产品白皮书 正式版本 V6.0

AI-Service 智能服务与运维白皮书

以知识引擎、智能坐席与多模态多 Agent 协同，构建工业车辆售后服务与智能运维闭环。

面向工业车辆主机厂、经销与服务网络、租赁公司及企业维保团队

知识引擎 智能客服坐席 多模态多 Agent 运维

AI-Driven · 原稿发布日期：2026 年 9 月 21 日

章节目录

产品定位与致读者

执行摘要

第一章 行业背景：工业车辆服务与运维正在走向智能化

第二章 平台概述：构建知识化、智能化、闭环化的服务与运维体系

第三章 知识引擎：让工业车辆知识真正可检索、可理解、可调用

第四章 智能客服坐席：将知识能力融入每一次服务交互

第五章 多模态多 Agent 智能运维：从问题描述到可执行处置建议

第六章 典型应用场景：面向业务问题组织智能能力

第七章 部署、集成、安全与运营治理

第八章 结语：以可信知识驱动服务能力持续进化

附录 A：核心术语表

附录 B：官网公开客户案例与平台数据

产品定位与致读者

产品定位

AI-Service 以计算机视觉、边缘计算、物联网和 AI 大模型技术为基础，面向制造业和物流仓储企业，提供工业设备 AI 资产管理与售后智能客服的完整解决方案。系统聚焦场内物流载具安全监控、驾驶员行为识别、设备上云、数字化运维和智能售后，帮助企业降低安全事故风险、提升资产利用率与售后维护效率。

致读者

工业车辆的服务与运维，正在进入以知识、数据和智能协同为核心的新阶段。产品型号多、部件结构复杂、维修服务链路长，决定了企业无法仅依靠少数资深人员的个人经验来保障持续、标准和高质量的服务。

基于多模态大模型+多Agent协同的智能运维系统，面向工业车辆售后与运维场景，将维修手册、操作手册、视频资料、故障代码、历史案例、配件信息和质保规则等企业知识资产进行组织、调用与持续优化，并将其服务于客户咨询、故障诊断、维修指导、配件推荐、质保索赔、工单协同与运营洞察等业务环节。

本白皮书基于现有产品资料编写，重点介绍平台的能力体系、典型应用场景与部署实施路径。文中“支持”“可配置”“可按需对接”等表述，均应以最终产品版本、项目范围和合同约定为准。

执行摘要

生成式人工智能正在改变企业服务客户、沉淀知识和支持一线人员的方式。对于工业车辆企业而言，真正有价值的 AI 应用不应停留在通用问答，而应深入到车型资料查询、故障分析、维修作业、配件适配、质保索赔和客户服务等高频、专业且可衡量的业务场景。

基于多模态大模型+多Agent协同的智能运维系统，以三类能力共同支撑智能化服务升级：

- **知识引擎**：将文档、视频、错误代码、维修案例等资料沉淀为可检索、可问答、可调用的企业知识资产；
- **智能客服坐席**：在网站、微信、APP、企业服务系统或售后小程序等服务触点中，为客户与一线人员提供智能问答、标准回复建议、人工协同、工单流转和运营分析能力；
- **多模态多 Agent 智能运维**：围绕“检索—诊断—方案—配件—质保”链路，支持文本、语音、图片等输入，辅助输出可执行的诊断步骤、维修方案、配件建议和质保处理建议。

平台面向工业车辆主机厂、经销与服务网络、租赁公司和大型终端维保团队等客户，帮助企业在以下方面建立长期能力：

1. 将分散的产品、维修和服务知识沉淀为统一资产；
2. 降低一线服务对个人经验的依赖，提升服务一致性；

3. 缩短资料检索和故障定位路径，提升维修与配件服务效率；
4. 通过会话、工单与使用反馈，形成知识持续更新与运营优化闭环。

第一章 行业背景：工业车辆服务与运维正在走向智能化

1.1 从产品交付走向全生命周期服务

工业车辆广泛服务于仓储物流、制造、港口、机场、建筑、冷链及园区等场景。随着客户对设备可用性、维修及时性、配件保障和全生命周期服务的要求持续提高，企业的竞争重点正从单一产品能力，延伸到覆盖售前、交付、使用、维保、配件与质保的综合服务能力。

在这一过程中，售后服务和运维团队需要持续面对车型多、配置多、部件多、故障现象复杂、客户问题表述不标准、服务区域广及人员能力差异大等挑战。传统依靠纸质手册、分散文件、系统查询和人工问询的服务方式，难以支撑规模化、标准化和高响应的服务目标。

1.2 工业车辆服务的五类典型挑战

知识分散，经验难以规模化复用

操作手册、维修手册、培训材料、视频资料、故障代码、历史工单和维修案例往往分散在不同部门、不同系统和个人设备中。资深人员掌握的经验难以被结构化沉淀，新员工则难以快速获得准确指导。

故障诊断依赖人工经验

同一故障现象可能涉及电机、转向、液压、控制器、充电机或其他多个系统。传统诊断过程依赖经验判断和多轮沟通，既增加停机时间，也难以确保不同服务人员的判断一致性。

维修方案标准难以统一

维修过程中，故障原因、处理步骤、注意事项和最佳实践需要被及时记录、审核和复用。若缺少结构化知识与可执行指导，服务质量容易随人员能力波动。

配件识别与适配效率不足

配件查询常涉及车型、车辆序列号、物料号、部件图片、适配关系和替代件信息。人工比对工作量大，且易出现识别不完整、推荐不准确或跨系统查询效率低等问题。

质保索赔沟通链路复杂

质保与索赔通常涉及车辆信息、铭牌、故障描述、维修记录、配件信息、企业政策及材料清单。若规则口径未被清晰固化，容易增加审核、沟通和处理成本。

1.3 智能化服务的目标不是替代人，而是放大专业能力

工业车辆服务具有高专业性和高责任属性。大模型和智能 Agent 的合理定位，不是替代服务工程师、坐席人员或技术专家，而是将企业已有的知识、规则、案例和流程转化为可快速调用的辅助能力。

通过“AI 预处理 + 人工复核与决策 + 业务系统协同”的方式，企业可以让更多一线人员更快获得可信的资料、建议和标准流程，将专家经验沉淀为可持续复用的组织能力。

第二章 平台概述：构建知识化、智能化、闭环化的服务与运维体系

2.1 平台定位

基于多模态大模型+多Agent协同的智能运维系统，是面向工业车辆售后服务与智能运维场景的能力体系。平台以企业知识资产为基础，结合大模型、多模态理解和多 Agent 协同能力，为客户服务、技术支持、维修作业、配件业务和质保索赔等环节提供智能辅助。

平台重点解决“知识难找、经验难传、问题难判、服务难协同、数据难回流”的问题，使企业从零散资料和个人经验驱动，逐步走向知识驱动、流程协同和持续优化。

2.2 三层能力体系

第一层：知识引擎——企业知识资产底座

知识引擎负责将各类工业设备资料转化为可组织、可检索、可问答、可发布的知识资产。它承接手册、视频、故障代码、案例、车型及标签等资料，为智能应用、坐席服务与运维决策提供知识支持。

第二层：智能客服坐席——客户服务与人工协同触点

智能客服坐席面向网站、微信、APP、企业服务系统或售后小程序等渠道，为客户、一线服务人员和管理人员提供多渠道会话、智能建议、人工转接、工单协同、客户信息查看和运营洞察能力。

第三层：多模态多 Agent 智能运维——业务处置与决策辅助能力

智能运维能力围绕工业车辆售后业务中的高价值场景，组合使用车型资料、维护手册、历史案例、故障代码、配件 BOM 与质保规则等信息，辅助完成故障诊断、维修建议、配件识别与推荐、质保索赔指导等工作。

2.3 业务闭环

建议将平台能力理解为以下闭环：

> 知识资料与业务规则沉淀 → 智能查询与服务协同 → 诊断、维修、配件、质保辅助 → 会话、工单与使用反馈回流 → 知识持续更新与运营优化

该闭环反映的是平台的业务能力组织方式。具体模块间的数据共享、接口对接和系统部署方式，应以实际项目建设方案为准。

2.4 产品构成、交付边界与依赖条件

AI-Service 不是单一的对话应用，而是由知识引擎、智能客服坐席、多 Agent 智能运维、业务系统集成及私有化运营治理共同组成的产品能力体系。为避免将标准产品能力、客户数据条件和项目定制内容混同，平台按以下方式组织交付。

产品能力	核心内容	交付边界与前提
知识引擎	文档、视频、错误代码、维修案例等资料的数据集管理；标签、索引、检索、问答、应用/Agent 发布及知识审核	属于基础产品能力；效果依赖资料质量、车型/版本等标签完整性，以及客户对内容的审核与持续维护
智能客服坐席	Web、移动端、企业服务系统等渠道接入；访客端、坐席工作台、会话转人工、智能建议、工单协同、服务监控与运营分析	属于可配置产品能力；渠道、品牌、组织、服务区域、转人工规则和服务话术需按客户业务配置
多模态多 Agent 智能运维	售前咨询、售后诊断、配件推荐、索赔查询、翻译服务等任务型 Agent 的意图识别、任务分解、知识检索、工具调用和结果整合	属于场景化能力；需结合可用的知识、规则、模型、权限和人工审批点进行配置与验证
业务系统集成	与 CRM、ERP/SAP、MES、服务工单、配件、库存、质保等系统交换授权业务数据	仅在客户完成接口授权、字段确认、权限裁剪和联调验收后提供；平台不替代客户业务系统的主数据与审批权威性
私有化部署与运营治理	内网部署、统一身份认证、角色权限、HTTPS、日志审计、监控、备份与环境管理	部署拓扑、资源规格、高可用等级、备份容灾目标及安全策略须基于实际环境和合同范围确认

上述分类用于说明产品能力的适用边界：基础产品能力可作为平台底座交付；配置实施能力需结合业务流程启用；系统集成能力依赖授权接口和数据质量；专项算法、全量历史数据治理及客户专属流程属于项目定制范围。最终交付内容以需求确认、技术方案和合同约定为准。

2.5 工业车辆大模型项目沉淀的行业能力

基于工业车辆大模型项目在售前、售后和服务协同场景中的建设经验，AI-Service 可将行业知识、业务规则与人工责任边界组合为以下产品化能力。各场景均应在已接入的资料、规则和业务系统范围内使用。

业务场景	典型输入	可调用的知识或工具	平台输出与人工责任边界
售前咨询与选型支持	客户需求、工况、机型关键词、目标区域	产品资料、工况适配表、竞品参数、CRM 历史记录、经授权的 SAP/库存信息	输出参数查询、对比摘要、选型或配置建议；价格、交期、认证状态及对外方案须由授权人员复核
售后诊断与维修支持	车型、故障代码、故障现象、现场文字/语音/图片	故障代码库、维修手册、历史案例、系统知识、故障排查规则	输出代码解释、可能原因、检查顺序、风险提示和维修资料；现场检测、安全操作及最终故障结论不得由模型自动替代
配件识别与推荐	车型、车辆序列号、物料号、配件名称或图片	配件目录、BOM、兼容性规则、替代关系、经授权的库存数据	输出候选配件、适配依据、替代件提示和安装参考；最终适配、价格、库存和交期以正式业务系统与人工核验结果为准
质保与索赔指导	车型、购车日期、故障类型、维修记录、铭牌或材料图片	质保规则、索赔流程、材料清单、相关案例	输出材料完整性提示、办理流程和注意事项；是否受理、质保判定、赔付金额及对外承诺须由具备授权的人员审批
多语言服务与资料辅助	咨询文本、产品/维修资料、目标语言	术语库、知识库、翻译模型	输出多语言问答或文档初稿；涉及技术规范、合同、法规和对外文件时应执行人工校对与发布审核

2.5.1 任务型 Agent 的协同方式

平台可将复杂请求拆分为“统一入口 → 意图识别 → 任务分解 → 知识与业务数据调用 → 结果整合 → 人工复核或工单流转”的处理链路。例如，售后诊断 Agent 可结合故障代码、车型资料和案例库生成排查建议；配件推荐 Agent 可在取得授权后查询兼容性规则和库存信息；索赔查询 Agent 则以有效质保规则和材料要求为依据提供流程辅助。

每一次任务处理应保留必要的会话上下文、知识引用、工具调用、权限校验和人工接管记录。对于低置信度、资料冲突、超出知识范围、涉及人身/设备安全，或需要价格、库存、质保审批等权威业务结论的请求，系统应提示转人工、发起工单或进入授权审批流程。

2.5.2 行业知识数据模型与生命周期

工业车辆知识的有效使用依赖可追溯的数据组织方式。除文件内容外，平台宜为知识建立来源、责任部门、车型/系列、系统、部件、故障类型、故障代码、适用工况、行业场景、版本、生效日期及适用范围等元数据；对配件、质保和案例资料，还应关联物料号、兼容性规则、政策版本或案例状态等业务属性。

建议采用以下生命周期管理机制：

> 资料盘点与授权 → 解析、标注与索引 → 待审核入库 → 发布给指定应用或 Agent → 使用反馈与问题回流 → 定期或触发式更新 → 失效归档与历史版本追溯

其中，正式手册、维修案例、业务规则、经验草稿和失效资料应采用不同的发布与访问策略。平台可以辅助检索、归纳和提示知识缺口，但不能自动将未经审核的内容作为正式维修、质保或对外服务依据。

2.5.3 人机协同服务流程

面向客户和一线服务人员，平台可将服务过程组织为“访客/服务人员发起咨询 → AI 理解与知识检索 → 标准问题即时辅助或人工接管 → 工单创建与流转 → 处理结果、满意度和问题标签回流 → 知识库和服务流程持续优化”。

坐席人员可在同一服务上下文中查看会话历史、AI 建议、相关知识、相似案例和已授权业务信息，并在必要时接管会话或创建工单。会话摘要、推荐回复、热点问题统计和知识使用情况可作为运营分析依据；客户资料、订单、库存、索赔等信息的展示范围，应遵循组织、角色、区域和最小必要权限原则。

第三章 知识引擎：让工业车辆知识真正可检索、可理解、可调用

3.1 面向工业设备知识管理的专属知识库

工业车辆企业拥有大量高价值资料，包括产品说明书、操作手册、维修手册、培训材料、故障代码表、维修案例、零部件图册和视频教程。知识引擎支持建立面向工业设备知识管理的专属知识库，为后续检索、问答与智能应用提供基础。

资料可按业务需要纳入通用文档、视频文件、错误代码等不同类型的数据集，并可添加描述、标签和索引信息，帮助模型和使用人员在具体场景中更快定位相关内容。

3.2 用车型、系统和案例标签组织专业知识

工业车辆知识具有强上下文属性。同一个问题的答案，往往与车型、部件、系统、配置、故障类型和使用环境有关。因此，知识建设不仅是文件上传，更需要建立适合业务的分类与标签体系。

平台可结合实际业务，围绕以下维度组织知识：

- 车型、车辆系列、配置与适用范围；
- 液压系统、转向系统、控制器、驱动系统、制动系统、充电系统等部件或系统；
- 故障现象、故障代码、维修动作与风险提示；
- 零部件名称、物料号、适配关系和替代关系；
- 质保规则、索赔流程、材料要求与审核要点。

通过标签化组织，企业可以让知识资产从“资料库”升级为能够支持具体业务判断和服务动作的“知识网络”。

3.3 结构化沉淀维修解决方案与案例

平台可按照故障描述、故障分析、解决步骤、总结、过程图片或视频、适用车型等结构化字段沉淀解决方案和维修案例。

这种结构化方式的价值在于：一方面，服务人员可以快速找到与当前问题相近的历史处理经验；另一方面，企业可以将优秀工程师的诊断逻辑和维修方法沉淀为组织资产，持续提升整体服务能力。

3.4 将知识发布为应用与 Agent

知识引擎支持构建简易应用和 Agent 代理，并为不同应用关联相应知识库。企业可根据客户服务、技术支持、内部培训、配件查询、维修指导等不同场景，配置不同的应用信息、输入引导和使用权限。

平台可支持分享链接、聊天授权、对话日志查看、图片辅助问答、语音输入和语音播放等能力，使知识能够以更自然的方式服务一线人员和客户。

3.5 知识治理是智能化成功的前提

企业在推进智能服务前，应建立知识治理机制：

1. 明确资料来源、责任部门和更新机制；
2. 区分正式手册、经验案例、内部草稿和失效资料；
3. 对关键信息建立车型、版本、生效日期和适用范围标识；
4. 对专业结论、维修建议和质保规则设置审核流程；
5. 通过使用反馈持续发现知识缺口和内容冲突。

第四章 智能客服坐席：将知识能力融入每一次服务交互

4.1 多渠道服务入口

客户咨询可能发生在官网、微信、APP、企业服务系统或售后小程序中。智能客服坐席可作为统一服务入口，帮助企业承接咨询、技术支持、服务请求和业务引导。

通过统一的坐席工作台，服务人员可以在会话中查看客户信息、历史记录、相关工单、配件订单及系统推荐的回复建议，减少多窗口切换和重复信息确认。

4.2 AI 辅助与人工协同并行

对于标准化、重复性较高的问题，系统可基于历史语料和知识库提供标准回复建议，坐席人员可结合实际情况进行确认和发送。对于涉及复杂诊断、特殊政策或高风险决策的问题，可转交人工专家或指定服务人员处理。

这种模式既能提升常见问题的响应效率，也能保留人工对复杂问题的专业判断和最终责任。

4.3 会话、客户、工单与业务信息协同

智能客服坐席可在服务过程中承接多类业务信息：

- 客户基础资料、联系方式和历史交互记录；
- 会话内容、咨询意图和处理过程；
- 工单草稿、历史工单及处理状态；
- 配件订单、索赔与售后相关信息；
- 客户标签、设备环境和服务偏好等信息。

当客户问题需要进一步处理时，系统可将会话中的关键信息带入工单，降低重复录入成本，并使后续处理人员能够理解完整上下文。

4.4 运营洞察：让服务数据反哺知识和流程优化

服务过程中产生的会话、工单和使用数据，是企业发现问题和优化能力的重要基础。平台可支持多维历史检索、热点问题聚类、关键指标观察、根因下钻和优化建议等能力。

企业可据此识别：

- 哪些问题咨询频率最高；
- 哪些车型、部件或故障类型需要补充知识；
- 哪些问题需要优化产品资料、培训内容或服务流程；
- 哪些坐席或服务区域需要更多支持；

- 哪些服务请求应进一步沉淀为标准案例。

4.5 全球化服务与治理能力

对于跨区域或海外业务场景，平台可提供多语言使用能力，并支持组织、业务线、服务区域等维度的授权管理。结合敏感词规则、品牌化配置、多网站配置及账号管理等机制，企业可以在扩展服务触点的同时保持统一的服务标准和管理边界。

第五章 多模态多 Agent 智能运维：从问题描述到可执行处置建议

5.1 以“检索—诊断—方案—配件—质保”为核心链路

工业车辆运维并不只是回答“某个部件是什么”或“某个代码代表什么”。真正有价值的智能化能力，应能把问题信息、企业知识、业务规则和后续动作串联起来。

平台围绕以下业务链路提供辅助能力：

> 检索相关资料 → 分析故障或业务问题 → 生成维修或处置建议 → 推荐相关配件 → 提示质保或索赔处理要点

该过程可结合企业实际资料和规则使用，并由服务工程师、技术专家或业务人员进行必要的核验与决策。

5.2 支持文本、语音与图片等多模态输入

一线服务场景中的问题输入往往不标准。用户可能以自然语言描述故障现象，也可能通过语音说明问题，或直接上传故障部位、零部件、铭牌等图片。

多模态输入能力可帮助系统更充分地理解上下文。例如：

- 服务人员输入故障现象，获得可能原因与检查方向；
- 上传零部件图片，辅助识别零件名称或物料信息；
- 上传车辆铭牌图片，辅助提取车辆序列号等信息；
- 使用语音快速描述现场问题，降低文字输入成本。

图片识别、型号判断、配件适配和序列号识别等结果，仍应结合正式产品资料、业务系统数据和人工核验后用于实际业务决策。

5.3 场景一：车型与维修手册查询

当服务人员需要查询某一车型的操作手册、维修手册、维护要求或技术说明时，平台可根据车型、关键词、系统标签和资料内容提供检索与问答支持。

这一能力可缩短人工翻阅资料的时间，也有助于新员工快速熟悉产品与服务规范。

5.4 场景二：故障代码查询与故障诊断

面对故障代码或异常现象，平台可结合故障代码表、维护手册、历史案例和维修规则，辅助解释代码含义、提示可能原因、给出检查顺序与处理建议。

例如，当车辆出现启动后轮子持续自转等现象时，系统可辅助服务人员从电机、转向、液压、控制器、充电系统等可能方向展开排查，而不应替代现场检测和安全规范。

5.5 场景三：步骤化维修指导

基于结构化维修案例与手册内容，平台可将专业知识组织为更便于执行的步骤化指导，包括检查前提、操作步骤、注意事项、所需工具和结果验证等内容。

通过将优秀维修经验沉淀为标准化方案，企业可以提升不同区域和不同层级服务人员的作业一致性。

5.6 场景四：配件识别、查询与推荐

配件业务是工业车辆服务链中的重要组成部分。平台可结合车型、车辆序列号、配件名称、物料号、配件图片和 BOM 等信息，辅助进行零部件识别、适配关系查询、相关配件推荐及替代件提示。

在实际业务中，配件价格、库存、交期、最新物料状态和最终适配关系，应以企业正式业务系统和授权规则为准。

5.7 场景五：质保与索赔指导

针对整机或配件索赔场景，平台可基于企业质保规则、索赔政策、案例和材料要求，辅助提示处理流程、所需材料、信息完整性和注意事项。

这一能力有助于减少口径不一致和材料遗漏，但最终质保判定、索赔审批和对外承诺应由具备授权的业务人员依据有效政策作出。

第六章 典型应用场景：面向业务问题组织智能能力

6.1 电池、驱动与制动系统维护问答

典型问题： 如何检查铅酸电池？锂电池、驱动系统和制动系统应如何维护？

平台能力： 基于维护手册、培训资料和历史案例，提供结构化的检查项目、维护周期、注意事项和相关资料入口。

业务价值： 将标准维护知识前置到一线，降低基础咨询对专家的依赖，提升服务口径一致性。

6.2 车型操作与维修资料快速定位

典型问题： 如何找到某一车型的操作手册或维修手册？

平台能力： 基于车型、关键词、系统标签和文档索引，辅助定位相关资料并提供内容问答。

业务价值： 减少人工翻找文件的时间，提高服务人员和培训人员的资料使用效率。

6.3 基于图片的零部件识别与物料查询

典型问题： 上传多路阀或其他零部件图片后，能否辅助识别零件名称和物料号？

平台能力： 利用图片输入能力，并结合零部件资料、BOM 和产品知识，辅助识别部件并引导进一步确认。

业务价值： 缩短部件初步识别和资料查询路径，为后续配件业务提供辅助信息。

6.4 车型、序列号与配件推荐

典型问题： 如何按车型、车辆序列号和配件名称查询适配件或替代件？

平台能力： 根据企业可用的车型资料、配件信息和 BOM 数据，辅助完成适配关系查询与推荐。

业务价值： 改善配件服务的响应效率，减少人工查询和沟通成本。

6.5 故障现象分析与维修建议

典型问题： 车辆开机后轮子持续自转，可能涉及哪些问题？

平台能力： 基于故障案例、手册、故障代码和系统知识，辅助给出排查方向、检查步骤和风险提示。

业务价值： 帮助一线人员快速建立诊断思路，并为专家进一步判断提供结构化信息。

6.6 质保索赔材料与流程指导

典型问题： 叉车索赔需要准备哪些资料？如何从铭牌中获取车辆序列号？

平台能力： 基于质保规则、索赔流程和材料清单，辅助提示所需信息与处理步骤。

业务价值： 降低材料遗漏和沟通往返，提高索赔业务的规范性与处理效率。

6.7 销售支持、行业资料与新人培训

典型问题： 建筑工地、冷库等不同应用环境应如何选择工业车辆？新员工如何快速学习锂电池产品知识？

平台能力： 将产品资料、行业场景、培训材料和标准问答沉淀为可调用知识，面向销售、渠道和新人提供学习与问答支持。

业务价值： 促进产品知识和行业经验的快速传播，提升客户沟通与培训效率。

第七章 部署、集成、安全与运营治理

7.1 分阶段建设路径

企业可根据业务目标和数据条件，采用循序渐进的建设方式：

1. **场景识别与资料盘点**：确定优先场景，盘点手册、案例、故障代码、配件资料和质保规则；
2. **知识建库与试用验证**：接入样本文档和表格，建设基础知识库，验证重点问答与业务场景；
3. **流程协同与运营优化**：将智能能力嵌入服务触点，完善人工协同、工单和运营分析机制；
4. **正式部署与按需集成**：根据项目范围，规划企业门户、微信小程序、公众号及 CRM、ERP、MES、EPC 等系统的集成工作。

试用阶段通常应聚焦资料接入、知识构建、场景验证和效果评估。接口开发、定制功能、专项算法优化及全量历史数据治理等事项，应在正式项目阶段依据双方约定推进。

7.2 部署与接入方式

AI-Service 采用“1+3+4”私有化部署架构体系，面向企业内网环境构建从智能应用、业务协同、数据服务到基础设施的完整能力栈。该架构适用于对产品资料、客户信息、服务记录、配件数据和运维数据具有较高数据安全与自主可控要求的制造业和物流仓储企业。

7.2.1 私有化部署总体架构



平台总体分为四层：

1. **应用层**：面向业务人员提供售前咨询 Agent、售后诊断 Agent、配件推荐 Agent 和索赔查询 Agent 等专业智能应用；
2. **业务能力层**：包括智能客服支撑系统、多 Agent 执行系统和垂直领域微调大模型。其中，智能客服支撑系统作为统一入口，可覆盖访客端 Web、访客端小程序、坐席工作台和监控后台；多 Agent 执行系统负责意图识别、任务分解、工具调用与结果整合；垂直领域微调大模型为工业设备专业问答、诊断与业务协同提供模型能力；
3. **数据服务层**：由统一知识库（Milvus）、业务数据库（PostgreSQL）和缓存服务（Redis/ES）组成，分别承载专业知识检索、业务数据管理及高频访问与检索加速；
4. **基础设施层**：由 GPU 服务器、应用服务器和存储服务器构成，为模型推理、业务服务运行及数据存储提供资源支撑。

7.2.2 Agent 协同与数据调用链路

针对客户咨询、现场报修、配件查询和索赔指导等请求，平台可按照“统一入口 → 意图识别 → 任务分解 → 知识与业务数据调用 → 结果整合 → 人工复核或工单流转”的方式组织处理：

1. 用户通过 Web、小程序、坐席工作台或其他业务入口提交文本、语音、图片及设备相关信息；

2. 智能客服支撑系统识别用户身份、会话上下文及业务意图，并将复杂任务交由多 Agent 执行系统编排；
3. 多 Agent 执行系统按任务需要调用垂直领域模型、统一知识库、业务数据库和经授权的外部业务工具；
4. 系统整合检索依据、业务信息和处置建议，输出可追溯的回复、诊断步骤、配件建议或索赔材料清单；
5. 对涉及现场安全、最终配件适配、质保审批、价格库存及对外承诺的事项，保留人工核验、授权审批或工单流转环节；
6. 会话、工单和处理反馈经脱敏、审核或按权限处理后，可用于知识更新、案例沉淀与运营分析。

这种设计使通用对话能力与行业知识、业务数据和人工责任边界相结合，避免将模型输出直接替代安全操作、业务审批或最终专业判断。

7.2.3 服务入口、环境配置与试用路径

平台服务入口可结合 PC、移动端、企业门户、微信小程序或公众号等形式进行设计。私有化交付时，建议按模型推理、应用服务和数据存储进行资源隔离或逻辑隔离，并依据并发量、知识库规模、文件解析量、图片/语音使用比例及保留周期进行横向扩展规划。

对于试用阶段，企业可依据安全、数据和管理要求选择在线 SaaS 试用或私有化试用；正式项目的服务器规格、GPU 型号与数量、存储容量、网络隔离、备份容灾、数据迁移、账号体系及系统接口方案，应在项目调研后确定。涉及客户生产业务的接入应遵循“最小必要权限、分环境验证、可审计回退”的实施原则。

7.2.4 RAG 检索与模型使用策略

AI-Service 采用 B/S 访问形态，以 HTTPS 对外提供知识库、坐席工作台和访客端等独立服务入口。平台以 RAG 知识库为核心：先通过向量检索、关键词/结构化条件过滤及重排序定位可信资料，再由大模型在检索上下文范围内生成回答、摘要或处置建议。该方式使知识库、坐席服务和行业 Agent 能够共享可追溯的企业知识底座。

模型能力按“轻量推理本地化、对话能力集中化”的原则进行组织：

- embedding、rerank 等轻量推理服务可基于 vLLM（或经验证的等效推理框架）部署在客户内网 GPU 节点，与 Milvus 等向量检索服务协同工作；
- 大模型对话、翻译等通用生成能力可经统一网关调用集中模型服务；如需调用外部 API，应在客户的数据出域、安全评估、脱敏和授权要求明确后实施；
- 模型权重、提示词、知识库、业务工具和会话数据应按权限隔离；业务微服务不得抢占模型推理所需的 GPU 资源；
- 产品不要求在每个客户环境中重复部署大参数模型，实际部署方式应根据数据安全、并发、成本和响应时延综合确定。

7.2.5 Kubernetes 容器化微服务参考架构

正式项目可采用以 Kubernetes 为核心的容器化微服务架构：控制面按多节点高可用方式部署，工作节点按 CPU 与 GPU 角色进行资源隔离。入口链路可按“WAF（如 samwaf）→ API 网关（如 APISIX）→ 应用服务”分层组织，由网关统一执行鉴权、路由、限流和审计；知识库、坐席与访客端可分别路由至对应的后端服务。

工作负载类型	适用组件示例	作用与部署原则
Deployment	坐席后端、访客端后端、RAG 检索/加载/工作器、网关、前端等无状态服务	支持弹性扩缩容、滚动发布与故障自愈，优先运行在 CPU 工作节点
StatefulSet	MySQL、PostgreSQL/TimescaleDB、MongoDB、Redis、Neo4j、Elasticsearch、RabbitMQ、RocketMQ、etcd、MinIO、Milvus 等	通过持久卷保存状态；应根据业务量选择组件组合，不要求每个项目部署全部中间件
DaemonSet	CNI、节点 DNS 缓存、监控采集器、NVIDIA GPU 设备插件等	为每个符合条件的节点提供基础网络、可观测性和 GPU 设备发现能力

GPU 节点可设置 `gpu=nvidia:NoSchedule` 等污点，并通过节点标签、`nodeAffinity` 与 `toleration` 将 embedding、rerank、向量检索 GPU 查询节点等推理负载绑定至 GPU 节点，从而与坐席、网关、知识管理等业务微服务实现物理或逻辑隔离。镜像可通过企业私有镜像仓库（如腾讯云 TCR）管理，并按需配置公共组件镜像加速代理；具体仓库产品与镜像安全策略由客户环境确定。

7.2.6 资源配置与基础运行环境参考

以下为单台物理服务器的参考配置基线，不构成统一采购规格。实际资源应根据并发会话数、知识库规模、文件解析量、图片/语音占比、模型调用模式和高可用目标完成容量评估后确定。

部件	参考规格	参考数量/说明
CPU	基础主频不低于 2.25GHz、睿频不低于 3.4GHz 的 64 核处理器	2 颗，共 128 核
内存	DDR4/DDR5，频率不低于 3200MHz，单条不低于 64GB	8 条，共 512GB
GPU	NVIDIA 单卡显存不低于 24GB	2 张
系统盘	NVMe SSD 480GB	2 块，建议 RAID1
数据盘	NVMe SSD 2TB	4 块，建议 RAID5 或按可靠性目标选择 RAID10
阵列与电源	支持 RAID 0/1/5/10 的阵列卡；冗余电源	阵列卡不少于 8 端口；双电源并预留功耗余量
网络与虚拟化	内外网互通、域名解析、千兆/万兆网口；虚拟化平台	可采用 Proxmox VE、VMware ESXi 或经确认的等效平台
操作系统与容器	Ubuntu 24.04 LTS 虚拟机、containerd、Kubernetes 1.3x 系列	以项目兼容性与安全基线为准

建议机房维持 18–27°C、推荐 22°C±2°C，相对湿度 40%–60%RH，温度变化率不高于 5°C/h；配置精密空调、温湿度监控，并按设备总功耗的 1.2–1.5 倍规划制冷冗余。网络层可采用 Calico CNI 与 VXLAN Overlay，为 Pod 提供独立网段；GPU 设备通过 `nvidia-device-plugin` 以 DaemonSet 方式上报给 Kubernetes 调度器。

7.3 安全、网络隔离与权限治理

工业车辆服务和运维知识中可能包含产品技术资料、客户信息、配件数据、服务记录和质保规则等敏感信息。因此，企业应从一开始建立覆盖身份、权限、网络、数据和模型调用的治理机制。

平台可结合项目需要提供或配置以下安全与治理能力：

- 多租户、组织、业务线和服务区域的数据隔离；
- 细粒度角色、功能和数据权限，以及单点登录（SSO）能力；
- HTTPS 传输、统一身份鉴权、接口限流、日志审计和账号授权管理；
- 面向访客端、坐席端和知识管理端的独立访问入口与会话边界；
- 按资料密级、适用对象和发布状态控制知识访问；
- 对高风险 Agent 工具、外部接口调用和人工接管过程保留审计记录。

在 Kubernetes 部署模式下，可按业务（如 `llm`）、向量检索（如 `milvus`）、日志（如 `loki`）、监控（如 `monitoring`）和 Agent 沙箱（如 `agent-sandbox`）划分命名空间。对多租户、高风险工具调用或 Agent 沙箱等工作负载，可使用 `NetworkPolicy` 实施入向白名单隔离；其余内部服务也应通过网络访问控制、网关鉴权及最小开放端口策略进行保护，命名空间本身不提供网络隔离。

集群内部服务优先使用 ClusterIP 进行互访，外部访问经 API 网关或在受控场景下使用 NodePort 暴露；所有外部入口应前置 WAF，并结合网关实施请求清洗、路由和访问控制。企业还应对知识内容实施分级分类，明确哪些资料可供客户使用、哪些仅限内部服务人员使用、哪些需要专家审核后才能发布。

7.4 数据持久化、检索服务与数据/模型边界

平台可采用统一持久化层承载数据库、知识库索引、模型权重、日志及备份数据。私有化环境中可采用 NFS 或经确认的企业级共享存储作为 StorageClass 后端，并为关键持久卷设置 `Retain` 回收策略，防止删除 PVC 时自动回收底层数据卷。

数据服务可按项目范围组合使用以下组件：

- 关系与业务数据：MySQL 8.0、PostgreSQL/TimescaleDB；
- 文档、会话与图谱等数据：MongoDB、Neo4j、Elasticsearch、Redis；
- 消息与异步任务：RabbitMQ（可包含延迟交换能力）、RocketMQ；
- RAG 检索：Milvus 集群，使用 etcd 维护元数据、MinIO 承载对象数据，并可配置 GPU Query Node 提升向量查询性能；
- 对象存储：按向量库、日志或业务资料分别配置 MinIO 实例或等效对象存储，避免不同类型数据相互影响。

上述组件均属于参考技术栈，应根据数据规模、并发、可用性、运维能力和既有 IT 标准选择，不应理解为每个客户项目的固定部署清单。模型权重可通过共享存储 PVC 提供给推理 Pod 挂载加载；GPU 节点仅承载获准的轻量推理和向量检索负载，不承载通用业务微服务。涉及外部模型或 API 的请求，应在客户明确数据边界、加密、脱敏、权限与审计策略后通过统一网关发起。

7.5 备份、容灾与可观测性

平台应在上线方案中明确备份对象、频率、保留期限、恢复责任和演练要求。可采用 Kubernetes CronJob 对中间件数据与应用配置执行定时备份；项目参考基线为每 2 日凌晨执行一次，备份产物写入 NFS 或对象存储持久卷。MySQL、PostgreSQL、MongoDB 等关键库还应定期执行逻辑备份，保留多个可恢复版本，并按客户制度进行离机或异地保存。

除备份外，建议建立恢复验证机制：对持久卷回收策略、数据库恢复、知识库索引重建、模型权重加载和关键业务配置恢复进行定期演练。具体 RPO、RTO、备份频率和异地容灾等级应在合同与项目方案中明确。

可观测性基线可包括 `node-exporter`、`metrics-server`、Prometheus 与 Grafana，用于采集节点、容器、GPU、服务和业务指标；日志可采用 Loki 与 Alloy 或客户既有日志平台统一采集。监控与告警应至少覆盖资源使用率、服务健康、接口成功率与时延、推理/检索性能、队列积压、存储容量、备份任务结果及异常访问行为。

7.6 效果评估建议

在试用与正式建设过程中，建议结合实际业务选择评估指标：

- 有效回答率、知识命中率与引用依据完整性；
- 故障定位和资料检索时间；
- 首次修复率与重复工单比例；
- 配件识别和推荐的准确性；
- 索赔处理时长与材料完整率；
- 坐席使用频次、人工介入比例与用户满意度；
- 知识更新及时性、问题修复率与人工接管闭环率；
- 平台可用性、接口成功率、备份成功率和恢复演练达标率。

具体目标值应结合客户现状、数据质量、业务流程、项目范围和已确认的服务等级共同确定。

第八章 结语：以可信知识驱动服务能力持续进化

工业车辆行业的智能化升级，不仅是部署一个大模型应用，更是对企业知识资产、服务流程、数据治理和组织协同能力的一次系统性建设。

基于多模态大模型+多Agent协同的智能运维系统，以知识引擎为底座，以智能客服坐席为服务触点，以多模态多 Agent 能力支撑故障诊断、维修指导、配件推荐和质保索赔辅助，帮助企业逐步形成面向全生命周期服务的智能能力体系。

未来，随着知识持续沉淀、业务数据不断回流和应用场景持续扩展，企业可以从“解决单个问题”走向“构建可持续进化的服务与运维能力”，在服务效率、客户体验、渠道协同和品牌专业形象方面建立长期竞争优势。

附录 A：核心术语表

术语	说明
工业车辆	面向仓储、物流、制造、港口、冷链、建筑等场景的工业装备，本白皮书重点覆盖叉车等相关应用场景。
知识引擎	用于组织、索引、检索、问答和发布企业专业知识的能力平台。
智能客服坐席	面向客户服务与一线服务人员的多渠道会话、AI 辅助、人工协同、工单与运营分析能力。
多模态	支持文本、语音、图片等多种输入形式的 AI 理解与处理能力。
多 Agent	由多个具备不同任务分工的智能代理协同完成检索、诊断、方案、配件、质保等任务的能力模式。
故障代码	设备通过控制系统记录或输出的异常代码，可结合手册、案例和现场检查进行分析。
配件 BOM	用于描述产品零部件构成、物料信息及关系的清单数据。
适配件	与特定车型、配置或部件具有适配关系的零部件。
替代件	在满足业务、技术和授权规则前提下，可作为原配件替代选择的零部件。
SSO	单点登录能力，用户可通过企业统一身份体系访问授权应用。

附录 B：官网公开客户案例与平台数据

B.1 使用说明

本附录根据爱动超越官网截至 2026 年 9 月公开展示的案例页面与指标页面整理，用于补充本白皮书的客户实践与数据依据。除特别注明外，文中数值均采用官网披露口径；项目成果受客户基础条件、数据质量、实施范围、统计周期和业务流程影响，**不构成对其他客户的结果承诺**。

以下内容为对官网公开资料的归纳改写，而非逐字转载。

B.2 平台规模与行业数据底座

依据爱动超越的核心指标与统计口径页面，截至 2026 年 7 月 1 日，平台累计接入工业车辆 **23 万台**、累计服务企业客户 **5000 家以上**、覆盖 **40 个以上国家和地区**；2026 年 1—6 月，平台日均处理运维工单与诊断请求超过 **3000 条**。

官网同时披露，平台已累计沉淀超过 **10 万条维修工单**、**5000 多类故障类型**、**30 亿条以上故障码**和 **20 万页多语种维修手册**。这些数据属于累计数据资产口径，与在线车辆数、模型覆盖故障数量及日均处理量并非同一统计维度。

在诊断与运营效能方面，官网披露的代表性口径包括：复杂场景故障诊断准确率超过 **95%**，故障语义解析准确率为 **97.3%**；复杂故障平均诊断时长由 24 小时以上压缩至 3 分钟内；远程一次性解决率为 **81%**；高频问题自动处理率为 **80%**；平台全年服务可用率为 **99.9%**。其中，不同指标具有各自的样本范围和定义，应以官网指标说明为准。

B.3 公开案例一：徐工集团——面向海量车型与全球服务网络的知识化升级

根据徐工集团案例公开信息，项目面向车型多、知识分散和全球化服务响应压力等问题，建设了由接入、服务、数据和业务组成的智能化架构。项目将 **198 款以上车型主数据**与 **5800 条以上核心服务知识**纳入企业级知识体系，并围绕检索增强生成（RAG）、多模态视觉交互和售前智能选型建设应用能力。

公开案例披露的应用效果包括：维修资料检索从约 30 分钟缩短至约 1 分钟、超过 60% 的故障可实现自动响应或拦截、新员工辅助培训周期由 6 个月缩短至 2 个月；项目在约 3 个月内完成从数据梳理、核心功能试运行到一期验收与推广的建设节奏。上述为该项目公开口径，适用条件不应外推至其他项目。

B.4 公开案例二：林德（中国）——车、云、人协同的全生命周期服务模式

根据客户案例页公开信息，林德（中国）采用“AI 销售助手 + AI 服务工程师助手”的双场景模式，并结合 ADAS 与 T-Box 数据基础，探索从安全感知、售前价值沟通到售后服务支持的软硬件协同。

案例公开展示了 7×24 小时智能服务、车—云—人数据协同和预测性维护建议等能力方向，并给出了单车全生命周期价值提升、服务收入增长与服务溢价等项目指标。该案例适合作为“从单一设备销售走向全生命周期服务”的业务叙事，不应将其中的具体商业指标视为跨客户的通用结论。

B.5 公开案例三：诺力集团——面向索赔与配件业务的精细化运营

根据客户案例页公开信息，诺力集团案例聚焦分散服务渠道、索赔审核和配件匹配等业务问题，通过统一接入服务渠道、索赔自动化 Agent 和配件匹配能力，推进全球售后精细化运营。

案例公开展示的关键指标包括：索赔审核成本降低 **90%**、配件推荐准确率超过 **90%**、知识库查询效率提升 **80%**、故障语义识别率为 **97.3%**。其中配件准确率、审核成本等均为该案例对应的业务范围和统计口径，不应脱离业务条件单独引用。

B.6 来源链接

- 爱动超越官网首页
- 客户案例总览
- 核心指标与统计口径