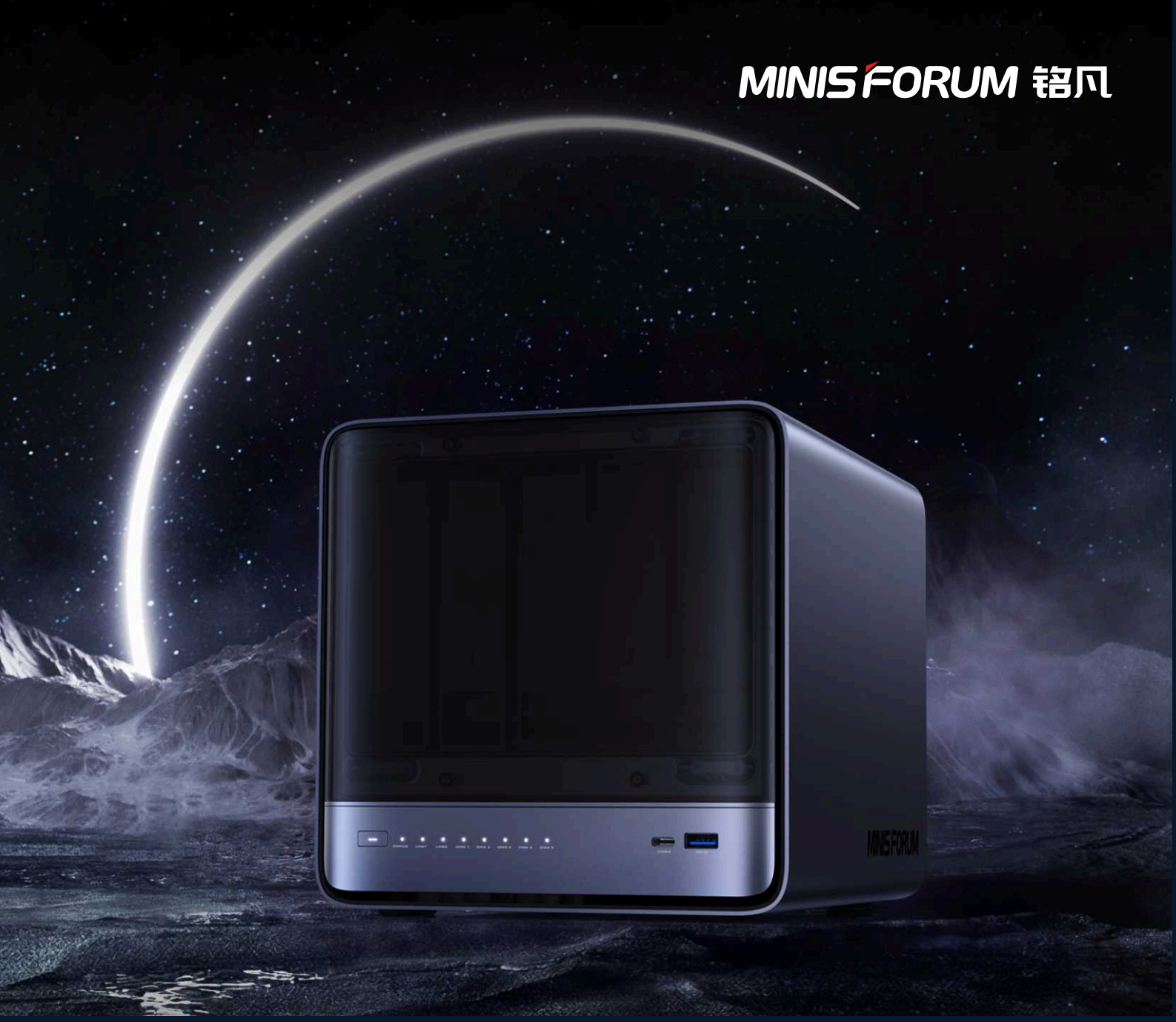


N5 MAX

本地AI知识库
解决方案白皮书

2026 年 8 月

01 /

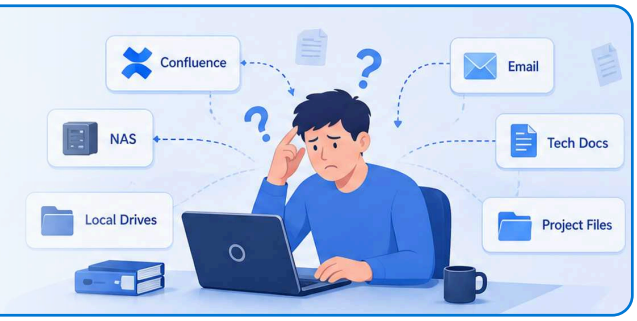
知识管理的三个痛点

1.1 数据上云，主权让渡——云端便利背后的隐私困境

企业将产品文档、技术白皮书、合同协议上传到公有云 AI 平台，每一次提问都在将核心数据暴露于第三方服务器。对于重视数据主权的团队而言，这不是效率工具，而是合规风险。

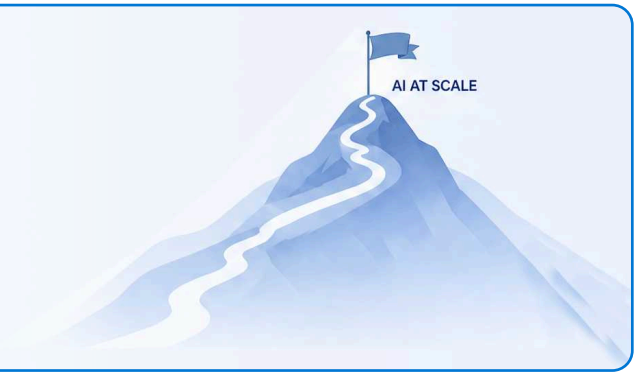
1.2 信息资产碎片化——多端离散数据的检索与沉淀困境

产品规格在 Confluence、项目文档在 NAS、客户合同在网盘、技术方案在本地硬盘——知识散落在五六个平台上，检索靠记忆、靠翻目录、靠问同事。你知道那份文档存在，但就是找不到。



1.3 本地大模型的工程化鸿沟——从拥有算力到真正用起来的距离

构建本地 RAG 知识库需要 GPU 采购、向量数据库搭建、Embedding模型选型、LLM 部署调优、前后端开发——对于非 AI 团队而言，这意味着一到两个工程师全职投入数月。即使勉强跑起来，后续的文档更新、索引重建、模型升级同样是持续负担。

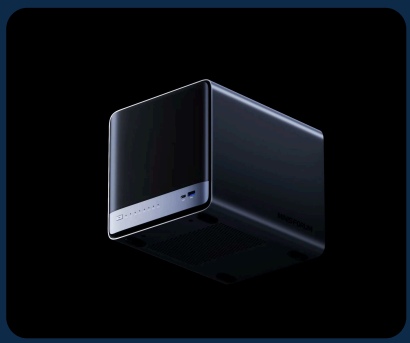


02 /

N5 MAX 本地知识库

开箱即用的私有 AI 本地知识库

N5 MAX 将本地知识库所需的全部能力——126 TOPS AI 算力、向量检索引擎、Embedding 模型、大模型运行时、文档解析管线、Web 管理界面——预装为一体化方案。通电开机，上传文档，即可开始提问。所有检索与推理在本地完成，文档不出设备，查询不留云端日志。



03 /

核心配置

配置类别	内容	客户价值
AI 引擎	AMD Ryzen™ AI Max+ 395（8 核 32 线程，最高 5.1 GHz），最高 126 TOPS	本地流畅运行 Qwen3.5/Qwen3.6 35B、Llama 3 等开源模型，RAG 问答响应 < 3 秒
统一内存	128GB 和 64GB LPDDR5x 8533 MT/s	大模型 + Embedding 模型 + 向量索引同时驻留内存，无需显存焦虑
存储架构	最高 200TB（5 × HDD + 5 × M.2 NVMe SSD）	百万级文档长期归档，SSD 承载热索引、HDD 存储原始文档与冷备份
文档解析	支持 MD / PDF / DOCX / XLSX / PPTX / CSV / TXT / HTML / JSON	多种格式一键导入，无需手动转格式或预处理
网络	双 10GbE 万兆 RJ45	团队多人查询不卡顿，大文件导入秒级完成

04 /

知识库核心能力

4.1 多格式文档一键导入

支持拖拽上传或指定 NAS 目录批量扫描导入。Markdown、PDF、Word、Excel、PPT、CSV 等格式自动解析——PDF 转文本、表格提取、样式清洗全部在后台自动完成，用户无需关心底层解析逻辑。

4.2 智能分块与混合检索

文档导入后自动进行语义分块（Smart Chunk），结合文档标题生成带上下文的 chunk 标签。检索阶段采用 BM25 关键词匹配 + 向量语义检索的混合策略（Hybrid Search），既保证专业术语精确命中，又覆盖同义词和语义相近的查询意图。

4.3 部门级多知识库隔离

支持按部门或项目创建独立知识库，每个知识库拥有独立的文档集和索引空间。查询时自动匹配对应知识库，也可跨库联合检索。权限控制确保不同团队的文档互不可见。

4.4 流式问答与引用溯源

用户以自然语言提问，系统通过 RAG 检索相关 chunk 并交由本地大模型生成回答。回答以 SSE 流式输出，像 ChatGPT 一样逐字呈现。每个回答附带引用来源（源文档名 + 原文片段），可一键跳转查看上下文。

4.5 周期性同步与增量更新

支持配置 NAS 目录周期性扫描，新增或修改的文档自动索引更新，无需手动触发全量重建。文档删除时同步清理对应 chunk 和向量索引，保持知识库内容与实际文件一致。

4.6 完全离线运行

知识库引擎的所有组件——文档解析、向量索引、混合检索、大模型推理——均在 N5 MAX 本地运行。断网环境下同样可以导入文档、构建索引、问答交互。联网仅用于可选的模型下载与系统更新。

05 /

场景与技术映射

场景类型	典型用例	技术支撑
产品文档问答	“N5 MAX 最多支持多少 TB 存储？” “BD895i 的 TDP 是多少？”	Ollama（Qwen 3.6 35B）+ Embedding 向量索引（768 维）+ BM25 混合检索 + RAG 管线
合同条款定位	“去年与 XX 公司签的合同，违约金条款在哪？”	PDF/DOCX 解析 + Smart Chunk 分块 + Hybrid Search + 引用溯源
技术方案检索 / 学习笔记复盘	“之前做过的那个双活架构方案，数据库选型部分怎么写的？” “上周整理的设计模式笔记，单例模式那部分再给我讲一下”	全文索引 + 语义检索 + 部门级知识库隔离 Markdown 解析 + Chunk 级引用 + 流式问答
法规合规查询	“最新的数据安全法里，跨境数据传输有什么要求？”	批量文档导入 + 周期性增量更新 + 引用溯源
多语言文档混合检索	中英文混合语料库中查找特定技术参数	多语言 Embedding 模型 + 跨语言语义检索

06 / 技术架构

知识库引擎架构

层级	组件	说明
Web UI	FastAPI + SSE Streaming	文档上传、知识库管理、问答交互的统一界面
检索层	Hybrid Search（BM25 + 向量）+ RRF 融合	关键词精确匹配 + 语义相似度检索，RRF 算法融合排序
索引层	向量索引 + 全文索引 + 元数据索引	768 维 Embedding 向量 + BM25 倒排索引 + 文档属性过滤
分块层	Smart Chunk + Front Matter 剥离	按语义边界智能分块，自动去除 AIGC 元数据，保留纯正文
解析层	PDF（pdftotext）/ DOCX / XLSX / PPTX / MD / CSV / HTML / JSON	多格式解析器链，自动识别文件类型并路由
模型层	llama.cpp（Core Runtime）+ Ollama（Optional Runtime）+ Embedding 模型 + LLM	本地模型管理、Embedding 向量化、大模型推理生成
存储层	SQLite（元数据/索引）+ ZFS（原始文档）	轻量数据库 + 企业级文件系统，兼顾查询性能与数据安全

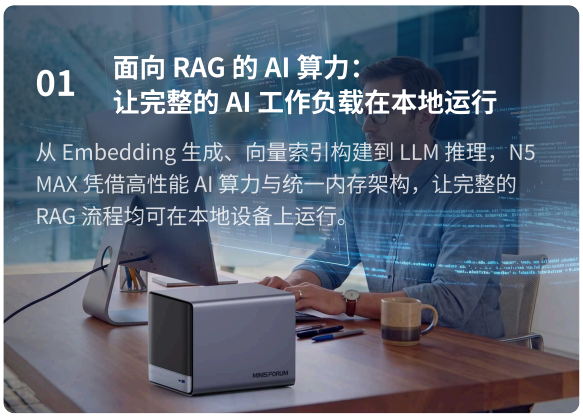
06 / 技术架构

查询流程

用户提问 → Hybrid Search（BM25 + 向量） → RRF 融合排序 → Top-K Chunk 提取 → Build RAG Context → 本地 LLM 生成 → SSE 流式返回 + 引用溯源

07 /

为什么选择 N5 MAX 做知识库



01 面向 RAG 的 AI 算力：让完整的 AI 工作负载在本地运行

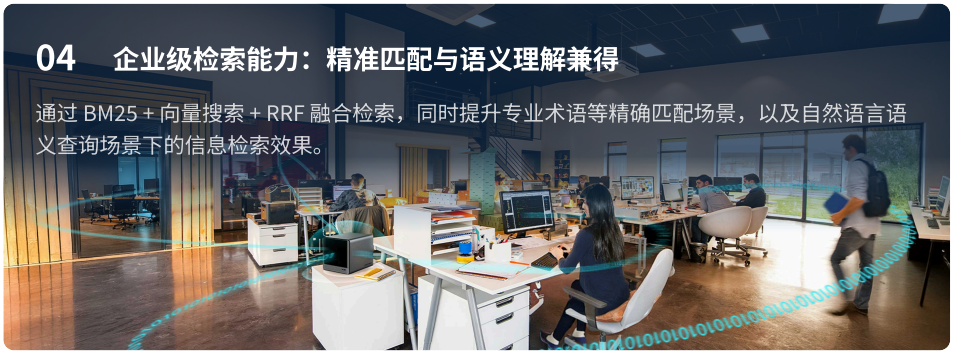
从 Embedding 生成、向量索引构建到 LLM 推理，N5 MAX 凭借高性能 AI 算力与统一内存架构，让完整的 RAG 流程均可在本地设备上运行。

02 一体化 RAG 技术栈：一台设备，集成完整 RAG 流程

从文档解析、智能分块、Embedding、索引构建，到信息检索与 LLM 推理，完整的 RAG 技术栈均集成于一台设备中。

03 大规模知识存储：为持续增长的知识库而打造

HDD + NVMe 存储架构将原始文档与活跃索引分离存储，在满足长期知识沉淀的同时，确保高频访问的 RAG 数据保持高速读写。



04 企业级检索能力：精准匹配与语义理解兼得

通过 BM25 + 向量搜索 + RRF 融合检索，同时提升专业术语等精确匹配场景，以及自然语言语义查询场景下的信息检索效果。

05 可追溯的 AI 答案：每一个答案，都有据可查

每个 AI 回答均基于检索到的源文档生成，并提供引用溯源能力，帮助用户快速定位答案所依据的原始内容。

08 / 对比

N5 MAX 本地知识库 vs 云端 AI 方案对比表

对比维度	N5 MAX 本地知识库	云端 AI 方案（如 ChatGPT + 第三方 RAG）
数据存储位置	本地设备，零云端依赖	第三方服务器
隐私合规风险	低（文档不出设备）	高（每次上传暴露数据）
部署方式	开箱即用，通电即用	需 API 配置、开发集成
长期成本	一次性硬件投入，文档量越大边际成本越低	按 Token/月订阅，文档量增长费用线性上升
离线可用	完全离线，断网可用	必须联网
响应速度	局域网内 < 3 秒	依赖公网带宽和 API 限速
文档格式支持	MD / PDF / DOCX / XLSX / PPTX / CSV / TXT / HTML / JSON	通常需预处理或仅支持纯文本

免责声明

01 知识库问答结果由本地大模型自动生成，可能存在不准确之处，不应作为法律、医疗、金融等专业决策的唯一依据。

02 文档解析准确率受原始文件质量影响（扫描件PDF、手写体、复杂排版等场景可能存在识别误差）。

03 检索效果与文档结构、命名规范、分块策略相关，建议按照产品文档模板规范编写，以获得最佳体验。

04 本文档内容不构成任何形式的商业承诺，MINISFORUM铭凡 保留对软件功能与产品规格进行迭代更新的权利。

版权声明

MINISFORUM

https://www.minisforum.cn/

bill.gai@miniscloud.com