Product Whitepaper Official edition V6.0

AI-Service Intelligent Service & Maintenance Whitepaper

Connect the knowledge engine, intelligent service desk and multimodal agents across industrial vehicle after-sales service and maintenance.

For industrial vehicle manufacturers, dealers, service networks, rental companies and maintenance teams

Knowledge engine | Intelligent service desk | Multimodal multi-agent maintenance

AI-Driven · Original publication: September 21, 2026

Contents

Positioning and introduction

Executive summary

Chapter 1. Industry context: intelligent service and maintenance for industrial vehicles

Chapter 2. Platform overview: knowledge-driven, intelligent service with continuous feedback

Chapter 3. Knowledge engine: making industrial vehicle knowledge searchable, understandable and usable

Chapter 4. Intelligent service desk: knowledge in every service interaction

Chapter 5. Multimodal multi-agent maintenance: from symptoms to actionable guidance

Chapter 6. Typical applications: organizing capabilities around business problems

Chapter 7. Deployment, integration, security and operational governance

Chapter 8. Conclusion: continuously improving service through trustworthy knowledge

Appendix A. Key terms

Appendix B. Public customer cases and platform data

Positioning and introduction

Product positioning

Built on computer vision, edge computing, IoT and large language models, AI-Service provides an integrated solution for AI-enabled industrial asset management and intelligent after-sales service in manufacturing, logistics and warehousing. It focuses on on-site logistics vehicle safety monitoring, driver behavior recognition, connected equipment, digital maintenance and intelligent after-sales support, helping businesses reduce safety risks and improve asset utilization and maintenance efficiency.

To our readers

Industrial vehicle service and maintenance is entering a new phase driven by knowledge, data and intelligent collaboration. The variety of models, complex component structures and long service chains mean that organizations cannot rely solely on a few experienced individuals to deliver consistent, standardized and high-quality service.

AI-Service organizes and applies enterprise knowledge from maintenance manuals, operating manuals, videos, fault codes, historical cases, parts information and warranty policies. It supports customer inquiries, fault diagnosis, repair guidance, parts recommendations, warranty claims, work-order collaboration and operational insights.

Based on existing product materials, this whitepaper describes the platform's capabilities, typical applications and implementation approach. References to supported, configurable or optional integrations are subject to the final product version, project scope and contractual agreement.

Executive summary

Generative AI is changing how enterprises serve customers, retain knowledge and support frontline staff. Valuable AI applications for industrial vehicles must extend beyond general chat into frequent, specialized and measurable activities: model documentation lookup, fault analysis, repair work, parts compatibility, warranty claims and customer service.

AI-Service brings together three capabilities:

- **Knowledge engine:** Turns documents, videos, fault codes and repair cases into enterprise knowledge assets that can be searched, queried and used by applications.
- **Intelligent service desk:** Provides knowledge-based answers, suggested replies, human collaboration, work-order routing and analytics across websites, WeChat, apps, enterprise service systems and after-sales mini programs.

- **Multimodal multi-agent maintenance:** Connects retrieval, diagnosis, repair planning, parts and warranty support. It accepts text, voice and images to assist with actionable diagnostic steps, repair plans, parts suggestions and warranty handling guidance.

For manufacturers, dealers and service networks, rental companies and large maintenance teams, the platform helps build lasting capabilities:

1. Consolidate scattered product, repair and service knowledge into a shared asset.
2. Reduce dependence on individual experience and improve service consistency.
3. Shorten information retrieval and fault investigation workflows to improve repair and parts service efficiency.
4. Feed conversations, work orders and usage feedback into continuous knowledge and operational improvements.

Chapter 1. Industry context: intelligent service and maintenance for industrial vehicles

1.1 From product delivery to lifecycle service

Industrial vehicles operate in warehousing, logistics, manufacturing, ports, airports, construction, cold chains and industrial parks. As customers demand better equipment availability, timely repairs, parts supply and lifecycle support, competition extends beyond individual products to comprehensive pre-sales, delivery, operation, maintenance, parts and warranty services.

After-sales and maintenance teams face many models, configurations and components, complex fault symptoms, inconsistent customer descriptions, wide service areas and varying staff experience. Paper manuals, disconnected files, separate system searches and informal consultations struggle to support scalable, standardized and responsive service.

1.2 Five common service challenges

Scattered knowledge and limited reuse of experience

Operating manuals, repair manuals, training materials, videos, fault-code tables and historical work orders are often scattered across departments, systems and personal devices. Expert experience is difficult to capture in a structured form, while new employees struggle to obtain accurate guidance quickly.

Diagnosis depends on personal experience

The same symptom may involve a motor, steering, hydraulics, controller, charger or several systems. Experience-driven diagnosis and repeated discussions increase downtime and make consistent judgments difficult.

Repair standards are hard to unify

Fault causes, procedures, precautions and good practices must be recorded, reviewed and reused promptly. Without structured knowledge and actionable guidance, service quality varies with staff experience.

Parts identification and compatibility checks are inefficient

Parts searches involve models, vehicle serial numbers, material numbers, component images, compatibility relationships and substitutes. Manual comparisons take time and can produce incomplete identification, incorrect recommendations or inefficient cross-system searches.

Warranty claims require complex communication

Claims involve vehicle information, nameplates, fault descriptions, repair records, parts data, company policies and supporting documents. Inconsistent policies increase review, communication and processing costs.

1.3 Augment professional expertise

Industrial vehicle service requires specialist knowledge and clear accountability. LLMs and agents should help service engineers, desk staff and technical experts access company knowledge, policies, cases and processes.

Combining AI preparation, human review and decisions, and business-system collaboration helps frontline staff obtain trustworthy information and standard procedures. It turns expert experience into an organizational capability that can be reused and improved.

Chapter 2. Platform overview: knowledge-driven, intelligent service with continuous feedback

2.1 Platform positioning

AI-Service supports industrial vehicle after-sales service and maintenance. Based on enterprise knowledge assets, it combines LLMs, multimodal understanding and multi-agent collaboration to assist customer service, technical support, repair work, parts operations and warranty claims.

It addresses knowledge that is hard to find, experience that is hard to share, problems that are hard to diagnose, disconnected service workflows and missing feedback. Organizations can move from isolated materials and individual experience toward shared knowledge, coordinated processes and continuous improvement.

2.2 Three layers of capability

Layer 1: knowledge engine - the enterprise knowledge foundation

The knowledge engine makes equipment documentation searchable, queryable and publishable. It organizes manuals, videos, fault codes, cases, models and tags to support intelligent applications, service desks and maintenance decisions.

Layer 2: intelligent service desk - customer service and human collaboration

Across websites, WeChat, apps, enterprise systems and after-sales mini programs, the service desk provides conversations, AI suggestions, human handover, work-order collaboration, customer information and operational insights to customers, frontline staff and managers.

Layer 3: multimodal multi-agent maintenance - task handling and decision support

For high-value after-sales scenarios, maintenance capabilities combine model documentation, manuals, historical cases, fault codes, parts bills of materials (BOMs) and warranty policies to assist diagnosis, repair recommendations, parts identification and claims guidance.

2.3 The service feedback loop

The platform's capabilities form the following cycle:

> Capture knowledge and business rules → Search and coordinate service → Assist diagnosis, repairs, parts and warranty → Feed back conversations, work orders and usage → Update knowledge and improve operations

This describes how business capabilities are organized. Data sharing, interfaces and deployment between modules depend on the actual project design.

2.4 Product components, delivery scope and prerequisites

AI-Service combines a knowledge engine, service desk, multi-agent maintenance, business-system integration and private-deployment governance. Its delivery boundaries distinguish standard capabilities, customer data prerequisites and custom project work.

Product capability	Core content	Scope and prerequisites
Knowledge engine	Dataset management for documents, videos, fault codes and repair cases; tags, indexing, retrieval, Q&A, application/agent publishing and review	Core capability; results depend on data quality, complete model/version tags, customer review and continuing maintenance
Intelligent service desk	Web, mobile and enterprise-system access; visitor interfaces, staff workspace, human handover, suggested replies, work orders, monitoring and analytics	Configurable capability; channels, branding, organizations, service regions, handover rules and response scripts need customer-specific configuration
Multimodal multi-agent maintenance	Intent recognition, task decomposition, retrieval, tool calls and result integration for pre-sales, diagnosis, parts, claims and translation agents	Scenario-specific capability; configure and validate against available knowledge, rules, models, permissions and human approval points
Business-system integration	Authorized data exchange with CRM, ERP/SAP, MES, work-order, parts, inventory and warranty systems	Requires interface authorization, field mapping, restricted permissions and integration acceptance; the platform does not replace source-system master data or approval authority
Private deployment and operational governance	Internal-network deployment, identity, role permissions, HTTPS, audit logs, monitoring, backups and environment management	Topology, resources, availability, recovery targets and security policies depend on the environment and contract

Core capabilities provide the foundation; configuration depends on business workflows; integrations require authorized interfaces and reliable data. Specialized algorithms, complete historical-data cleanup and customer-specific processes are custom project work. Final delivery follows confirmed requirements, technical proposals and contracts.

2.5 Industry capabilities developed through industrial vehicle LLM projects

Experience from pre-sales, after-sales and service collaboration combines industry knowledge, business rules and human responsibility into the following capabilities. Use each scenario within the available documentation, policies and connected systems.

Scenario	Typical input	Knowledge or tools	Output and human responsibility
Pre-sales consultation and selection	Customer requirements, working conditions, model keywords, target region	Product documents, operating-condition suitability, competitor specifications, CRM history and authorized SAP/inventory data	Specifications, comparisons and configuration suggestions; authorized staff must verify prices, delivery dates, certification and external proposals
Diagnosis and repair support	Model, fault code, symptoms and field text/voice/images	Fault codes, manuals, cases, system knowledge and troubleshooting rules	Code explanations, possible causes, inspection sequence, risk alerts and references; AI cannot replace field checks, safe procedures or final diagnosis
Parts identification and recommendation	Model, serial number, material number, part name or image	Parts catalogs, BOMs, compatibility, substitutes and authorized inventory	Candidate parts, compatibility evidence, alternatives and installation references; final fit, prices, stock and delivery depend on official systems and human checks
Warranty and claims guidance	Model, purchase date, fault type, repair records, nameplate or document images	Warranty rules, claims procedures, document checklists and cases	Completeness checks, procedures and precautions; authorized staff approve acceptance, eligibility, payment amounts and external commitments
Multilingual service and documentation	Inquiry text, product/repair documents and target language	Terminology, knowledge bases and translation models	Multilingual answers and document drafts; technical specifications, contracts, regulations and external documents require human proofreading and publication review

2.5.1 How task-specific agents collaborate

Complex requests can follow a shared entry point, intent recognition, task decomposition, knowledge and business-data calls, result integration, and human review or work-order routing. A diagnosis agent combines codes, model documentation and cases; a parts agent checks compatibility and inventory after authorization; a claims agent uses valid policies and document requirements.

Each task should retain relevant conversation context, knowledge references, tool calls, permission checks and human handover records. Low confidence, conflicting documents, missing knowledge, human/equipment safety issues, or authoritative decisions about price, stock and warranty should trigger human support, a work order or authorized approval.

2.5.2 Industry knowledge model and lifecycle

Traceable knowledge needs more than file contents. Metadata should identify the source, responsible department, model/series, system, component, fault type/code, operating conditions, industry, version, effective date and scope. Parts, warranty and case materials should also link to material numbers, compatibility rules, policy versions or case status.

A recommended lifecycle is:

> Inventory and authorization → Parsing, tagging and indexing → Pending review → Publication to selected applications or agents → Usage feedback → Scheduled or triggered updates → Retirement and version history

Official manuals, repair cases, policies, draft experience notes and obsolete documents need different publication and access rules. The platform can assist retrieval, summarization and gap identification, but must not automatically treat unreviewed material as an official repair, warranty or customer-service authority.

2.5.3 Human-AI service collaboration

A typical service flow is: customer or technician inquiry, AI interpretation and retrieval, assistance for standard questions or human takeover, work-order creation and routing, feedback on outcomes/satisfaction/tags, and improvement of knowledge and processes.

In the same service context, staff can review history, AI suggestions, related knowledge, similar cases and authorized business information before taking over or creating work orders. Summaries, suggested replies, common questions and knowledge usage inform analytics. Customer, order, inventory and claim information must follow organization, role, region and minimum-necessary access rules.

Chapter 3. Knowledge engine: making industrial vehicle knowledge searchable, understandable and usable

3.1 A dedicated knowledge base for industrial equipment

Industrial vehicle businesses hold valuable operating manuals, repair manuals, training materials, fault-code tables, cases, parts catalogs and instructional videos. The knowledge engine organizes these into a dedicated base for retrieval, Q&A and intelligent applications.

Documents, videos and fault codes can be managed as different dataset types with descriptions, tags and indexes, helping models and people find relevant information in specific situations.

3.2 Organizing knowledge by model, system and case

The answer to an equipment question depends on model, component, system, configuration, fault type and operating environment. Building knowledge requires a business-relevant classification and tagging scheme, not just uploading files.

Knowledge can be organized by:

- Vehicle model, series, configuration and applicability.
- Components or systems such as hydraulics, steering, controllers, drives, brakes and charging.
- Symptoms, fault codes, repair actions and risk alerts.
- Part names, material numbers, compatibility and substitutes.
- Warranty policies, claim processes, document requirements and review criteria.

Tags turn a document repository into a knowledge network that supports specific business judgments and service actions.

3.3 Capturing structured repair solutions and cases

Solutions and repair cases can record structured fields: fault description, analysis, resolution steps, conclusions, process images/videos and applicable models.

Staff can find similar previous cases quickly, while expert diagnostic reasoning and repair methods become reusable organizational assets that improve overall service capability.

3.4 Publishing knowledge as applications and agents

The engine supports simple applications and agents linked to relevant knowledge bases. Organizations can configure application details, input guidance and permissions for customer service, technical support, training, parts searches and repair guidance.

Shareable links, chat authorization, conversation logs, image-assisted Q&A, voice input and voice playback help frontline staff and customers use knowledge naturally.

3.5 Knowledge governance is a prerequisite

Before deploying intelligent service, organizations should:

1. Define information sources, responsible departments and update procedures.
2. Distinguish official manuals, experience-based cases, internal drafts and obsolete materials.
3. Label critical information with model, version, effective date and applicability.
4. Review professional conclusions, repair recommendations and warranty rules.
5. Use feedback to identify missing knowledge and conflicting content.

Chapter 4. Intelligent service desk: knowledge in every service interaction

4.1 Multichannel service entry

Inquiries arrive through websites, WeChat, apps, enterprise systems and after-sales mini programs. A unified service desk can handle consultations, technical support, service requests and business guidance.

Within a shared workspace, staff can see customer details, history, work orders, parts orders and suggested replies during a conversation, reducing window switching and repeated information checks.

4.2 AI assistance and human collaboration

For standardized, repetitive questions, the system can suggest replies based on historical conversations and the knowledge base; staff verify and send them as appropriate. Complex diagnosis, unusual policies and high-risk decisions can be referred to experts or designated staff.

This improves responses to common questions while retaining human professional judgment and accountability for complex ones.

4.3 Connecting conversations, customers, work orders and business information

The service desk can bring together:

- Customer profiles, contact details and interaction history.
- Conversation content, intent and processing history.
- Draft and historical work orders and their status.
- Parts orders, claims and after-sales information.
- Customer tags, equipment environment and service preferences.

When a request needs further work, relevant conversation details can populate a work order, reducing duplicate entry and preserving context for the next person.

4.4 Operational insights for knowledge and process improvement

Conversations, work orders and usage data help identify issues and improve service. The platform can support multidimensional historical search, clustering of common questions, metric monitoring, root-cause drill-down and improvement suggestions.

Businesses can identify:

- The most frequently asked questions.
- Models, components and fault types needing more knowledge.
- Product documents, training or service processes needing improvement.
- Staff or service regions needing more support.
- Requests that should become standard reference cases.

4.5 Global service and governance

For cross-region and overseas services, multilingual capabilities can be combined with access controls by organization, business line and service region. Sensitive-term rules, branding, multi-site configuration and account management help expand service access while maintaining consistent standards and governance boundaries.

Chapter 5. Multimodal multi-agent maintenance: from symptoms to actionable guidance

5.1 Retrieval, diagnosis, repair planning, parts and warranty

Industrial maintenance requires more than naming a component or explaining a code. Useful intelligence connects problem descriptions, enterprise knowledge, business rules and follow-up actions.

The platform assists this workflow:

> Retrieve references → Analyze the fault or business issue → Suggest repairs or handling steps → Recommend related parts → Highlight warranty or claims requirements

Use the workflow with actual enterprise documentation and policies, with necessary checks and decisions by service engineers, technical experts or business staff.

5.2 Text, voice and image input

Field descriptions are often inconsistent. Users may describe symptoms in natural language, record a voice message or upload an image of a faulty area, component or nameplate.

Multimodal input helps establish context, for example:

- Enter symptoms to obtain possible causes and inspection directions.
- Upload part images to assist identification of names or material information.
- Upload nameplates to assist extraction of vehicle serial numbers.
- Describe field problems by voice to reduce typing.

Image recognition, model identification, parts compatibility and serial-number extraction must be checked against official documents, business-system data and human verification before operational decisions.

5.3 Scenario 1: model and repair-manual lookup

When staff need a model's operating manual, repair manual, maintenance requirements or technical instructions, the platform can retrieve and answer questions using the model, keywords, system tags and document content.

This reduces manual searching and helps new staff learn products and service standards.

5.4 Scenario 2: fault-code lookup and diagnosis

For fault codes or abnormal behavior, the platform combines code tables, manuals, cases and repair rules to explain codes, suggest causes, order inspections and recommend next steps.

For example, if wheels continue rotating after startup, the system can help staff investigate motors, steering, hydraulics, controllers or charging systems. It must not replace field testing or safety procedures.

5.5 Scenario 3: step-by-step repair guidance

Structured cases and manuals can be organized into actionable guidance covering preconditions, procedures, precautions, tools and result verification.

Standardizing good repair experience helps staff across regions and experience levels work more consistently.

5.6 Scenario 4: parts identification, lookup and recommendation

Parts are essential to industrial vehicle service. The platform can use models, serial numbers, part names, material numbers, images and BOMs to assist identification, compatibility checks, recommendations and substitute suggestions.

Prices, stock, lead times, current material status and final compatibility remain subject to official business systems and authorization rules.

5.7 Scenario 5: warranty and claims guidance

For vehicle or parts claims, the platform can use warranty rules, claim policies, cases and document requirements to suggest procedures, required materials, completeness checks and precautions.

This reduces inconsistent guidance and missing documents. Authorized staff must make final warranty decisions, approve claims and make external commitments under valid policies.

Chapter 6. Typical applications: organizing capabilities around business problems

6.1 Battery, drive and brake maintenance Q&A

Typical questions: How do I inspect a lead-acid battery? How should lithium batteries, drive systems and brakes be maintained?

Platform capability: Uses manuals, training materials and cases to provide structured checklists, intervals, precautions and reference links.

Business value: Brings standard maintenance knowledge to frontline staff, reducing basic inquiries to experts and improving consistency.

6.2 Quickly locating operating and repair information

Typical question: How can I find the operating or repair manual for a specific model?

Platform capability: Uses model information, keywords, system tags and document indexes to locate references and answer related questions.

Business value: Reduces file-search time and improves information use by service and training staff.

6.3 Image-based parts identification and material lookup

Typical question: Can a photo of a multi-way valve or another component help identify its name and material number?

Platform capability: Combines image input with parts documentation, BOMs and product knowledge to assist identification and guide confirmation.

Business value: Shortens preliminary identification and information searches, supporting subsequent parts service.

6.4 Model, serial number and parts recommendations

Typical question: How can compatible or substitute parts be found using a model, serial number and part name?

Platform capability: Uses available model documentation, parts information and BOM data to assist compatibility searches and recommendations.

Business value: Improves parts service responsiveness and reduces manual searches and communication.

6.5 Symptom analysis and repair suggestions

Typical question: What could cause wheels to keep rotating after the vehicle starts?

Platform capability: Uses cases, manuals, codes and system knowledge to suggest investigation directions, checks and risk alerts.

Business value: Helps frontline staff form a diagnostic approach and provides structured information for expert assessment.

6.6 Warranty claim documents and procedures

Typical questions: What documents are needed for a forklift claim? How can the vehicle serial number be extracted from a nameplate?

Platform capability: Uses warranty rules, claim processes and checklists to identify required information and processing steps.

Business value: Reduces missing documents and repeated communication, improving consistency and efficiency.

6.7 Sales support, industry information and onboarding

Typical questions: How should industrial vehicles be selected for construction sites or cold stores? How can new employees quickly learn about lithium-battery products?

Platform capability: Makes product documents, industry scenarios, training materials and standard Q&A available to sales teams, channels and new staff.

Business value: Spreads product knowledge and industry experience faster, improving customer communication and training.

Chapter 7. Deployment, integration, security and operational governance

7.1 A phased implementation approach

Organizations can proceed according to business goals and data readiness:

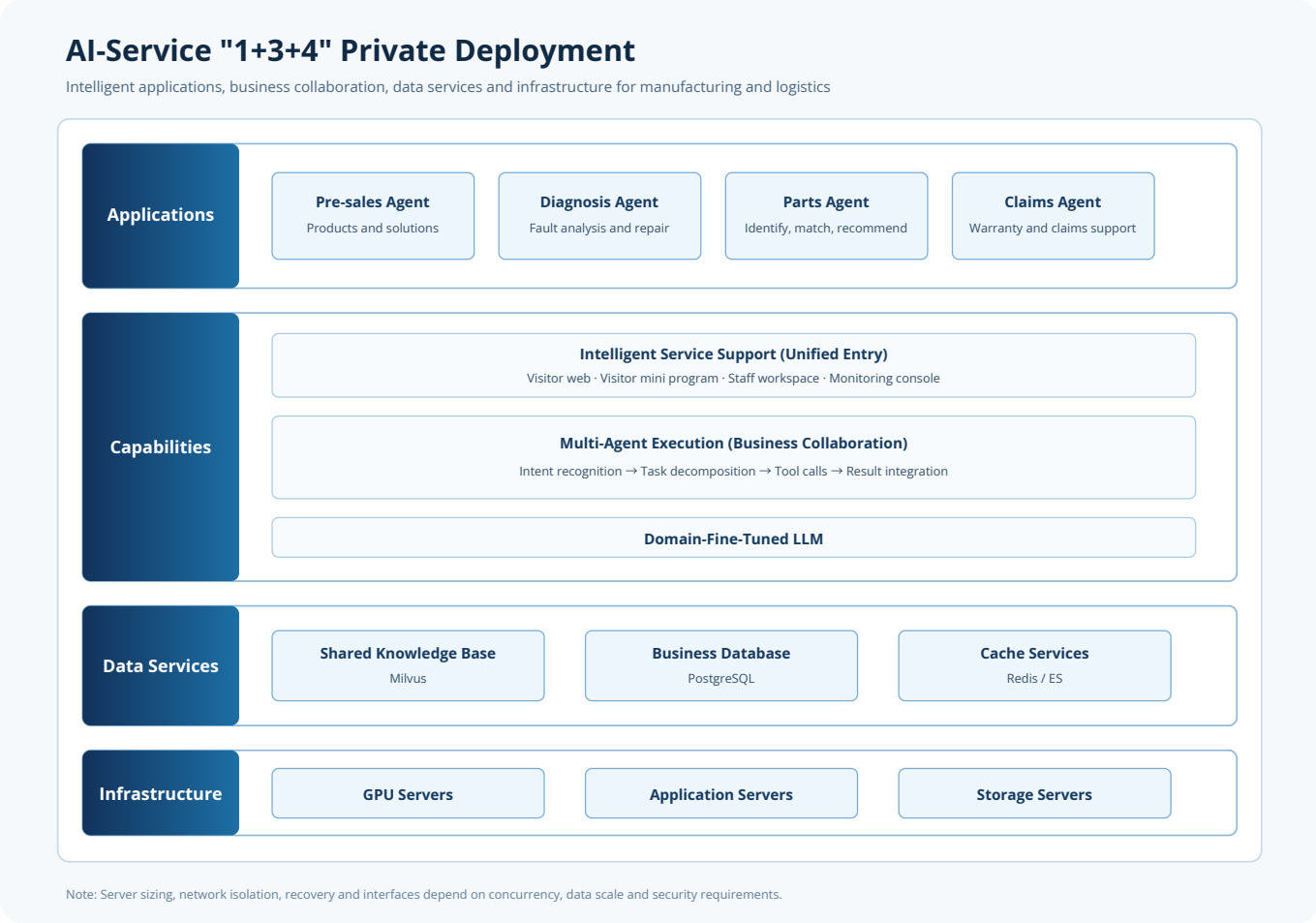
1. **Identify scenarios and inventory materials:** Prioritize use cases and collect manuals, cases, codes, parts information and warranty rules.
2. **Build knowledge and run trials:** Ingest sample documents and tables, create an initial knowledge base and validate key questions and scenarios.
3. **Coordinate processes and improve operations:** Embed intelligence in service touchpoints and improve human collaboration, work orders and analytics.
4. **Deploy and integrate as required:** Plan enterprise portal, WeChat mini program/official account, CRM, ERP, MES and EPC integrations within scope.

Trials should focus on ingestion, knowledge construction, scenario validation and evaluation. Interface development, custom functions, specialized algorithm optimization and full historical-data cleanup belong to the formal project scope agreed by both parties.

7.2 Deployment and access

AI-Service uses a "1+3+4" private-deployment architecture to connect intelligent applications, business collaboration, data services and infrastructure within enterprise networks. It serves manufacturers, logistics and warehousing businesses requiring control and security over technical documents, customer information, service records, parts and maintenance data.

7.2.1 Private-deployment architecture



The platform has four layers:

1. **Application layer:** Specialized pre-sales, after-sales diagnosis, parts recommendation and claims agents.
2. **Business capability layer:** A service support system, multi-agent execution system and domain-fine-tuned LLM. The unified service entry covers visitor web and mini program interfaces, staff workspace

and monitoring console. Agents handle intent, task decomposition, tools and results; domain models support specialist Q&A, diagnosis and business collaboration.

3. **Data services layer:** A shared knowledge base (Milvus), business database (PostgreSQL) and cache services (Redis/ES) support retrieval, business data and high-frequency access/search acceleration.
4. **Infrastructure layer:** GPU, application and storage servers support inference, business services and data persistence.

7.2.2 Agent collaboration and data access

Customer inquiries, field repair requests, parts searches and claims guidance can follow a shared entry point, intent recognition, task decomposition, authorized data calls, result integration, and review or work-order routing:

1. Users submit text, voice, images and equipment information through web, mini programs, staff workspaces or other channels.
2. The service support system identifies users, conversation context and intent, then delegates complex requests to agent orchestration.
3. Agents call domain models, knowledge bases, databases and authorized external tools as needed.
4. The system combines references, business information and recommendations into traceable answers, diagnostic steps, parts suggestions or claims checklists.
5. Field safety, final parts compatibility, warranty approval, prices, inventory and external commitments retain human checks, authorized approvals or work orders.
6. Conversations, work orders and feedback can improve knowledge, cases and operations after appropriate de-identification, review or permission checks.

This combines general conversational ability with industry knowledge, business data and human responsibility, keeping model output from replacing safe procedures, approvals or final professional judgment.

7.2.3 Access, environment configuration and trials

Access can be provided through PCs, mobile devices, enterprise portals, WeChat mini programs or official accounts. Private deployments should separate inference, applications and storage physically or logically, planning scale according to concurrency, knowledge size, parsing volume, image/voice use and retention.

Trials may use online SaaS or private deployment according to security and management requirements. Server specifications, GPU types/counts, storage, network isolation, recovery, migration, accounts and interfaces should be determined after assessment. Production integration should use minimum-necessary permissions, separate validation environments and auditable rollback.

7.2.4 RAG retrieval and model strategy

AI-Service uses browser/server access over HTTPS, with separate knowledge-base, staff and visitor interfaces. Retrieval-augmented generation (RAG) first locates trustworthy material through vector retrieval, keyword/structured filtering and reranking, then generates answers, summaries or suggestions within that context. Knowledge applications, service desks and agents can share a traceable enterprise foundation.

Model capabilities follow a strategy of local lightweight inference and centralized conversational services:

- Embedding and reranking can run through vLLM or a validated equivalent on internal GPU nodes, alongside Milvus retrieval.
- General conversation and translation can use centralized models through a shared gateway; external APIs require agreed data-transfer, security, de-identification and authorization conditions.
- Model weights, prompts, knowledge, business tools and conversations must be permission-isolated; business services must not consume GPU capacity reserved for inference.
- Large models need not be deployed separately for every customer; deployment depends on security, concurrency, cost and latency.

7.2.5 Kubernetes microservices reference architecture

Formal deployments can use Kubernetes microservices with a multi-node highly available control plane and separate CPU/GPU worker roles. An ingress chain of WAF (such as samwaf), API gateway (such as APISIX) and applications provides authentication, routing, limits and auditing. Knowledge, staff and visitor interfaces can route to separate backends.

Workload type	Example components	Purpose and deployment principles
Deployment	Staff/visitor backends, RAG retrieval/loaders/workers, gateways and frontends	Stateless services supporting scaling, rolling updates and recovery; prefer CPU workers
StatefulSet	MySQL, PostgreSQL/TimescaleDB, MongoDB, Redis, Neo4j, Elasticsearch, RabbitMQ, RocketMQ, etcd, MinIO and Milvus	Persist state on volumes; select components for actual workloads rather than deploying every component
DaemonSet	CNI, node DNS cache, monitoring collectors and NVIDIA GPU device plugin	Provide networking, observability and GPU discovery on applicable nodes

GPU nodes can use taints such as `gpu=nvidia:NoSchedule`, labels, `nodeAffinity` and `toleration` to bind embedding, reranking and GPU vector-query workloads to GPUs, separating them from service desks, gateways and knowledge management. Images may use private registries such as Tencent Cloud TCR and optional public-image caching; registry products and security policies depend on the customer environment.

7.2.6 Resource and runtime reference

The following is a reference baseline for one physical server, not a universal purchasing specification. Assess capacity using concurrent conversations, knowledge size, parsing load, image/voice share, model access and availability targets.

Component	Reference specification	Quantity or notes
CPU	64 cores; base frequency at least 2.25GHz and turbo at least 3.4GHz	2 processors, 128 cores total
Memory	DDR4/DDR5, at least 3200MHz, at least 64GB per module	8 modules, 512GB total
GPU	NVIDIA, at least 24GB VRAM per card	2 cards
System disks	NVMe SSD 480GB	2 disks; RAID1 recommended
Data disks	NVMe SSD 2TB	4 disks; RAID5 or RAID10 according to reliability goals
RAID and power	RAID 0/1/5/10 controller; redundant power	At least 8 controller ports; dual power supplies with headroom
Networking and virtualization	Internal/external connectivity, DNS, gigabit/10-gigabit ports; virtualization	Proxmox VE, VMware ESXi or a confirmed equivalent
OS and containers	Ubuntu 24.04 LTS VMs, containerd and Kubernetes 1.3x	Subject to project compatibility and security baselines

Reference room conditions are 18-27°C, preferably 22°C ±2°C; relative humidity 40%-60% RH; temperature change no more than 5°C/hour. Use precision cooling and temperature/humidity monitoring, with cooling capacity planned at 1.2-1.5 times equipment power. Calico CNI and VXLAN Overlay can provide Pod networks; `nvidia-device-plugin` can report GPU resources to Kubernetes as a DaemonSet.

7.3 Security, network isolation and access governance

Knowledge may include sensitive product specifications, customer details, parts, service records and warranty rules. Establish governance for identity, permissions, networks, data and model access from the outset.

Depending on project needs, the platform can provide or configure:

- Data isolation by tenant, organization, business line and service region.
- Fine-grained roles, function/data permissions and single sign-on (SSO).
- HTTPS, shared authentication, rate limiting, audit logs and account authorization.
- Separate visitor, staff and knowledge-management entry points and session boundaries.

- Knowledge access by classification, audience and publication status.
- Auditing for high-risk agent tools, external interfaces and human handovers.

Kubernetes deployments may separate business (`llm`), vector retrieval (`milvus`), logs (`loki`), monitoring (`monitoring`) and agent sandboxes (`agent-sandbox`) into namespaces. For tenants, high-risk tools and sandboxes, `NetworkPolicy` can enforce ingress allowlists. Other services also require network controls, gateway authentication and minimum exposed ports; namespaces alone do not enforce network isolation.

Prefer ClusterIP for internal services and controlled external access through gateways or, where appropriate, NodePort. Place external entry points behind a WAF with gateway filtering, routing and access control. Classify which knowledge is customer-facing, internal-only or requires expert approval.

7.4 Persistence, retrieval and data/model boundaries

A shared persistence layer can store databases, knowledge indexes, weights, logs and backups. Private environments may use NFS or approved enterprise shared storage behind a StorageClass. A `Retain` policy on critical volumes avoids automatic backing-volume removal when a PVC is deleted.

Data services can combine components according to scope:

- Relational/business data: MySQL 8.0 and PostgreSQL/TimescaleDB.
- Documents, conversations and graphs: MongoDB, Neo4j, Elasticsearch and Redis.
- Messaging and asynchronous tasks: RabbitMQ, optionally with delayed exchange support, and RocketMQ.
- RAG retrieval: Milvus with etcd metadata, MinIO object data and optional GPU Query Nodes.
- Object storage: separate MinIO instances or equivalent storage for vector data, logs and business materials as appropriate.

These are reference technologies, not a fixed customer deployment list. Choose based on data size, concurrency, availability, operational skills and IT standards. Shared PVCs can supply model weights to inference Pods. GPU nodes host authorized lightweight inference and vector retrieval rather than general business services. External model/API requests require agreed data boundaries, encryption, de-identification, permissions and auditing through a shared gateway.

7.5 Backup, recovery and observability

The deployment plan should define backup scope, frequency, retention, recovery ownership and exercises. Kubernetes CronJobs can back up middleware and configuration; the project reference baseline is every two days in the early morning to NFS or persistent object storage. Critical MySQL, PostgreSQL and MongoDB databases also need periodic logical backups, multiple recoverable versions and offline or off-site copies under customer policy.

Regularly test volume retention, database recovery, knowledge-index rebuilding, weight loading and critical configuration restoration. Recovery point/time objectives (RPO/RTO), frequency and off-site recovery levels must be agreed in the contract and project plan.

Observability may include `node-exporter`, `metrics-server`, Prometheus and Grafana for nodes, containers, GPUs, services and business metrics, plus Loki and Alloy or an existing logging system. Alerts should cover utilization, health, API success/latency, inference and retrieval performance, queues, storage, backup results and abnormal access.

7.6 Evaluation recommendations

Choose measures suited to the actual trial or project:

- Effective answer rate, retrieval hit rate and completeness of supporting references.
- Fault investigation and document retrieval time.
- First-time fix rate and repeat work orders.
- Parts identification and recommendation accuracy.
- Claims processing time and document completeness.
- Staff usage, human intervention and satisfaction.
- Knowledge freshness, issue resolution and completed human handovers.
- Availability, API success, backup success and recovery exercise results.

Agree targets using the customer's baseline, data quality, workflows, scope and confirmed service levels.

Chapter 8. Conclusion: continuously improving service through trustworthy knowledge

Intelligent industrial vehicle service requires more than an LLM application. It involves enterprise knowledge, service workflows, data governance and collaboration.

With a knowledge engine, intelligent service desk and multimodal agents, AI-Service assists diagnosis, repairs, parts and warranty claims, helping organizations develop lifecycle service capabilities.

As knowledge, business feedback and applications grow, organizations can move from solving isolated issues to continually improving maintenance, service efficiency, customer experience, channel collaboration and professional reputation.

Appendix A. Key terms

Term	Definition
Industrial vehicle	Equipment for warehousing, logistics, manufacturing, ports, cold chains and construction; this paper focuses on forklifts and related applications.
Knowledge engine	A platform for organizing, indexing, retrieving, querying and publishing specialist enterprise knowledge.
Intelligent service desk	Multichannel service, AI assistance, human collaboration, work orders and analytics for customers and frontline staff.
Multimodal	AI understanding and processing of text, voice, images and other input types.
Multi-agent	Agents with different responsibilities collaborating on retrieval, diagnosis, repair planning, parts and warranty.
Fault code	An abnormal condition recorded by equipment controls, interpreted alongside manuals, cases and field inspections.
Parts BOM	A bill of materials describing components, material information and their relationships.
Compatible part	A part compatible with a particular model, configuration or component.
Substitute part	A replacement that meets business, technical and authorization requirements.
SSO	Single sign-on: access to authorized applications through a shared enterprise identity.

Appendix B. Public customer cases and platform data

B.1 How to use this appendix

This appendix summarizes AI-Driven website case studies and metrics publicly presented as of September 2026. Unless noted otherwise, figures follow the website's definitions. Results depend on customer conditions, data quality, implementation scope, measurement periods and workflows; **they are not promises of outcomes for other customers**.

The following is a paraphrased summary of public website information.

B.2 Platform scale and industry data

According to the Core Metrics & Measurement Basis page, as of July 1, 2026, the platform had cumulatively connected **230,000 industrial vehicles**, served **5,000+ enterprise customers** and covered **40+ countries and regions**. In January-June 2026 it processed an average of **3,000+ maintenance work orders and diagnostic requests per day**.

Accumulated assets include **100,000+ repair work orders**, **5,000+ fault types**, **3 billion+ fault-code records** and **200,000 pages of multilingual repair manuals**. These cumulative assets differ from concurrent online vehicles, model fault coverage and daily processing volume.

Reported measures include complex-fault diagnosis accuracy above **95%**, fault semantic parsing accuracy of **97.3%**, average complex-diagnosis time reduced from over 24 hours to under 3 minutes, remote first-contact resolution of **81%**, automatic handling of frequent issues of **80%**, and annual availability of **99.9%**. Each has a distinct population and definition; consult the metric page.

B.3 Case 1: XCMG - knowledge for many models and a global service network

The XCMG case describes a four-layer access, service, data and business architecture addressing model variety, scattered knowledge and global response demands. It incorporated **198+ models** and **5,800+ core service knowledge items**, with RAG, multimodal visual interaction and intelligent pre-sales selection.

Reported outcomes include reducing manual retrieval from about 30 minutes to about 1 minute, automated responses or interception for over 60% of faults, and reducing assisted onboarding from 6 months to 2 months. Data preparation, core trials, initial acceptance and rollout took about 3 months. These project-specific results should not be extrapolated.

B.4 Case 2: Linde (China) - lifecycle service connecting vehicles, cloud and people

The customer case page describes an AI sales assistant and AI service engineer assistant combined with ADAS and T-Box data, linking safety perception, pre-sales value communication and after-sales support.

The case presents 24/7 intelligent service, vehicle-cloud-human collaboration and predictive-maintenance suggestions, with project indicators for lifecycle value, service revenue and service premiums. It illustrates lifecycle service beyond equipment sales; commercial indicators are not universal cross-customer conclusions.

B.5 Case 3: Noblelift - claims and parts operations

The customer case page describes fragmented service channels, claim reviews and parts matching. Unified channels, a claims automation agent and parts matching support more disciplined global after-sales operations.

Reported indicators include **90%** lower claims review costs, parts recommendation accuracy above **90%**, **80%** higher knowledge-query efficiency and **97.3%** fault semantic recognition. These figures apply to the case's business scope and measurement definitions, not to other contexts without qualification.

B.6 Sources

- AI-Driven website
- Customer case studies
- Core Metrics & Measurement Basis