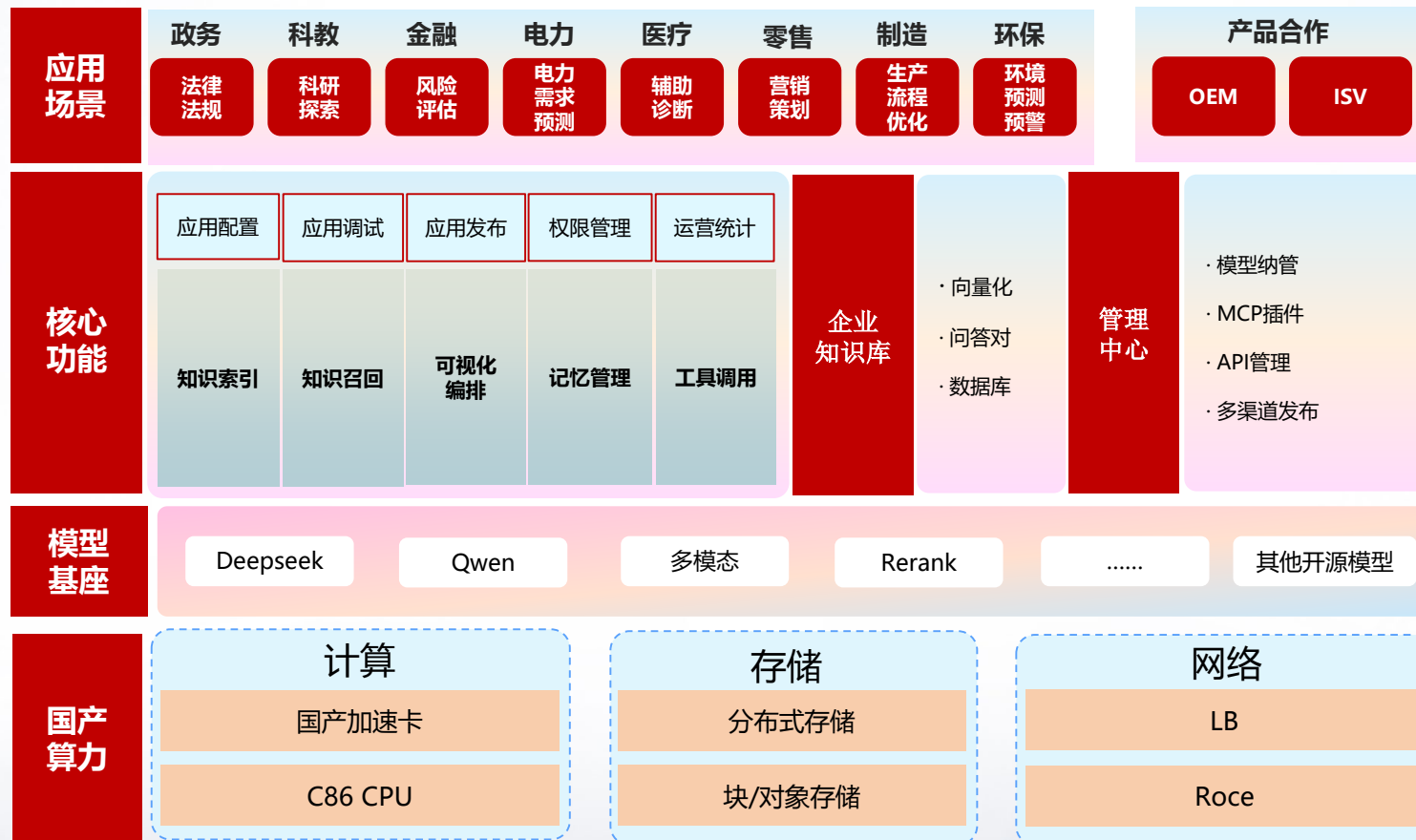
- 创新 RAG 检索 + 独特 QA 结构，生成带引用来源的答案，大幅减少 AI “幻觉”
- 多维度调试工具（搜索测试、全局变量、调试预览），确保输出可控可优化
- 代码可控（全自研），数据处理透明可追溯，满足合规与安全需求

低成本落地

- 无需专业 AI 人才，降低人力与技术投入成本
- 支持二次开发与无限扩展，API 定制无需修改源码，适配业务增长
- 覆盖医疗、金融、制造等数十个行业，开发者生态持续丰富场景

DAP核心功能矩阵

□ 微服务架构设计，混合索引技术，多模型兼容

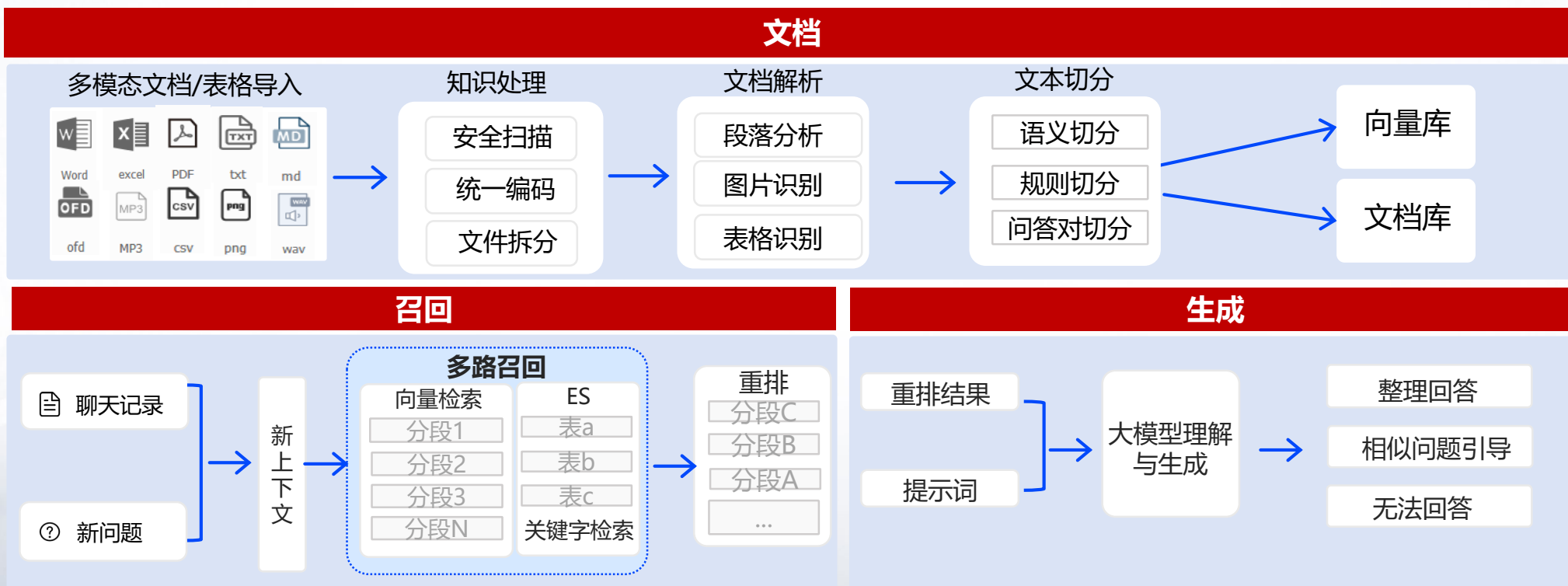


DAP核心能力-知识融合增强

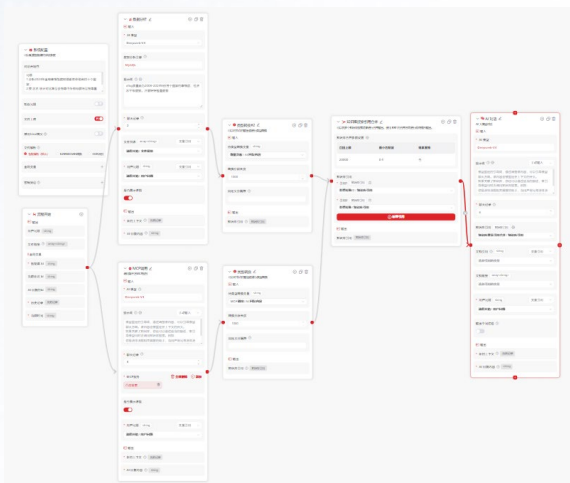
从知识库中检索相关数据，再由大模型基于检索结果生成更准确、可靠的输出

关键节点

- **知识导入**：支持多类型文档导入（.docx/.pdf/.txt/.md/.pptx/.xlsx/.csv/.ofd/.mp3/.wav/.m4a/.flac等）问答对导入。
- **解析切分**：在解析阶段对文档进行版面分析、元素提取。在切分阶段将长文本分割为短文本分段。
- **知识召回**：将用户问题转化为向量，与向量库中文档分段进行相似度匹配，再与原始文档分段一同召回。
- **理解与生成**：大模型根据检索内容总结生成答案。



DAP核心能力-智能体应用构建



丰富的组件库

内置大量开箱即用的功能节点：

基础交互：用户输入解析、上下文管理、多轮对话控制；知识调用：知识库检索、文档提取、相关度排序；工具集成：API 调用、数据库查询；逻辑处理：条件判断、循环执行；内置常用智能体模版，可直接微调复用，减少重复开发。

零代码可视化

提供拖拽式工作流编排界面，用户无需编写代码即可构建复杂智能体。

支持对话流、插件调用、用户交互节点等模块，像“积木”一样组装 AI 应用。

可观测可优化

全链路监控与日志：记录智能体的每一次交互与执行过程，包括：用户输入、节点调用顺序、工具返回结果、模型输出、异常触发点等，通过可视化关键指标，便于开发者快速定位问题。

持续迭代机制：支持版本管理，可保存不同时期的智能体配置，便于对比迭代效果并快速回滚。

- AI 能力
 - AI 对话
 - 知识库检索
 - MCP调用
 - 意图识别
 - 文本内容提取
 - 数据分析
- 工具
 - HTTP
 - 文本拼接
 - 文件上传节点
 - 指定回复
 - 文档解析
 - IF 判断器
 - 读取/存储
 - 变量更新
 - 代码运行
 - 智能体调用
 - 获取当前时间
 - 数据库操作
 - 知识库|用转换器
- 其他



DAP落地路径

场景定位

聚焦高重复、高价值、可量化的核心业务痛点，如客服应答、知识管理。避免盲目追求“科幻式想象”，确保项目落地的实用性和价值。

数据治理

清理冗余数据，统一数据格式，打通部门系统壁垒，为智能体提供高质量的数据输入，奠定坚实的数据基础。

工具选型

结合业务场景需求，选择适配的模型，完成DAP部署与环境配置，确保系统能够快速上线并投入使用。

选取典型场景

在单一部门的客服咨询或营销策划场景进行试点，快速验证智能体的实际效果，积累经验。

效果监测

建立量化评估指标，如准确率、效率提升率、错误率，与人工处理基线数据对比，客观评估智能体的性能。

迭代优化

根据试点反馈，调整知识库内容与 workflow 逻辑，持续优化智能体，提升其适配性和用户体验。

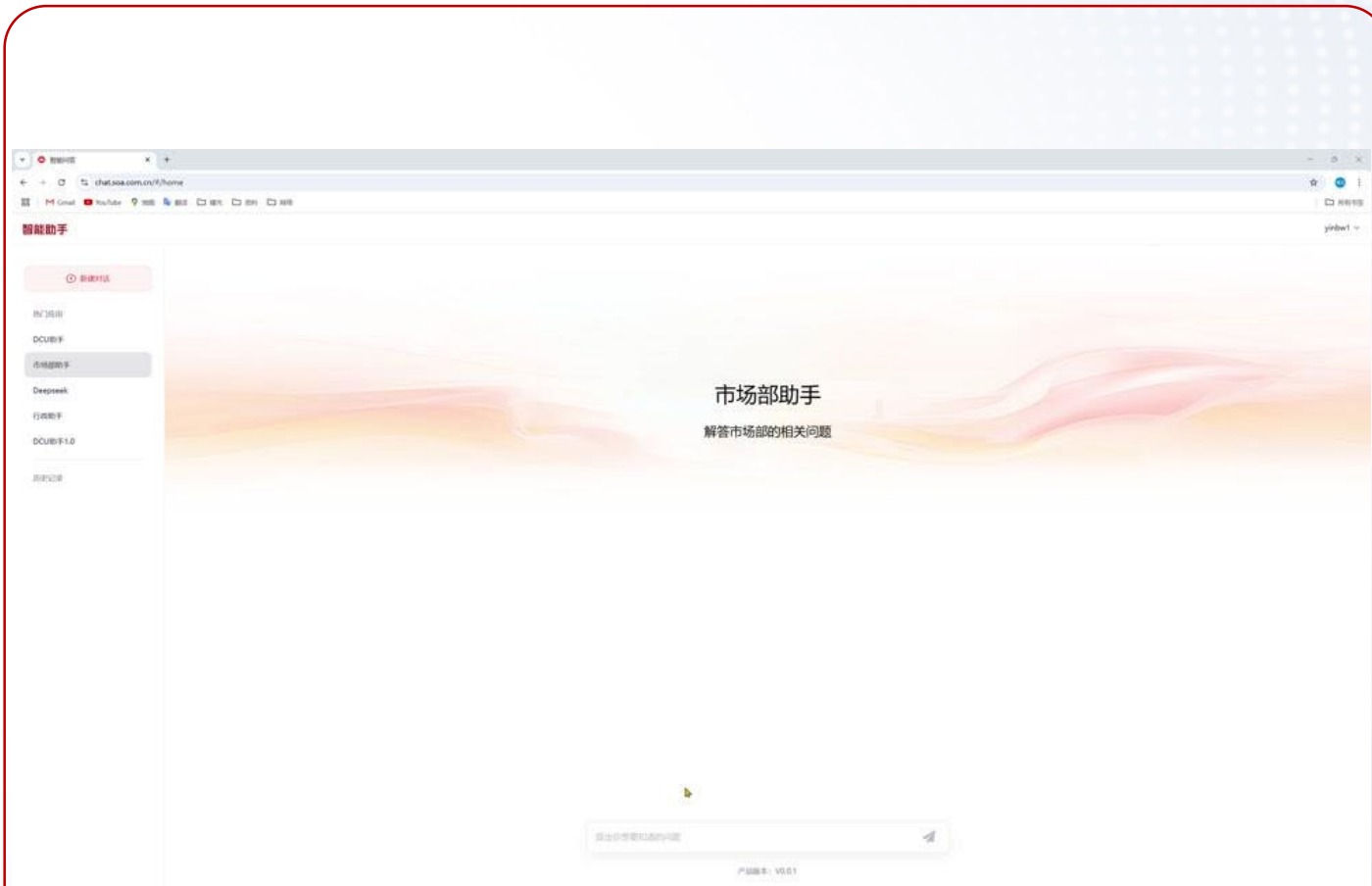
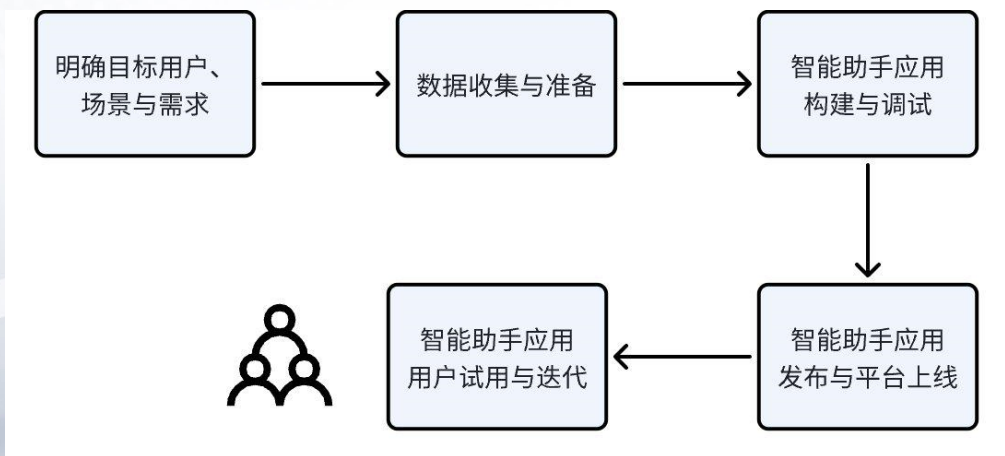
推广落地

分批推广：按业务线分批推广，避免一次性全量上线导致问题集中爆发。开展操作培训，实现人机协作，提升整体工作效率。

DAP落地案例-市场部助手

项目简介：随着公司业务的发展，市场部负责组织和管理的活动日益增多，涉及的内容也愈发复杂。为了提高工作效率，减少员工在查找活动规则、流程及物料获取等方面的时间成本，并确保信息的一致性和准确性，开发了一款智能问答助手。该助手充分利用了DCU和大模型能力优势，能够快速响应并提供精准的答案，帮助市场部同事更高效地完成日常工作。

应用构建流程：



应用演示

DAP落地要点

技术风险

多系统集成兼容性问题、复杂场景下问答准确率不足，可能导致项目无法顺利落地。

提前完成系统兼容性测试，通过多模型对比选型提升专业领域准确率，利用MCP协议扩展适配能力。

管理风险

最终用户对新工具的抵触、缺乏明确的价值评估标准，可能影响项目的推广和应用。

强化员工培训与价值宣导，建立“效率+质量+业务影响”三维评估体系，提升员工对智能体的认知和接受度。

合规风险

金融、医疗、等强监管行业的算法公平性与数据隐私问题，可能导致合规风险。

嵌入算法公平性验证与数据隐私保护环节，确保输出内容合规可控，满足强监管行业的要求。

生态支持

模型仓库

模型仓库

适配加速卡的优化模型库，助力开发者快速部署 AI 与科学计算任务

运行过程

综合排序 | 最新发布 | 下载数量

请输入关键词搜索

加速卡

K100AI | BW1000 | Z100L

算法类别

人脸识别 | 对话问答 | 多模态 | OCR | 代码生成

目标检测 | 文本分类 | 文本理解 | 视频生成

语音合成 | 语音识别 | AIIC | 多轮对话

多模态输入 | 图像分类 | 图像识别 | 以文生图

推荐系统 | 时序预测

框架类型

ait | bladefac | deepspeed | dgl

diffusers | fastertransformer | fastllm

jax | libai | lmdesploy | migraphx

ollama | oneflow | onnuruntime

paddle | pytorch | sglang

tensorflow | tgi | transformers

triton | bvm | unlosho | vllm

qwen3-30b-a3b_vllm

qwen3-30b-a3b是一个非思考模式 (non-thinking mode) 的新模型。仅配置30参数，就能取得媲美 Gemini 2.5-Flash (non-thinking)、GPT-4o等顶尖闭源模型的性能。

推理 | 对话问答 | vllm | 制造 | 医疗 | +3

更新于 2025-09-17 17:21:11

deepseek-v3.1_vllm

DeepSeek-V3.1 是一个支持思考模式和推理思考模式的混合模型。

推理 | 对话问答 | vllm | 制造 | 教育 | +1

更新于 2025-09-17 17:43:13

deepseek-v3.2-exp_vllm

DeepSeek-V3.2-Exp模型是一个实验版本，作为迈向下一代架构的中间步骤。

推理 | 对话问答 | vllm | 制造 | 广告 | +2

更新于 2025-10-04 15:53:10

deepseek-ocr_pytorch

DeepSeek 推出了全新的纯文本生成模型 DeepSeek-OCR，

推理 | OCR | pytorch | 交通 | 制造 | +3

更新于 2025-10-22 10:33:19

longcat_sglang

类似开源模型推理，一个强大且高效的推理模型。拥有总计 5600 亿个参数，采用了创新的专家混合 (MoE) 架构。

推理 | 对话问答 | sglang | 制造 | 表演 | +2

更新于 2025-09-08 10:05:26

Step3_pytorch

Step3 是一个先进的多模态推理模型，基于混合专家架构构建，拥有 321B 总参数，是token数达388参数。

推理 | 对话问答 | pytorch | 广告 | 教育 | +1

更新于 2025-08-06 15:28:58

DAP主页

DAP

首页 | 功能 | 定价 | 下载 | 文档 | 教程 | 联系我们

AI智能体对话

DAP 智能平台是面向智能时代的企业应用级金融 AI 开发应用平台。基于自主创新软硬件协同优化技术，搭载“数据治理-算法研发-模型部署-应用集成”全流程链条

在线体验 | 立即下载

产品架构

赋能千行百业

互联网 | 能源 | 交通 | 媒体 | 通信 | 金融 | 医疗

普通用户 | 企业用户 | 合作伙伴

通过WebAPP使用对话功能 | 调用API接口集成至自有系统 | 系统log、积分、背景等自定义功能

开发者社区版块

开发者社区

请输入搜索内容

首页 | 消息 | 发现 | zhangy5

全部 1069

官方技术指导 12

模型讨论区 231

DTK开发 424

人工智能 204

前沿算法讨论 7

DAP讨论区 9

科学计算应用 40

基金与大赛 60

FAQ 82

所有 | 精华 | 已关注 | 类型 | 排序 | 发布

推荐内容

- 【光合基金分享】一种提高... 203
- 【光合基金分享】HIP编程... 159
- 【光合基金分享】OpenChi... 208
- 在Fortran程序中使用HIPFO... 131
- 【光合基金分享】如何在for... 228

标题

5.2节点vllm6.2部署DeepSeek-R1-Distill-Qwen-7B模型，服务可正常启动，如下：

查看更多

扫一扫访问移动端 | 光合开发者社区

光合社区：<https://forum.sourcefind.cn/>
DAP相关介绍，软件包下载，功能试用，使用问题咨询

谢谢！