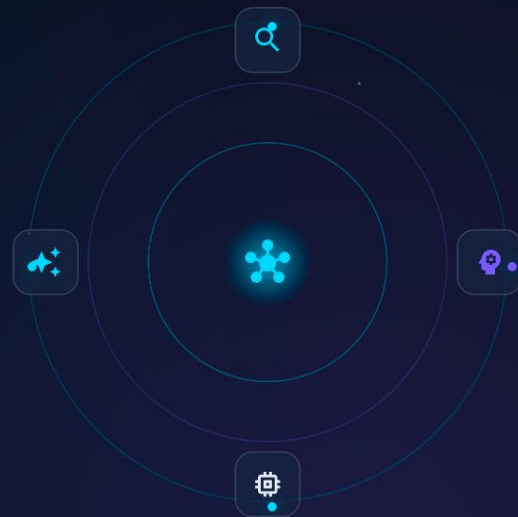


搜索驱动智能

为什么每个 AI 系统最终都需要搜索引擎？

AI 时代企业搜索平台建设实践与思考

" 2026 年，如果你还在讨论 AI 需不需要搜索，
你可能已经落后了。"



几乎所有先进 AI 系统都在做搜索

行业现状 / 2026

AI 越先进，越依赖搜索。搜索不是 AI 的附属功能，而是 AI 系统的核心能力。



ChatGPT Search

4 亿+

周活用户

搜索已深度集成 ChatGPT 主线，成为默认信息获取通道。



Google AI Overviews

60%

问题查询触发 AI 摘要

覆盖 15 亿+ 月活用户，AI 摘要改写搜索结果呈现。



Perplexity

2.5 亿+

月查询量

增长最快的 AI 原生搜索引擎，重新定义搜索体验。



Anthropic Claude

Deep Research

AGENT 自主迭代检索

Agent 自主规划多轮检索与推理，进入研究型搜索范式。



Microsoft Copilot

M365 + Windows

深度集成

搜索能力内嵌于 Microsoft 365 与 Windows 系统层。



市场体量

全球 AI 搜索日查询量已达 **17 亿次** (2026 Q1)，较 2025 年 **5.9 亿次** 接近 **3 倍** 增长。



市场格局

搜索市场没有单一赢家——用户在 ChatGPT、Perplexity、Google 之间多平台切换。



行业共识

Gartner 将 Agentic AI 列为 2025 年十大战略技术趋势首位，Agent 的核心能力之一就是搜索。



搜索不再只是搜索引擎公司的事——它是所有 AI 系统的底层能力。

大模型到底缺什么？

能力鸿沟 / 范式跃迁

大模型很强，但知识是**静态的、通用的、过时的**。企业需要**动态的、专属的、实时**的知识。

维度	训练知识（大模型自带）	企业知识（需要搜索获取）
时效性	🕒 截止到训练时间，无法更新	🔄 持续更新，实时变化
覆盖面	🌐 公开互联网数据	🏢 企业内部数据，模型看不到
准确性	⚠️ 通用知识尚可，专业领域幻觉率高	🔍 需要事实来源验证
可控性	🚫 不可控	✅ 可控，有据可查，可追溯

👤 纯 LLM 幻觉率 **12-18%**，RAG 检索验证可降低 **40%+**

💰 RAG 部署成本仅为微调的 **1/10**

🚀 2026 年新认知 · RAG 四大新范式

传统 RAG（向量库 + 检索 + 生成）正在被超越

🏠 **Graph-RAG**
从向量相似度走向知识关系推理

🔄 **Agentic RAG**
检索从单次变成『思考→检索→再思考→再检索』的循环

🧠 **长期记忆系统**
AI 开始拥有持续记忆，而非每次重新检索

∞ **无检索推理**
部分场景 RAG 被更高层架构吸收

RAG 不会消失，但正在从『独立架构』变成『AI 的基础能力层』。

🔗 不是模型不够强，而是模型不知道你的企业发生了什么。 搜索，是填补这道鸿沟的唯一路径。

中国企业的数字现实

数据现状 / 中国场景

企业的数据从来不在一个地方。但中国企业的数字分布，和硅谷报告画的不一样——不是 291 个 SaaS，而是 **本地系统 + 内部工具 + 部分云协作** 的混合体。



本地部署系统

ON-PREMISE SYSTEMS

ERP / OA

CRM / SCM

数据库 / 数仓

日志系统

占比最大 · 数据主权

~60%



内部协作平台

INTERNAL COLLABORATION

内部 Wiki

文档管理系统

知识库

流程引擎

知识沉淀 · 高价值

~25%



部分云服务 / SaaS

CLOUD / SAAS

飞书 / 钉钉 / 企微

代码仓库

对象存储

部分行业 SaaS

增长最快 · 混合部署

~15%



数据孤岛

01

系统间割裂，格式不统一，权限各自为政



找人比找数据难

02

员工每天在多个系统间切换：找文档、找代码、找流程、找知识



AI 想用，用不上

03

大模型无法直接访问防火墙后的数据



实时性差

04

关键业务数据持续变化，训练数据早已过时

中国企业的数字不在『291 个 SaaS』里，而是在 ERP、OA、Wiki、数据库、代码仓库和部分云协作工具里。数据孤岛的严重程度，不亚于海外——甚至更甚。

AI 工程范式的三次跃迁

范式演进 / HARNESS ENGINEERING

2026 年, AI 工程的重心已从『怎么写提示 词』转移到『怎么构建可靠的 Agent 运行框架』。搜索, 是这个框架的核心 组件。

阶段 01 2022-2024

Prompt Engineering

提示工程

优化单次模型调用的输入

阶段 02 2025

Context Engineering

上下文工程

每一步该给模型什么上下文 (搜索/检索的核心起点)

阶段 03 2026

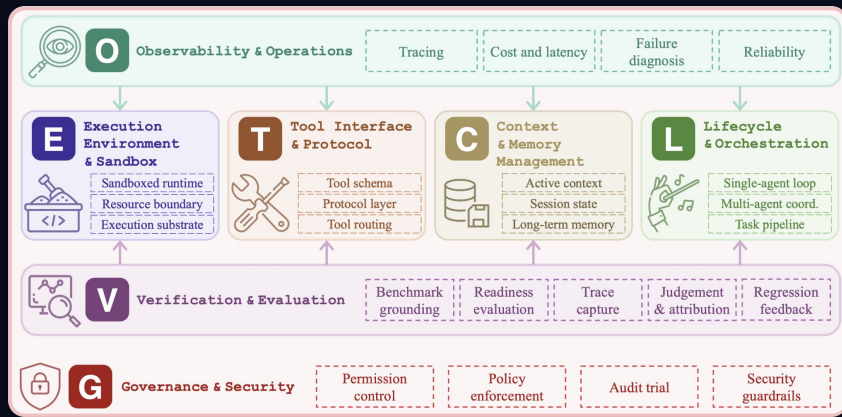
Harness Engineering

驾驭工程

模型外层的完整运行框架 (状态 + 工具 + 反馈 + 约束)

来源: CMU / Yale / Amazon, 2026

层	名称	与搜索的关系
E	执行环境与沙箱	—
T	工具接口与协议	搜索是 Agent 的核心工具
C	上下文管理	搜索是上下文管理的数据源
L	生命周期与编排	多轮检索的编排
O	可观测性	检索质量监控
V	验证	检索结果验证
G	治理	数据权限控制



<https://picrew.github.io/LLM-Harness/>

🔑 搜索是 Context Management 层的核心数据源。没有搜索, Agent 就没有上下文; 没有上下文, Agent 就没有可靠性。

A Timeline of Harness Systems

Key milestones across Execution, Tools, Context, Orchestration, Observability, Evaluation, and Governance



从『龙虾』到 Agent: AI 从『聊天』走向『执行』

AGENT 时代 / OPENCLAW

2026 年最大的变化, 不是模型更强了, 而是 AI 开始『做事』了。而做事的前提, 是能搜到正确的信息。



起源

2025 年底, **Peter Steinberger** 发起
OpenClaw 开源本地自主 AI Agent 框架



速度

GitHub Star **72 小时** 突破 **5 万**, 刷新
AI 开源项目最快增长纪录



别名

因图标形似龙虾, 被中文社区称为『**龙虾**』



扩散

2026 年春节引爆全球, 带动大量国产『**龙虾**』工具涌现

范式对比 · 从问答到执行

Traditional AI

2023-2024

问答型 AI · **不需要搜索**

用户提问



模型回答



结束

Agent AI · 2026

2026

执行型 AI · **每一步都依赖搜索**

用户下达任务



Agent 理解意图



搜索/检索知识



推理/制定计划



执行/调用工具



验证/检查结果



返回/完成任务

Memory · Retrieval ·
Knowledge · Context 正在
融合

4 大关键词



Memory

Agent 跨会话记忆 → 搜索提供可检索的长期存储



Retrieval

从单一检索到多路召回 + Rerank + 迭代优化



Knowledge

知识图谱 + 向量索引的边界正在模糊



Context

Context Window 有限, 搜索是大模型的『外部记忆』

搜索是 AI 的知识操作系统

核心枢纽 · 架构视角

在真实 AI 系统中, 搜索不是可选项——它是连接用户意图与企业知识的核心枢纽。

AI 系统真实架构

数据流: 从用户意图 → 知识检索 → 生成答案



知识连接器

打通所有企业数据源, 统一接入



统一索引层

一份索引, 多处消费, 消除数据孤岛



检索引擎

针对不同场景智能选择最优检索策略



Agent 记忆

为 AI Agent 提供可检索的长期记忆

搜索基础设施：兵家必争之地

赛道全景 · 2026

搜索基础设施正在成为 AI 时代最热的赛道之一。传统数据库系、Lucene 系、向量系三条路 线在 converge。

传统数据库系

PostgreSQL 阵营

ParadeDB (Tantivy) Neon Lakebase pgvector
Supabase Oceanbase ...

Lucene 系

倒排索引经典阵营

Elasticsearch OpenSearch Easysearch Solr

向量数据库系

AI 原生阵营

Milvus Qdrant Weaviate Pinecone / Chroma
...

融合趋势：DB + Search + AI 正在收敛

- PostgreSQL 在『吃掉』搜索 (ParadeDB: PG 里做 ES 级搜索)
- 搜索引擎在『吸收』向量能力 (ES 8.x / OpenSearch)
- 向量数据库在『长出』全文检索 (Milvus 2.5+ 混合检索)
- Agent 框架在『内嵌』搜索 (OpenClaw / LangChain)



AI 的数据入口

所有 AI 应用都需要搜索获取外部知识



市场规模

RAG 市场 **\$19.6亿** → **\$98.6亿**
(2025→2030), CAGR **38.4%**



技术收敛点

DB、搜索、向量三者的边界正在模糊



生态锁定

搜索基础设施迁移成本极高，强生态粘性

搜索引擎的演进路线

演进路径 · 三阶段

搜索正在从信息检索工具, 演进为 AI 的**知识基础设施**。

Keyword Search

2010s · 关键词搜索

阶段 1

特征

- 倒排索引
- 词频匹配
- 用户必须知道『该搜什么』

代表 Elasticsearch / Solr / Lucene

Hybrid Search

2024-2025 · 混合搜索

阶段 2

特征

- 全文 + 向量 + 语义
- Rerank 精排
- Query Rewrite · AI 自动改写查询

代表 Elastic 8.x / OpenSearch / Milvus / Qdrant / Easysearch

Agentic Search

2026+ · 智能体搜索

阶段 3

特征

- AI 自主规划检索策略
- 多轮迭代
- Deep Research 范式
- 搜索即推理, 推理即搜索

代表 OpenAI Deep Research / Perplexity Sonar / Claude Deep Research / Google AI Mode



从『找信息』 → 『做研究』

Deep Research 让搜索从单次查询变成多轮探索



从『被动检索』 → 『主动规划』

Agent 自己决定搜什么、搜几次、什么时候停



从『返回链接』 → 『生成答案』

用户不再需要点击链接



从『关键词匹配』 → 『意图理解』

不再依赖用户用词精确

AI 时代对搜索的新要求

能力模型 · 四维评估

传统搜索追求『找得到』，AI 搜索需要『找得全、找得准、找得懂、找得快』。

传统搜索
找得到



AI 搜索 四维能力模型

1 找得全

2 找得准

3 找得懂

4 找得快

全文 + 向量 + 混合 + Rerank / 多路召回 → 融合排序 → 精排 → 返回 / 意图理解 → 查询改写 → 迭代优化

维度	传统搜索	AI 搜索要求	实现策略
 找得全	关键词匹配	覆盖所有内容，不遗漏	全文检索，毫秒级
 找得准	TF-IDF，相关性有限	语义匹配，理解意图	向量检索，突破关键词
 找得懂	用户必须用对词	理解用户想问什么	混合检索 + Rerank
 找得快	秒级可接受	毫秒级 P99	分布式 + 近实时刷新

QUERY REWRITE 示例

用户问『去年 Q3 华东区销售额怎么样』 → AI 改写为多个子查询，意图识别 + 查询扩展 + 多路召回。

极限科技：做中国的搜索基础设施

公司概况 · INFINI Labs

搜索基础设施不能一直靠国外。在中国，需要有公司把这件事接过来。



极限科技 · 让搜索更简单

成立

2021年12月，北京

定位

专注于大数据搜索与分析的基础软件公司

研发

北京总部+长沙研发中心

核心团队

来自Elasticsearch原厂及中文社区，多年ES源码开发经验

使命

『让搜索更简单』，打造极致易用的数据探索与分析体验



API

SDK

Mem0

MCP

Skills

Coco AI



全文搜索

向量搜索

语义搜索

多模态搜索

地理空间位置搜索

聚合引擎

规则引擎

模式引擎

异步搜索

数据加密与脱敏

异地容灾

国密集成

...



Phytium 飞腾



Loongson



极限科技：做中国的搜索基础设施

公司概况 · INFINI Labs

搜索基础设施不能一直靠国外。在中国，需要有公司把这件事接过来。

500 万 +

Easysearch 累计下载部署量

85%

中国 ES 用户已部署我们的开源工具到生产环境

权威认可

国家高新技术企业 (科技部认定)

ISO三大管理体系认证 (质量/信息安全/IT 服务)

信通院基础能力专项测评 (搜索型数据库检验通过)

泰尔实验室 检验测试

2023星河案例 · 数据库标杆

2025金猿奖 · 年度国产化优秀代表厂商



国家高新技术企业



中关村高新技术企业



国家双软技术企业



北京大数据协会大数据企业



信通院数据库应用创新实验室
搜索行业工作组副组长单位
中国信通院数据库应用创新实验室
成员单位



北京软件和信息服务协会
会员单位



2024年度IT168技术卓越奖
极限科技 Easysearch



2023年度星河案例奖
(极限科技 + 中国移动)

Easysearch: 企业级分布式搜索引擎中间件

产品介绍 / 性能数据

ES 国产化的核心不是『换个牌子』, 而是在 **生态兼容** 的基础上做到 **性能超越** 和 **自主可控**。

企业级分布式搜索型数据库, 兼容 Elasticsearch 生态, 面向 AI 场景深度优化。



墨天轮搜索型数据库排行榜长期第一名

INFINI Easysearch 是一个企业级分布式搜索引擎中间件, 实现非结构化数据检索、全文检索、向量检索、地理位置信息查询、组合索引查询、多语种支持、聚合分析, 提供企业级的安全和加密, 并提供完善的周边管理工具。

兼容 ES 及其生态

完善的企业级功能

内核及功能增强

原生中文

国产自主可控

开箱即用

能力架构

全文搜索 · 倒排索引

向量检索 · 语义匹配

混合检索 · RRF + Rerank

统一搜索与知识访问层

实践案例 · 中国一汽

背景

50+ 套 ES 集群, 1.5PB+ 数据, 支撑营销、管理、财务、研发、自动驾驶等核心系统

痛点

部分集群写入掉节点, 当天数据写不完, ES 7.3.2 免费版无商业支持

方案

Easysearch 替代 ES, ZSTD 压缩 + 快照备份 + 冷热分离, 近两天热数据在本地, 历史数据在 S3 按需搜索

效果

磁盘从 130TB 降到 60TB (节省 50%), 节点从 14 降到 8 (节省 40%)

50%

磁盘节省
130TB → 60TB

40%

节点节省
14 → 8

搜索基础设施选型与实践

方法论 / 落地实践

选对引擎只是第一步，**怎么落地**才是真功夫。

搜索选型的现实约束

- 信创合规：** 关键行业搜索基础设施纳入自主可控范围
- ES License 动荡：** 2021 SSPL/ELv2 → 2024 追加 AGPL，开源合规复杂度持续上升
- 数据安全：** 核心数据不出域，私有化部署是刚需
- 成本压力：** 商业授权 + 硬件成本，大规模集群 TCO 高
- 迁移风险：** 业务不可中断，API/DSL 兼容是硬门槛

选型方法论 · 五维评估



兼容性
API/DSL



生态
运维工具 / 社区



性能
吞吐 / 延迟



自主可控
私有化 / 信创



合规性
License / 安全

原菱形结构：兼容性在上 · 生态与性能居中 · 自主可控在下 · 合规性贯穿

实际效果



企业问答

找文档平均 15 分钟 自然语言秒级获取答案



统一搜索

翻 10 个系统 一个入口搜到全部

实践中遇到的真问题

问题	核心矛盾	解决方向
兼容性	换引擎 ≠ 重写所有查询	API/DSL 兼容是迁移硬前提；网关代理，业务无感知切换
性能	高压缩和快查询传统不可兼得	ZSTD 等新压缩算法正在打破这个 trade-off
中文场景	插件式方案维护成本高	引擎层原生支持，持续工程投入
运维门槛	不是每个团队都有专职 DBA	降低复杂度，内置 Web 管控 + 自动化运维

应用到 AI 场景的典型架构

- ↓ 员工自然语言提问 / Agent 自主检索
- ↓ 统一搜索平台（全文检索 + 向量检索 + 混合检索 + Rerank）
- ↓ 企业知识索引层
- ↓ GitLab · Wiki · 飞书 · ERP · 数据库 · 对象存储 ...



代码助手

全局搜索无结果 跨代码库语义检索精准定位



知识助手

新员工入职培训 2 周 自助问答即时上手

企业搜索平台建设方法论

四阶段模型 / 关键经验

搜索平台建设是分阶段演进的体系，关键是在每个阶段**做对的事情**。



▲ 各阶段常见误区

阶段	常见误区
数据统一接入	贪多求全，试图一次接入所有系统
多策略检索	只做向量检索，忽视全文检索的基础作用
AI 增强	过度依赖模型，忽视检索质量的基础工程
Agentic 搜索	跳过前三阶段，基础不牢

💡 四条关键经验

- 1 搜索质量 > 模型能力**
RAG 效果 **70%** 取决于检索质量，**30%** 取决于模型
- 2 先做基础再做 AI**
没有好的索引和检索，再强的模型也是『巧妇难为无米之炊』
- 3 评估驱动迭代**
建立 Recall、MRR、NDCG 评估体系，数据驱动优化
- 4 从『检索准确率』到『系统可靠性』**
2026 年评价维度已变：任务完成率、决策正确率、长期一致性更重要

趋势前瞻：2026 下半年值得关注的方向

趋势展望 / H2 2026

搜索正在从『AI 的一个功能』变成『AI 的基础设施』。

01



Graph-RAG 走向工程化

从论文到落地。**知识图谱 + 检索**的融合不再停留在原型，企业级产品开始支持多跳推理和关系链验证。

▲ 多跳推理 · 关系链验证

02



Agentic RAG 成为标配

单次检索已经不够。Agent 需要『**思考→检索→再思考→再检索**』的多轮循环，检索深度决定推理质量。

∞ 多轮迭代 · 推理深度

03



AI 的『长期记忆』依赖检索层

OpenAI Memory、Claude Memory 等长期记忆系统，**底层都是检索**。记忆不再是 KV 存储，而是可搜索的语义空间。

☁ 可搜索语义空间

04



Harness Engineering 重塑 AI 工程范式

2026 年之后，AI 工程的重心不是调 Prompt，不是 Fine-tune，而是构建**可靠的 Agent 运行框架**。搜索是其中 Context Management 层的核心。

💎 Context Management 核心

05



搜索 / DB / AI 三层融合加速

PostgreSQL 在『吃』搜索 (ParadeDB)，搜索引擎在『吸』向量 (ES 8.x)，向量数据库在『补』全文 (Milvus)。终局是统一为『**AI 知识基础设施**』。

✳ 三路线收敛 · 统一基础设施

06



中国搜索基础设施走向自主

ES 国产化从政策驱动转向**技术驱动**。信创合规是入场券，性能超越和生态兼容是核心竞争力。

🌟 信创入场 · 性能超越

AI 决定答案如何生成，搜索决定答案从哪里来。

01

搜索是 AI 的知识操作系统

从 Agent 到 LLM 的每一次推理，都依赖 Search 提供的高质量上下文。

02

搜索基础设施是兵家必争之地

PG 系、Lucene 系、向量系三条路线在 converge，ES 国产化是真实趋势。

03

搜索正在从信息检索演变为 AI 基础设施

关键词搜索 → 混合搜索 → Agentic Search，搜索即推理，推理即搜索。

04

没有搜索的 AI，只是一个会说话的大模型

搜索能力的上限，决定了 AI 系统的智力天花板。

谢 谢

ITPUB 直播分享 · 2026

极限科技 INFINI Labs · 让搜索更简单