

企业AI建设规划 和落地指南

- ▶ 快速落地 **DeepSeek**
- ▶ 战略规划到场景落地，**8大**详细路径解析



DeepSeek 横空出世，以低成本、高性能、强开源的三重属性颠覆大模型产业链的传统逻辑，为我们提出了发展 AI 技术的另一个解法——并非一味地追求算力膨胀。各行各业都在讨论和积极尝试 DeepSeek，用户对于智能化的认知与实践正经历着新一轮的深刻变化，也影响着用户的 AI 规划和落地路径。



全面拥抱DeepSeek 建设AI的共识空前一致

各行业用户进行 AI 建设之前，往往要进行大量战略对齐工作，而在 DeepSeek R1 火爆之后，用户建设 AI 的共识达到空前一致，原本因技术不确定性而存在的观望态度迅速消解，取而代之的是对 AI 建设的战略聚焦。

资源的重新配置体现为研发预算的倾斜、跨部门协作机制的建立，以及对长期技术投入的容忍度提升。这种自上而下的共识推动 AI 项目从边缘试点向核心业务渗透，形成了一种“技术驱动战略”的新型决策逻辑。



DeepSeek重塑用户AI建设路线图

一个典型的 AI 建设整体路线图，包括 8 个关键步骤，才能实现 AI 项目的全面落地与持续运营。



01 把DeepSeek提升为战略级目标

目标应与用户的战略紧密相连，确保 AI 的建设为实现整体战略目标服务。用户在建设 AI 时，经营目标、运营目标和部署目标相互依存，形成一个完整的目标体系。用户在部署大模型时需要清晰地识别其适用场景与局限，不能一刀切，而应制定合理和差异化的目标。因此，了解何时适合采用 AI、何时避免使用 AI，以及针对特定场景选择何种 AI 技术至关重要。

DeepSeek 大模型的突破性进展，抬高了大模型能力天花板，AI 适用场景进一步扩大。它能够模拟人类专家级的逻辑推演过程，展现出极高的逻辑严密性，其核心在于将传统 AI 的“规则响应”升级为“认知推演”，这使得 AI 能够处理非结构化信息交织的模糊性场景。

当这种接近人类高阶思维的能力注入业务场景，能力的质变直接打破了“AI 仅适用于标准化场景”的固有认知，推动应用边界向战略规划、创新设计等高阶领域延伸。

02 评估现状，了解AI的落地成熟度

在明确了 DeepSeek 的战略目标之后，我们需要了解用户自身所处的阶段，发现薄弱环节，并为下一步的建设方向提供指导。我们主要从 4 个方面来评估现状。



战略与场景

要确保用户内部已达成目标的对齐，同时需要优先选择符合战略目标，并且具备可行性的场景进行实施。



数据就绪度

AI 项目的成功，一定是数据驱动的。无论是传统 AI 还是生成式 AI，高质量的数据是 AI 模型项目的基础。但对于生成式 AI 来说，更重要的一点是数据的规模和多样性。比如 DeepSeek R1 模型能力强，一方面原因是算法创新，如 MLA、强化学习等，更重要是数据质量高、场景覆盖广。另外，想要长久保持 AI 模型的高质量服务，还需要不断采集数据、反馈数据，对模型进行迭代更新。



AI 项目的落地实施，离不开团队

需要具备理解业务并转化为项目的核心架构，以及一些工程化的角色，如数据标注、清洗、模型训练和调参、工程优化等角色。DeepSeek 的优异性能，让越来越多中小企业可以直接应用，进一步降低了工程化团队的要求，也会使得 AI 项目落地更加容易。



资源就绪度

指的是 AI 算力，包括计算、存储、网络等，以及为了实施项目所需要的一些基础框架、开发工具等，目的是提供对 AI 全生命周期的支撑。随着 DeepSeek 的规模化应用，AI 行业重心从训练转向推理，未来推理算力的消耗和供给将会极大增加，推理算力占比将从 30% 提升到 70%，用户在算力选择方面也会更加便利。



03 选择DeepSeek应用场景

如何选择 AI 应用场景，归根结底都是围绕**降本增效、提升体验和模式创新**三个方面来展开的，这些也是各行业智能化转型的终极目标。

首先关注**战略匹配度**，看是否符合这三个方面的目标，其次是**技术支撑**，重点考虑是否是生成式 AI 的应用场景，以及通过哪些技术路径来支持场景落地。

另外就是**数据和 AI 基础设施**：数据层面，需要具备精标注数据、反馈数据，以及垂类语料（无标注数据）条件；算力方面提供 AI 计算支撑；基础大模型方面，这是应用 AI 的最基础条件，用户需要考虑大模型的开源 / 闭源路线。目前开源路线已得到极大发展，尤其是 DeepSeek 生态已快速形成并在规模化扩张，芯片厂商、云厂商、软件厂商都在主动接入或适配，未来将会持续扩张。拥抱 DeepSeek 是一种更高性价比的选择。

04 初步评估可行性

选定场景后，我们需要对某个特定场景进行初步可行性判断，这个判断需要**业务、技术和工程角色**共同参与评估。



业务如何判断呢？DeepSeek 的优异性能和获得的便利性，让我们可以随时进行场景验证。例如，业务方是合同审核部门，场景是审核合同里的关键内容是否缺失，那么**可以基于 DeepSeek 进行初步验证**，输入一段合同文本，输出就是几个关键字段和它的值，而目标就是提取这些信息，并且内容没有丢失。这是没有任何技术和工程参与优化的结果。这个工作业务团队是完全可以胜任的，如果让技术部门来判断，可能会存在业务知识缺少的情况，无法准确判断业务场景的可行性。



技术部门则重点关注基础模型本身的能力，也就是大模型在预训练阶段基本定型的，主要**包括它在预训练阶段用的知识、采用了哪些语言、有没有用一些带有逻辑推理的语料**。在这些方面，DeepSeek 模型也表现出极为出色的能力，尤其是 671B 满血版，经历了 SFT 冷启动、COT 数据、GRPO 强化学习算法及全场景强化学习等多个创新点，模型已具备高效及强大的思考能力。



最后是**工程角度**，在 DeepSeek 出现之前，使用大模型需要掌握一些 Prompt（提示词）优化技巧，有时还需要结合 few shot（示例），以及思维链等方式，来激发大模型本身的能力。而 DeepSeek 本身已经表现出了以接近人类专家的水平思考过程，这就极大降低了 Prompt 设计工作的复杂程度。

各行业用户内部，知识问答的需求将会爆发式增长，因为接入 DeepSeek 来实现企业 RAG（检索增强）生成这种方式，整体的工程化落地难度也大大降低，越来越多用户会开始建设知识问答应用，以服务内部或外部客户。

同样的，DeepSeek 本身对语言理解能力以及对指令的遵循能力变得更好，所以进行大模型 SFT（Supervised Fine-Tuning）微调甚至进行二次预训练，这种场景将会大幅减少，对其工程化团队的要求将大幅降低。

05

准备DeepSeek落地所需资源

在确定初步可行之后，即可开始准备大模型落地所需资源。主要从 4 个方面准备：与建设场景相匹配的数据、算力、开发工具和团队。

 准备数据

根据选择场景的不同，需要准备的数据也会有很大不同：

RAG 场景



文档预处理(如文档格式转换等)、文档解析(PDF 识别, 版式识别)、OCR(图像转文字)、文档切分(切分为小的片段)、数据增强(提取 QA、生成摘要)。

大模型微调场景



对于少数具备较强 AI 技术能力的团队，在微调场景需要准备：精标注 QA 数据（问答对形式）、准备指令数据（如人设遵循、输出格式要求等）。需要确保数据准确性（例如做过滤低质量，去掉重复数据等）、数据多样性（要能够覆盖各种真实应用场景）、数据一致性（数据不冲突）、数据增强（生成 QA、摘要）等。

二次预训练场景



这种需求只有在某个垂类领域的头部企业才会选择，包括：专业领域的语料（如金融、医疗等专业术语，这些领域使用语境等）。需要确保数据多样性（如覆盖场景种类足够多）、数据规模足够大（一般要达到数十亿级 token 以上）、一定的配比数据(增加通用领域数据做二次预训练, 如 1:1 配置)、数据处理能力(去除重复数据、低质量数据、隐私数据、不合规数据等)。

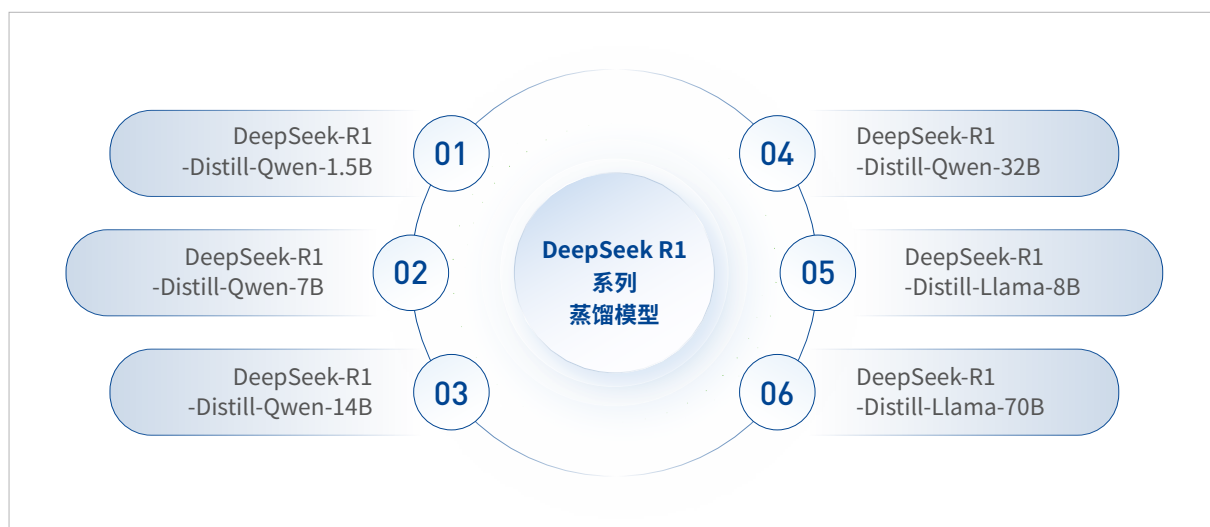
准备算力

在大模型建设中，算力的多少决定了训练速度、效率和模型表现，是推动模型快速落地的关键资源。那么要准备多大的算力才能满足实际应用需求？这里有一些经验，可以用于快速进行算力数量估算（此处仅考虑 GPU 算力）。

对于训练场景，全参微调所需的总显存一般为模型参数量（以 B- 十亿为单位）的 20 倍，例如 70 亿（7B）参数大模型，至少需要 140G 显存，按照单卡 80G 显存，则对应的 GPU 卡至少需要 2 张。而 130 亿（13B）参数大模型，至少需要 260G 显存，按照单卡 80G 显存，则对应的 GPU 卡至少需要 4 张。

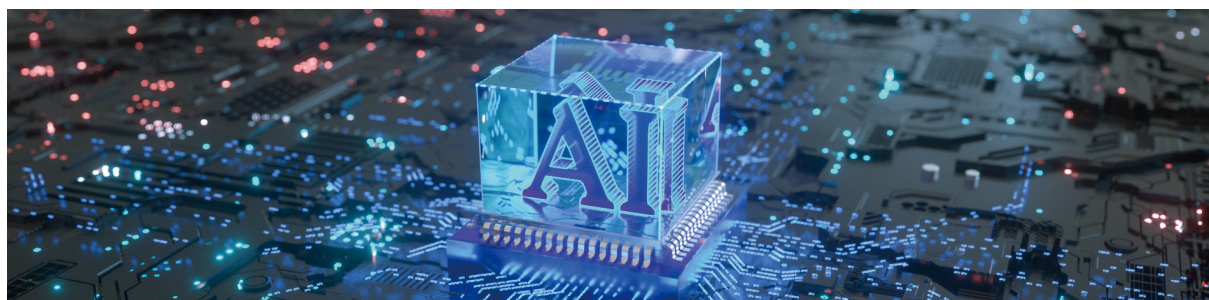
准备工具

基础大模型的选择，往往决定了应用效果的基线。对于大多数用户来说，首选是 DeepSeek R1 系列蒸馏模型，在同等参数量下，相比其它开源大模型，它能够取得更好的效果。而在 6 个蒸馏模型中，32B 参数量大模型能够在效果和落地成本方面取得较好的平衡，可以作为应用首选。



如果是为了解决特定场景下的应用目标问题，则需要 AI 应用开发工具，例如开发一个 RAG 应用、Agent 智能体应用等，需要评估应用开发工具对场景效果的支撑。

在模型微调和部署工具选择上，基于长远考虑，首先关注对资源的利用率和整体性价比，其次需要考虑工具的完整性、兼容性、易用性、可靠性、以及安全性。例如，在性价比方面，在相同的基础大模型条件下，平台对于算力资源消耗和模型服务性能如何；使用时学习门槛是否足够低，操作是否足够简便，过程是否自动化等。





搭建团队

大模型项目更依赖于数据驱动的效果调优、模型微调 and 二次预训练等技术，而传统系统更关注业务流程的实现与系统的长期稳定性。用户在推动大模型项目落地时，应根据首先落地的场景，来做差异化的团队配置。

例如，初期只选择内部流程自动化场景，只需要做 Prompt 优化和系统对接就可以，所以在 AI 团队上重点配置提示词优化工程师、工程开发工程师；而在 RAG 应用场景，就需要文档数据预处理、检索优化等工程师。

06 分阶段实施

阶段一：场景为王 & 运营提效

场景为王，指的是选择最简单、最容易做出效果的场景去建设。当前，基于 DeepSeek 模型的 RAG 应用正成为各行业智能化转型的关键突破口。DeepSeek 推动大模型技术逐步走向产业化规模化落地，用户亟需找到投入产出比高、见效快的 AI 应用场景。RAG 技术凭借其“知识检索 + 智能生成”的双重能力，正成为用户构建智能系统的首选方案。

对于希望快速实现 AI 价值的用户，建议优先选择知识密集度高、流程标准化强的场景切入，例如技术文档问答或产品知识库建设。通过小步快跑的方式，用户可在 3 个月内看到明显的效率提升，为后续更复杂的 AI 应用奠定基础。在实施过程中，要特别注意建立知识更新闭环，持续优化检索策略和生成质量，最终形成具有特色的智能知识中枢。

相较于其它通用大模型，DeepSeek 模型在构建企业 RAG（检索增强生成）应用中的优势不仅体现在成本降低和技术性能提升，更能通过开源生态为用户提供从基础应用到深度智能的全链路支持。

例如推理成本降低，**DeepSeek 在复杂推理任务（如金融研报生成）中实现了 70% 的成本降低**。其蒸馏模型（如 DeepSeek-R1）在更小规模下仍能保持高性能，减少了对高端计算资源的依赖；

长思维链与自我验证，通过强化学习和模型蒸馏技术，**DeepSeek 能在小参数模型中实现长推理链与自我验证**，满足用户对复杂场景的高准确率要求；

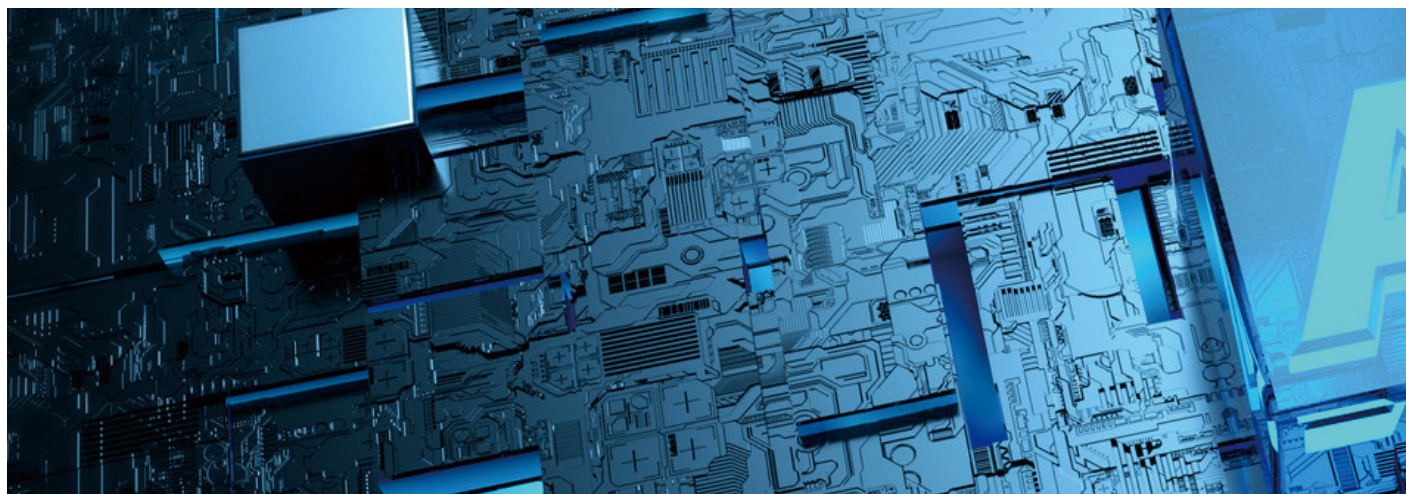
开源商用授权，**DeepSeek 模型提供开源商用授权政策**，用户可免费用于微调、量化及衍生开发，降低技术门槛和法律风险。

以知识库检索问答（RAG）方案为例，用户正将 AI 应用从单点工具升级为全域知识中枢。某处于装备制造产业链下游的传统型企业，需要对接上千家上游元器件供应商。企业有数百名产品设计人员，主要工作是结合工艺参数的需求，进行物料选型，再给后续环节使用。

物料总计有 3000 余个，对应 1 万多份物料技术规范书文档。员工通过手动整理参数表格并结合人工核对来完成物料筛选，完成一个场景的物料选型往往需要数天时间，严重影响生产效率。当 DeepSeek 出现后，用户可以通过对话的方式，让大模型准确回答其所需要的物料，使得整个过程缩短到分钟级，准确率达到 90% 以上。用户借助该能力中枢，正尝试对多个内部业务系统进行智能提效，如应用于研发、生产、供应链、销售、售后服务系统等，大幅提升效率和竞争力。另外需要注意的是，在企业级 RAG 应用上线后，需同步进行运营提效工作，持续收集场景反馈数据，对落地场景效果持续优化，才能在初始落地阶段产生明显价值。

阶段二：场景扩展

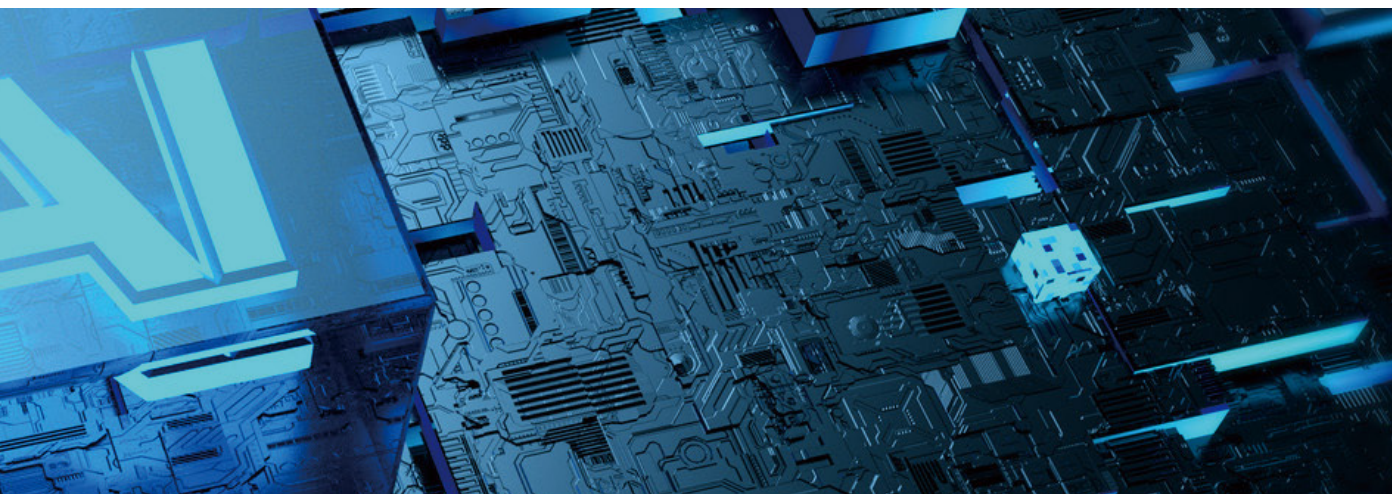
基于第一阶段的成功，用户要深入到核心业务场景去解决复杂的场景问题，发挥 AI 的价值，**从辅助性角色变成核心角色**。例如销售机器人、个性化服务机器人、运维机器人等。这个阶段的特点是复杂度更高，往往不是单靠大模型能完成的，需要借助小模型、智能体等来协作完成。



阶段三：持续优化

不管是什么样的 AI 模型，部署后，其固有知识相对不会变化，但业务场景不是一成不变的，如果不持续优化模型，它的效果会持续衰减。另外，随着落地的场景越来越多，并发量越来越大，它所消耗的资源也成倍增加，如果不对它的底层资源进行优化，后续将有非常大的成本支出。

综上，我们可以通过不断采集、反馈、优化，形成数据闭环，来持续对场景做提效。构建数据反馈闭环是确保企业大模型适应业务变化的基础，优化后的模型重新部署到业务场景中，继续收集新的数据，形成持续反馈的闭环。这不仅让模型适应业务的变化，还能不断提高其在核心场景中的应用效果。



阶段四：赋能核心业务

这是 AI 落地的终极目标，AI 大量融入到关键业务链条中，深入核心业务，比如制造业的研产供销服等环节。未来，用户的 AI 建设不断深入，将会覆盖到全部链条，例如——

- ◆ 在研发板块，可以充分利用大模型的理解以及调用外部工具、生成代码的能力，辅助进行产品工艺设计和参数选择；
- ◆ 在生产板块，可以通过语音自然语言交互，对机器人进行操作；
- ◆ 在销售板块，大模型结合数字人，能够为用户提供更好的体验和服务响应；
- ◆ 在设备运维方面，能够及时给出设备故障的维修方案；
- ◆ 在运营方面，通过大模型提供个性化的数据分析和运营报告。

.....

07 评估

对大模型项目建设评估不仅是对一次项目执行结果的评估，更是对阶段性工作的评估，包括但不限于项目业务价值、用户体验、成本效益、模型性能与合规性、安全性等多个维度的综合评估。通过建立持续的监控与反馈机制，优化运营成本，分析模型的扩展潜力，并结合未来发展战略，用户可进一步确保大模型项目的成功和长期价值。



08 持续运营

只有坚持长期在数据和场景方面的运营深耕，才有助于构建持续竞争优势，这个过程可重点关注数据飞轮持续化和场景运营精细化。

数据飞轮持续化

- ◆ 整合多源数据（销售、客服、供应链……），统一数据质量标准，设置专门的数据治理团队；
- ◆ 设置模型效果评估标准（基于标准测试集和基于业务反馈结果）；
- ◆ 收集线上 / 线下反馈数据；关键指标监控（性能指标、业务指标，如转化率、复购率）；
- ◆ 明确实施微调优化的准入标准；
- ◆ 通过持续的数据积累、数据反馈闭环，推动模型的不断优化和迭代。

场景运营精细化

- ◆ 一把手牵头，发动各个业务部门梳理业务场景，细化到最小粒度场景；
- ◆ 对全量场景做价值度评估，排优先级；
- ◆ 建立业务部门和 AI 团队的场景对接机制，明确场景提出 - 评估 - 立项 / 关闭 - 建设 - 评估 - 运营的整个闭环；
- ◆ 建立效果评估机制，对用户反馈做闭环跟踪；
- ◆ 挖掘新场景，鼓励内部做业务场景与 AI 结合的创新。

在这场由 DeepSeek 引发的智能化浪潮中，各行业用户的数字化转型路径都在发生变化，深信服将在这场变革中，助力用户更顺利地实现 AI+ 云化升级。

