

DeepSopho 产品白皮书

连接知识、技能与协作的企业智能中枢

滴普科技



DeepSophon 以本地安全为底座，凭借场景落地、模型路由、效果评估、组织赋能五大核心能力，顺应企业 AI 应用趋势与 Agent 落地高速增长的行业趋势，助力企业完成从 AI 工具试用向全员岗位生产力重构的转型。

核心结论	
 安全可信，本地优先	端到端数据不出域，多层防护体系，满足企业合规与隐私要求。
 场景驱动，生产力落地	覆盖核心业务场景，AI 员工深度融入流程，显著提升人效与质量。
 模型路由，智能最优	多模型动态路由，按任务特性选择适配模型，降本增效。
 业务闭环，价值可衡量	从任务执行到效果评估形成闭环，驱动持续优化与价值沉淀。
 组织进化，能力复制	知识沉淀与能力复用，沉淀数字资产，加速组织智能化进化。



关键洞察



AI 员工从“工具辅助”走向“岗位生产力引擎”，成为组织增长新动能。



场景深耕与流程再造是生产力落地的关键，ROI 持续提升。



模型路由与成本优化并重，实现性能、成本与安全的平衡。



业务数据闭环驱动决策优化，形成可复制的增长飞轮。



组织与人才能力升级是规模化复制的核心保障。

企业 AI 进入规模化落地窗口

高普及•低规模化•强治理•本地算力

驱动趋势：从能力跃升到落地深化



Agentic AI 加速落地

从对话问答到自主决策，
业务闭环能力显著增强



业务流程自动化
端到端闭环场景增多



AI Infrastructure 升级

算力、存储、网络与模型能力持续
迭代，支撑更大规模应用



算力需求爆发
推理成本持续优化



Local-first 成为主流

数据安全、合规与时延要求
推动本地化部署优先



本地部署占比提升
混合云成为常态



🔍 关键发现



普及率持续提升

企业 AI 采用率持续提升，
同比 + 16pct



规模化仍低

规模化落地仍处早期，
仍有较大空间



治理优先级上升

安全合规要求上升



本地算力需求强劲

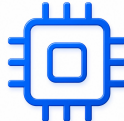
本地/混合部署需求增加

💡 启示



聚焦落地场景

优先选择可规模化复制
的行业与业务场景



布局基础设施

关注算力、存储、网络
与模型平台的综合能力



强化安全治理

数据安全、权限治理与
合规能力是刚需赛道



把握本地机遇

本地化部署与混合云
将带来长期增量市场

从能不能用 AI，到能不能安全、稳定、低成本交付 AI 员工

关键维度	传统云端 AI 助手	风险	DeepSophon 本地 AI 员工	机会	业务影响
 安全合规	数据上传云端，存在泄露风险 难以满足行业合规与审计要求	高风险	数据本地处理，端到端可控 满足合规与审计，企业数据不出域	合规可靠 安全可信	降低合规与数据安全风险
 弱网离线	依赖稳定网络，弱网体验差 断网不可用，影响业务连续性	高风险	支持弱网与离线运行 保障业务连续，可随时可用	随时可用 业务连续	保障关键业务连续稳定
 成本可控	按量付费，长期成本不可控 Token 消耗与调用费用高	高成本	一次部署，长期使用成本低 本地推理，外部 API Token 调用减少/可控	长期成本可控 ROI 更优	显著降低 TCO 与运营成本
 行业冷启动	通用模型，行业知识适配慢 冷启动周期长，落地效率低	低效率	内置行业知识与最佳实践 开箱即用，快速冷启动落地	快速落地 价值更快	加速行业落地与规模复制



↓ 90+
数据泄露风险



99.9%+
离线可用率



↓ 60+
长期 TCO 降低



↓ 3x+
长期 TCO 降低

战略结论：需求已从 “能用 AI” 迁移至 “如何安全、稳定、低成本交付 AI 员工”。DeepSophon 本地 AI 员工是企业级 AI 落地的更优解。



核心能力

AI 助理

知识管理

内容创作

数据分析

流程自动化

集成开放

产品价值链

连接与整合
连接人、信息与系统
打破孤岛，统一入口

理解与洞察
AI 理解业务与数据
提供洞察与建议

协作与执行
多人协同与任务驱动
提升效率与协作体验

沉淀与复用
知识与流程沉淀复用
形成组织资产

进化与创新
能力持续升级与扩展
驱动业务创新增长

设计原则

安全合规为先

用户体验至上

开放兼容集成

标准化与可配置

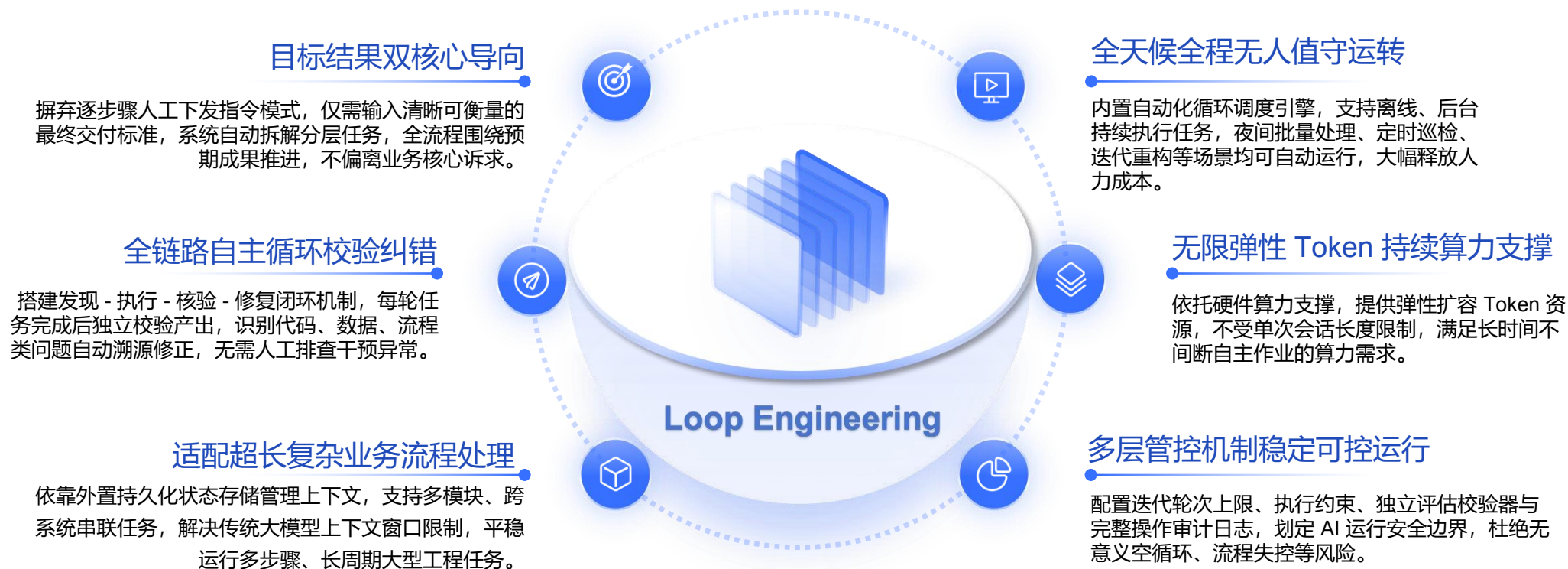
可观测与可运营

BS 系列算力底座 · DeepSophon Optimizer 模型服务 · Loop Engineering系统搭建 · 业务工作台



统一的算力底座 · 强大的模型服务 · 灵活的AI员工配置 · 安全可信的运营保障 驱动行业智能化转型，助力业务创新与增长

在 DeepSophon 中构建自主循环系统，将 AI 从被动的问答工具，升级为能够自主规划、执行并完成复杂任务的智能代理。



依托 Loop Engineering 六大核心能力，DeepSophon 打破传统 AI 单次交互局限，以自治闭环体系承载长周期复杂业务，兼顾无人值守自动化效率与全流程运行管控，实现从问答式工具到自主智能代理的全面升级。

以BS200本地算力与BS300高性能推理协同，重构模型、系统与应用交付，实现场景驱动、架构协同、模型闭环、应用落地与安全可控的企业级智能底座，支撑企业AI稳定高效规模化运营



场景驱动

架构协同

模型闭环

应用落地

安全可控

DeepSophon 以统一的知识本体为中心，整合业务领域、实体关系、行业知识、指标度量、规则约束和文档知识，解决企业数据分散、知识难复用的问题，帮助企业快速构建面向不同岗位和场景的 AI 能力。

专业 Skills 分类

客户服务	运营管理	供应链管理	设备运维	财务管理	人力资源	数据分析
咨询解答	经营分析	需求预测	故障诊断	预算管理	招聘筛选	数据清洗
工单处理	指标监控	库存优化	预防维护	成本核算	绩效分析	数据建模
客户洞察	异常预警	采购计划	巡检分析	资金预测	培训规划	可视化分析
满意度分析	报表生成	供应商管理	备件管理	合规审计	员工服务	洞察报告

价值亮点

- 统一的知识本体，打通数据孤岛，沉淀组织智慧资产
- 专业 Skills 封装最佳实践，开箱即用快速落地
- 行业模板加速复制，降低实施门槛与交付周期
- 持续进化与协同共建，形成正向增长的知识飞轮

DeepSophon Optimizer 作为DeepSophon 的核心算力引擎，围绕模型路由、Agent 层卸载与硬件定向优化构建六大技术优势。通过Token 经济性、弹性卸载、Agent 适配、服务化集成、运维可观测与BS硬件深度优化，实现低成本、高性能的企业级推理底座，驱动业务高效落地。



Token 经济性

更低成本 · 更高吞吐 · 更优 ROI

- 多层压缩与精简推理
- 长上下文稀疏化处理
- 动态 KV 缓存复用



Offloading

负载更均衡 · 延迟更低 · 体验更稳定

- 模型层弹性卸载
- KV/Embedding 分流
- 本地-边缘-云协同



运维可观测

可观测 · 可告警 · 可追溯

- 全链路指标采集
- 实时告警与根因分析
- 可视化仪表盘与报表



Agent 适配

更强 Agent 能力 · 更快落地 · 更少改造

- 任务级能力映射
- 工具/函数自动路由
- 记忆与上下文托管



服务化集成

易集成 · 可复用 · 快速对接业务系统

- 统一 API 网关
- 标准化服务协议
- 多租户与权限隔离



硬件定向优化

性能更强 · 资源利用更高 · TCO 更优

- 针对 BS 系列深度优化
- 精度/并行/显存跨调优
- 算子融合与内核优化



一体化引擎驱动业务创新，以更低成本、更高性能、更好体验，构建企业级智能化核心底座

让每个任务跑在最合适的模型与算力上



AMD芯片，一体化设计，电源内置，极致液冷散热

稳定功耗 80W-165W

极限功耗 120W-176W



项目	参数
尺寸	199mm × 199mm × 199mm
重量	5300g
材质	全铝合金
强度	1.0mm SECC
iGPU	AMD® Radeon™ 8060S Graphics
统一内存	LPDDR5X - 8000MHz 128GB
显存	96GB Max
SSD	PCIE 4.0 2TB
SoC	AMD® Ryzen™ AI Max + 395
电源	1U 400W 92 + 白金效率
高速接口	USB4(Display) × 2
多显接口	DP1.4 × 1 + HDMI2.1 × 1
高速网络	10GbE × 1 + 2.5GbE × 1
无线网络	WiFi-7 + 蓝牙5.4
操作系统	Ubuntu 24.04

NVIDIA芯片，分体设计，小巧外观，支持堆叠级联

极致性能

在FP4精度下提供高达1PFLOP 的AI性能。

支持堆叠

支持2台设备直连，AI模型支持规模高达4050亿参数推理。



项目	参数
尺寸	150mm x 150mm x 50.5mm (~1.2L)
重量	1200g，成套包装材料重量2400g
CPU	20 core Arm, 10 Cortex-X925 + 10 Cortex-A725 Arm
GPU	Blackwell 架构
CUDA Core	6144
Tensor Core	第五代，数据格式支持TF32/FP16/BF16/INT8/FP8/FP6/FP4
统一内存	128GB LPDDR5x 统一内存
存储	2x M.2 NVMe SSD (2TB & 4TB 选项)
电源	240W Type-C适配器（外置）
IO接口	4 个 USB4 Type C
显示接口	1x HDMI 2.1a
高速网络	200G x 2 QSFP (ConnectX-7), 1x 10GbE RJ-45
无线网络	WiFi-7 + 蓝牙5.4
操作系统	Ubuntu 24.04

BS300D堆叠级联方案

大参数规模

284B总参数规模，支持复杂任务理解与推理。

稳定并发

支持多任务同时处理，多路使用下保持稳定响应。



满血DeepSeek-V4 Flash

01 算力规模

BS300x 2

284B
总参数规模

稳定并发

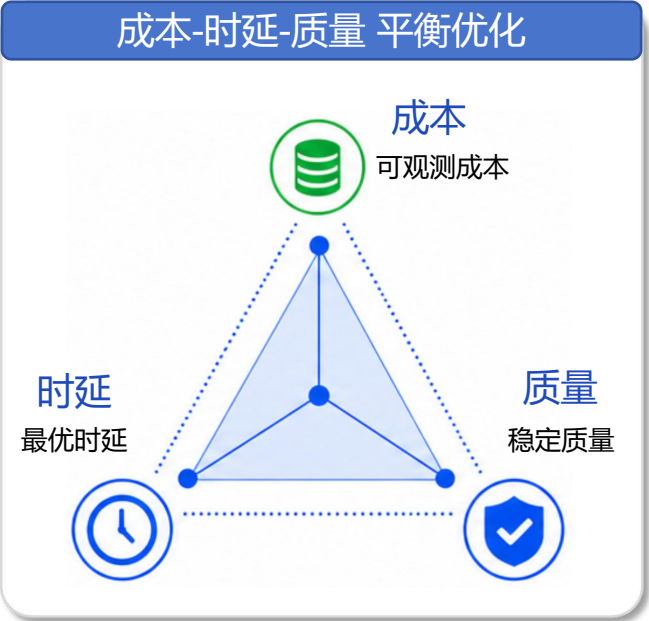
多任务同时处理，多路使用下保持稳定响应

02 互联

200G x 2 QSFP
(ConnectX-7)高速网络连接

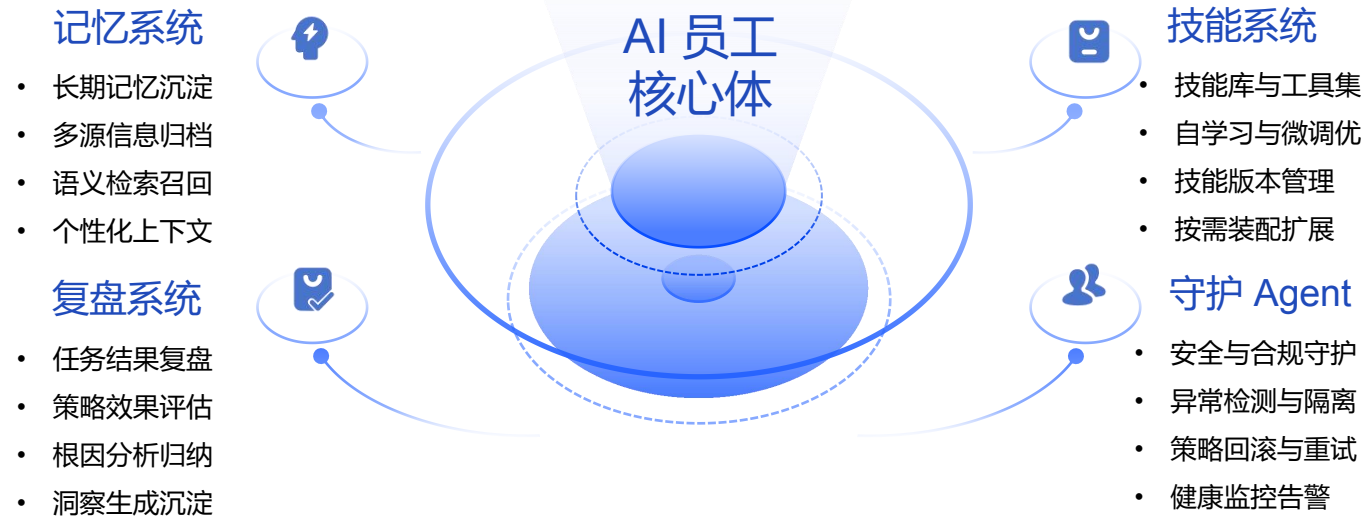
256G
最高显存

本地优先、云边协同、可观测成本



以任务为中心的智能调度，动态平衡成本、时延与质量，释放算力价值，支撑业务高效稳定运行

持续进化 AI员工通过记忆沉淀、技能迭代、复盘优化与守护自愈，持续提升能力、降低风险、保障业务稳定运行，形成正向增强的闭环生态。



自愈与异常处理机制				
异常类型	检测方式	自愈策略	恢复目标	审计留痕
模型异常	输出质量监控	切换备用模型，动态降级	返回可用结果	日志+指标
工具失败	执行状态监控	重试/改用替代工具，降级流程	完成任务或降级完成	日志+事件
数据缺失	输入完整性校验	补全数据/启用默认值，请求人工补充	数据可用	日志+快照
策略失效	效果指标偏离	回滚至稳定策略，自动重新规划	恢复策略有效性	日志+版本
系统风险	风险规则检测	隔离执行/熔断保护，告警通知	保障系统安全	日志+告警



以自进化驱动组织知识与能力的持续增长，让 AI 员工持续优化、稳定性持续提升。

围绕企业全生命周期，AI 员工提速增效



精细运营
统一运营，精准决策



产业服务
全域协同，资源整合



企业赋能
AI 员工，提速增效



安全合规
数据可控，全程留痕

园区运营中枢

数据驱动 | 智能决策
协同高效 | 安全可控

四大 AI 员工，覆盖园区治理与企业服务核心场景



智管家
园区运营助手

- 智能监测与预警
- 政策落地与执行跟踪
- 事件处置与闭环管理
- 数据看板与决策建议

治理提效 | 风险可控



智招
招商服务助手

- 企业画像与线索挖掘
- 招商政策智能匹配
- 项目跟进与进度管理
- 投资分析与决策建议

精准招商 | 转化提升



智办
企业服务助手

- 办事指南智能问答
- 材料预审与表单填报
- 审批流程智能协同
- 办件进度主动推送

办事高效 | 体验升级



智库
产业智库助手

- 产业研究与趋势洞察
- 规则解读与影响评估
- 产业链图谱与分析
- 专题报告智能生成

洞察前瞻 | 决策有据

方案
价值成效



5秒

智能问答响应时长
平均响应时长缩短至 5 秒以内



-90%

重复咨询量下降
AI 智能分流，重复咨询下降 90%



-97%

办事材料错误率下降
材料预审 + 智能校验，错误率下降 97%



100+名

人效提升
单个 AI 员工相当于 100+ 名专员

综合收益



治理更精细
数据驱动，科学决策



服务更高效
办事提速，体验升级



招商更精准
线索优选，转化提升



企业更满意
全周期陪伴，成长无忧



运营更低成本
自动化助力，降本增效



生态更繁荣
产业集聚，协同共赢

数据敏感型事件处置协同方案



以“AI 事件处置专员”为核心，连接接处警数据、视频监控、警综平台、互联网、电子卷宗、地理空间和物联网设备等多源数据，形成警务数据与能力底座。

统一资源池化与协同调度

事件接入

智能受理

分级研判

指令下达

警情多源接入

智能分级与派单

事件电话

响应设备

互联网

视频监控

移动警务

I级（重大）

II级（较大）

III级（一般）

IV级（轻微）

智能派单

指挥中心

属地单位

业务部门

联勤单位

ZC研判

线索汇聚

智能研判

关联分析

研判报告

智能线索分析

案件关系图谱

人像识别

关系挖掘

语义理解

车辆轨迹

行为分析

时空分析

关联实体

关联关系

关键线索

风险等级

事项办理

事项办理

文书生成

流程审批

归档管理

警能办公助手

协同与审批

公文起草

法律条文检索

会议纪要生成

报表自动生成

知识问答

部门提报

审核审批

会签流转

归档存证

已提交

进行中

进行中

已完成

数据与能力底座

事件数据

视频监控数据

业务平台数据

互联网数据

内部业务数据

电子档案数据

地理空间数据

物联网设备数据

医疗支付风控解决方案



基于医疗本体数据集与智能规则引擎，构建 AI 风控专员，实现医疗支付审核精准识别、证据可追溯与处置闭环，提升审核效率并降低支付风险。

全量数据驱动 · 智能规则引擎 · 精准异常识别 · 可解释可追溯

数据汇聚与治理



医疗机构数据

- 电子病历
- 医嘱处方
- 检验检查
- 手术操作等



支付数据

- 药品目录
- 支付政策
- 限定条件
- 历史结算等

AI 风控专员（智能体）

语义理解与结构化



语义理解与结构化

语义理解与结构化

语义理解与结构化

警情接入

智能受理

分级研判

指令下达

规则引擎与校验

政策规则

支付规则

限定条件匹配

临床规则

适应症校验

用法用量校验

组合规则

多规则联动

深度校验

校验结果与证据链



合规通过



风险预警



异常处置

定位原因 | 依据条款
影响金额 | 处置建议

智能审查与闭环管理

药品审查

适应症匹配
用法用量校验
禁忌与重复用药
超限定用药识别

医保限定

年龄限定
适应症限定
诊断限定
支付条件匹配

耗材审查

耗材编码校验
适用范围校验
数量与规格校验
高值耗材专项规则

人工复核

高风险案件
争议案件
系统建议
证据链展示

处置与反馈

拒付 / 部分拒付
改进建议
规则优化
模型迭代

典型应用成效



异常识别率提升
30%+



处置准确率提升
25%+



人工审核效率提升
60%+



损失挽回率提升
20%+

AI 轮机员：知识驱动·经验沉淀·故障快速定位·决策可靠可追溯



船舶机电设备维护知识库

设备手册

- 厂家手册
- 操作指南
- 维护保养

图纸与资料

- 系统图
- 结构图
- 布置图

校验记录

- 参数校验
- 性能测试
- 追溯数据

备件与物料

- 备件清单
- 物料规格
- 库存信息

合规文档

- 公约规范
- 检验要求
- 船级社规则

语义检索与问答

故障诊断与定位

维修方案推荐

风险预警与建议

AI 轮机员

多模态理解

推理分析与根因归因

工单生成与记录

知识反馈与持续优化

多角色协同，价值闭环

轮机长
决策与统筹

- 全船设备健康总览
- 关键风险识别与决策
- 维护计划审批与优化

大管轮
管理与优化

- 设备状态监控分析
- 维护策略制定与监督
- 备件计划与成本控制

二管轮
执行与落地

- 维修工单执行指导
- 操作步骤与注意事项
- 现场问题快速求助

三管轮
操作与记录

- 日常巡检与数据记录
- 异常上报与初步处理
- 知识反馈与经验沉淀

知识沉淀 → 经验复用 → 能力提升 → 安全运营

应急维修智能 workflow

异常发现
报警/巡检/上报

问题描述
语音/文字/拍照

智能检索
知识库秒级匹配

诊断定位
根因分析与确认

方案推荐
步骤/工具/备件

执行维修
指导执行与记录

验证闭环
效果验证与归档

THANKS



官 网: www.deepexi.com

媒体访问: pr@deepexi.com