高管指南：如何将生成式 AI 融入运营

从实验到实施：如何将生成式 AI 应用于实际场景

elastic | The Search AI Company

目录

开启您的生成式 AI 之旅

生成式 AI 是 2023 年兴起的极具颠覆性的技术。据预测，在未来几年里，生成式 AI 将深刻影响各行各业的方方面面，但有多少人能说他们已经破解了生成式 AI 的密码，现在就能让它为己所用？

当有些公司还在生成式 AI 浪潮中探索方向时，有些公司已经率先开始体验到这项技术带来的成果了。例如，在 Cisco，支持工程师可以迅速从类似的支持案例、内部论坛以及与客户问题相关的知识文章中找到归纳整理的相关答案。Cisco 已经充分享受到了生成式 AI 带来的益处，其重新设计的搜索解决方案成功解决了 90% 的支持请求，并每月为支持工程师节省了 5,000 小时的工作时间[1]。

在电子商务领域，您或许已经目睹过生成式 AI 的实际应用。生成式 AI 可以分析客户的过往购买记录、浏览历史和偏好，以在用户与聊天机器人互动时为他们生成个性化的产品推荐。在后端系统，使用生成式 AI 有望增强客户的参与度和留存率，提升欺诈检测能力，以及其他更多方面的改进。

为了揭开生成式 AI 各种功能的神秘面纱，并确定采用何种方式将它应用到您的业务中，您需要有一份关于如何激活数据的分步指南。在这本电子书中，我们将引导您**完成从初步设想到成为 AI 专家的转变之旅。**您可以将这份指南看作一个路线图，帮助您利用生成式 AI 实现业务成果的突破性提升。

99%

的组织认为，生成式 AI 有潜力推动组织内部或外部的变革[2]

但只有

32%

的领导者认为，自己有能力在组织中实施 AI[3]

[1] 数据来源：Elastic — Cisco 在 Google Cloud 2024 上利用 Elastic 打造 AI 驱动型搜索体验
[2] 数据来源：Elastic — Elastic 生成式 AI 报告 (2024)
[3] 数据来源：Russell Reynolds — 拥抱未知：面对不确定性，领导者如何运用生成式 AI (2024)。

下面是本指南将为您介绍的内容：

第 0 步

了解生成式 AI 能提供哪些帮助。您想用生成式 AI 实现什么目标？

第 1 步

确定您所期望的理想结果。对于您的用例来说，成功的实施应该呈现出怎样的效果？

第 2 步

评估影响并衡量成效。考虑哪些组织流程会受到影响，以及受到的影响程度如何。

第 3 步

选择一种实施策略。探索您的选项。

第 4 步

启动测试并采用迭代方法。

第 5 步

制定管治标准并解决数据安全问题。

第 6 步

为团队设定一个与目标相匹配的时间表。

不过，下面我们先来回顾一下基础知识。

第 1 部分：了解生成式 AI 的发展前景

您无需成为生成式 AI 方面的专家，也能制定出一份实施计划。然而，了解其中涉及的各个组件，将有助于您在整个实施过程中做出明智的战略决策。下面，让我们来逐一介绍这些构建块。

什么是生成式 AI?

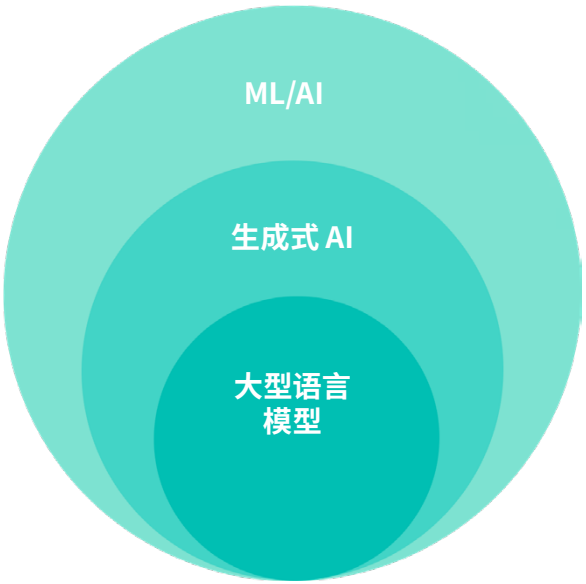
生成式 AI（即生成式人工智能）是指能够根据提示生成输出结果的深度学习模型。重要的是要明白，这项技术的生成能力取决于其能否根据已有数据，预测并生成符合统计学概率的输出结果，而这种能力是通过 Machine Learning 实现的。**数据是将生成式 AI 融入运营的核心，也是其成功实施和取得成效的关键。**稍后我们将对此进行更详细的介绍。

什么是 Machine Learning?

Machine Learning（机器学习，ML）是人工智能的一个重要分支，它通过使用算法从数据中学习和获取知识。这些算法会在监督式、半监督式或无监督式的学习环境中解析数据并“学习”其中的模式和相似性，进而使其能够做出决策。Machine Learning 是赋予生成式 AI（如大型语言模型）持续“学习”能力的底层技术。

什么是大型语言模型 (LLM)?

大型语言模型 (LLM) 是 Machine Learning 领域中的一种计算模型，是一种专门处理人类语言的生成式 AI。经过大量公共语言数据集的训练后，LLM 能够执行各种自然语言处理 (NLP) 任务，例如识别、分析、总结、预测、翻译或生成文本等。在将生成式 AI 融入实际运营的过程中，LLM 是让生成式 AI 能够用自然（或人类）语言进行交流的关键。



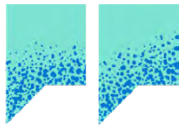
让我们来谈谈幻觉现象

幻觉是指 LLM 生成的不正确或误导性的结果。您可能对 ChatGPT 有时不那么准确的回答保持明智的判断，这是明智之举。输出看似合理，但真的靠谱吗？如果 LLM（ChatGPT 是基于 LLM 构建的）找不到答案，它往往会编造一个。当探讨在企业应用程序中使用 LLM 时，这个盲点是需要重点考虑的。您如何确保生成的输出既相关又准确呢？这时检索增强生成 (RAG) 技术就派上用场了。

什么是检索增强生成 (RAG)?

您可以将检索增强生成 (RAG) 视为对抗幻觉的一道防线。通过使用由矢量数据库支持的高度相关搜索，从特定的数据集或数据上下文中检索信息，从而对 LLM 生成的输出进行增强或“校验”。例如，通过 RAG，在响应用户查询时，组织会搜索自己的策略文档并向 LLM 提供相关回应，以便 LLM 使用组织的策略来回答用户的问题。除了作为对抗幻觉的防线外，**RAG 还可让您在自己的专有数据集上使用生成式 AI** — 这是它最大的优势。

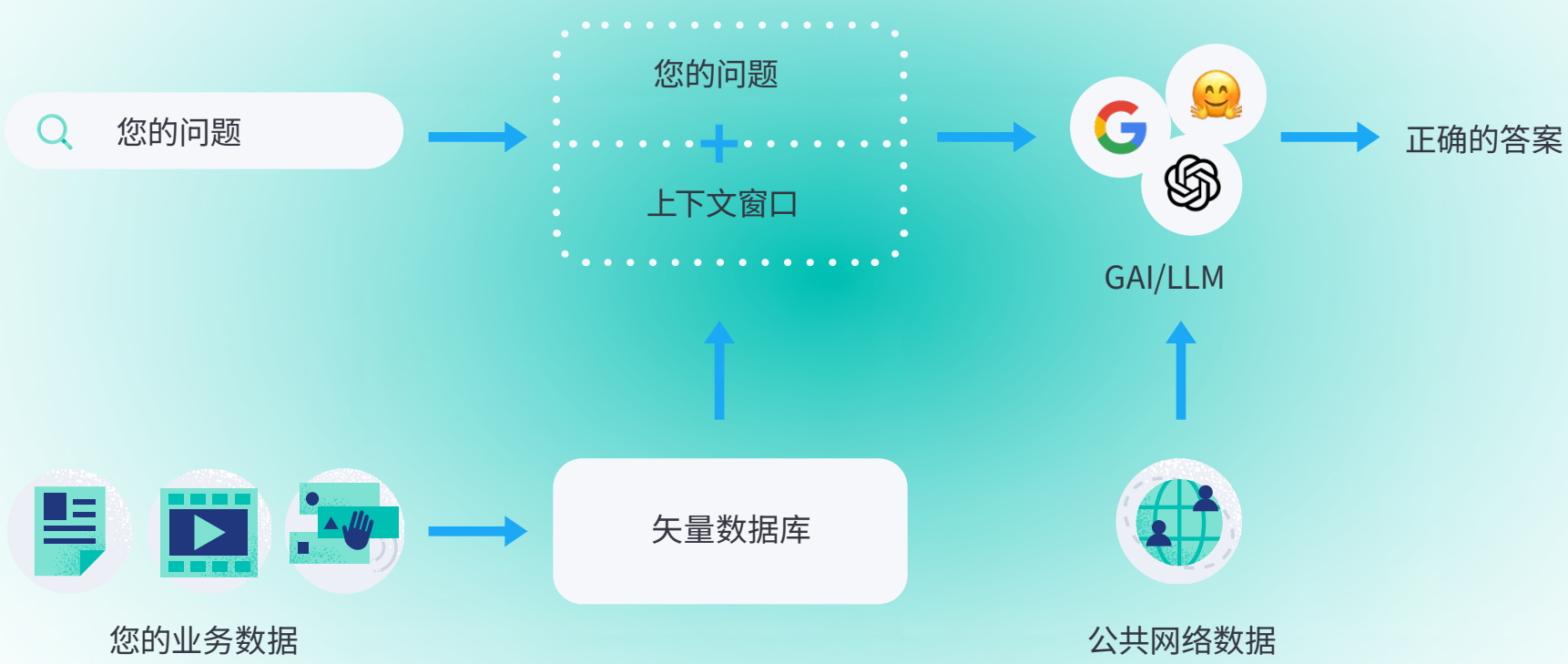
在为业务应用程序运用生成式 AI 的背景下，RAG 的重要性体现在多个方面：它不仅可以提供更优质、更相关的结果，并且提供了一种快速启动或利用自己专有数据的方法。另外，与训练或构建自己的 LLM 相比，RAG 也更具成本效益。因此，RAG 是生成式 AI 集成成功的关键。RAG 突破了传统 LLM 在多方面的界限，可助力打造“下一代搜索引擎”。



> **“借助 RAG 打造下一代搜索引擎”**
>
> **Baha Azarmi**
> **Elastic 全球客户工程副总裁**

检索增强生成 (RAG)

借助 RAG，您可以在自己的专有数据集上使用生成式 AI。



RAG 是一种回答用户问题的全新方式：

常规搜索

一位用户搜索**一个词**

居家办公政策

系统执行一个查询

```
query = {
 “bool” : {
 “should” : [
{
 “text_expansion” :{
 “ml.inference.text_expanded_
predicted_value” :{
 “model_id” :model_id.
 “model_text” :question
}
```

显示若干结果

文档标题	添加日期
员工行为规范	01/01/2010
IT 使用政策	01/01/2015
主页工作正在进行中	01/01/2022
等等……等等……	

用户选取一个文档并读取其中内容

员工行为规范

Lorem ipsum Lorem ipsum
Lorem ipsum Lorem ipsum
Lorem ipsum Lorem ipsum
Lorem ipsum Lorem ipsum
Lorem ipsum Lorem ipsum
Lorem ipsum

不使用 RAG 的生成式 AI

一位用户问**一个问题**

我们的居家办公政策是什么

系统执行一个查询

```
query = {
 “bool” : {
 “should” : [
{
 “text_expansion” :{
 “ml.inference.text_expanded_
predicted_value” :{
 “model_id” :model_id.
 “model_text” :question
}
```

得到的答案与您的领域没有关联

当您有员工采用混合办公模式或其他办公模式时，制定居家办公政策是必要的。

使用 RAG 的生成式 AI

一位用户问**一个问题**

我们的居家办公政策是什么

系统执行一个查询

```
query = {
 “bool” : {
 “should” : [
{
 “text_expansion” :{
 “ml.inference.text_expanded_
predicted_value” :{
 “model_id” :model_id.
 “model_text” :question
}
}
]
{
 “match” :{
 “text” : question
}
```

结果以带有上下文的形式提供

文档标题	添加日期
员工行为规范	01/01/2010
IT 使用政策	01/01/2015
主页工作正在进行中	01/01/2022
等等……等等……	

LLM 从搜索结果中得出答案

公司鼓励员工居家办公，前提是他们能够有效地……
等等……等等……

什么是矢量数据库?

矢量数据库用于存储矢量嵌入，也就是字词、图像或视频的数字表示形式。这些嵌入具有多维性，且支持语义搜索；语义搜索是一种旨在探寻查询意图和上下文含义的搜索方式。相比之下，文本搜索则仅查找与搜索查询中的关键字相匹配的结果。

在使用 RAG 提供的上下文中，矢量数据库能够根据提供给生成式 AI 的提示进行快速的语义搜索。这种高效的搜索能力也正是 RAG 得以成功的原因。

生成式 AI 在处理 NLP 任务方面表现出色，但传统的关键字搜索不能直接处理自然语言，因此无法提供最佳结果来供生成式 AI 使用。因此，通过使用矢量数据库为生成式 AI 提供与原始提示语义相似的搜索结果，生成式 AI 可以生成更加相关的答案。您可以将矢量数据库想象成一个知识库，可使生成式 AI 利用准确信息来回答问题。

然而，生成式 AI 并不局限于矢量数据库。使用 RAG，生成式 AI 可以接入关系数据库、图形数据库、基于文档的数据库，甚至关键字搜索引擎。最适合您的数据库往往取决于数据的特性、所使用的特定算法，以及系统的性能需求。例如，关系数据库适合存储结构化数据，而图形数据库则非常适合存储具有复杂关系的数据，而传统搜索引擎则适合进行全文搜索。

“所有方式最终都会趋向于混合搜索。”

Serena Chou，
Elastic 产品管理总监

尽管**语义搜索**能提供与查询含义相匹配的结果，但**关键字搜索**在将结果与查询中的确切关键字匹配方面仍然发挥着重要作用。混合搜索是一种结合使用矢量（常用于语义搜索）和关键字搜索技术的实践方法，旨在为生成式 AI 提供尽可能相关的搜索结果。

总而言之：混合搜索解决方案最有可能为您组织中的生成式 AI 体验提供最相关的搜索结果。

生成式 AI 能做什么

我们已经探讨了诸多底层技术和基本概念，但生成式 AI 具体能做什么呢？

创建

生成式 AI 通过学习训练数据中的各种模式来“创建”或生成输出。它通过对现有数据进行迭代，能够生成新的想法、图像和见解等。

总结

凭借强大的自然语言处理能力，生成式 AI 可以分析文本并进行总结。需要在短时间内审阅冗长的文档？让生成式 AI 来拯救你吧。

发现

生成式 AI 的关键在于其底层的搜索技术。这使得生成式 AI 工具能够接收查询指令，搜索庞大的私有或公共数据集，并据此作出回复。

自动化

假设您的组织使用两个不同的云平台提供不同的服务。每个云平台都会生成不同格式的日志。通过将这些数据自动转换为相同的格式并将其与 AI 进行映射，您的团队可以使用生成式 AI 就数据进行总结和提问。这样，您的 IT 团队就能够摆脱繁重的任务，专注于监测和管理系统。

第 0 步：列明部署生成式 AI 的理由并确定可能实现的目标

在这么多可能实现增值方式中，锁定一个起点至关重要。那么，如何充分发掘生成式 AI 的潜力来为您的团队赋能，满足客户不断变化的期望，并将公司推向新的高度呢？那就是锁定生成式 AI 能为组织带来最大价值的那个领域。

请考虑以下问题：

1 我想解决什么问题？

您的业务中是否存在效率特别低下的环节？您的员工主要在哪些重复性任务上投入了大量的时间？他们是否经常需要在内部数据库或外部搜索引擎中查找已存在的信息？

例如：

您的员工在查找信息方面 — 无论是查找项目的最新信息还是人力资源相关信息 — 是否会遇到困难？您的安全团队是否有可以实现自动化的任务，以便让他们有更多时间采取更为主动的行动？在您的客户服务流程中，是否存在信息熵增现象，比如客服工程师并不总是能实时与客服代理沟通已知问题和最新的解决方案，导致信息不一致或延迟？

温馨提醒：

您需要与那些工作流程将受到影响的团队密切合作。举例来说，如果您打算更新效率低下的 HR 流程，那么非常有必要从一开始就让 HR 部门参与进来，这有助于获得利益相关者的批准，并获得更多支持。

2 我能通过知识库来解决这个问题吗?

知识库是一个集合，包含各类信息性内容，无论是支持文章还是内部流程，都可以被收纳其中。

请考虑您可以从哪些内容中提取信息，以及是否可以通过利用这些内容来更有效地解决问题。仅实现任务自动化和个性化回复是否就足够了呢?

例如: 您已经发现，员工需要花相当多的工作时间来查找与 HR 相关的信息。由于您的团队没有 HR 专家，员工通常会被引导到公司内部网，在那里他们不得不阅读大量的政策文件，例如了解自己今年还剩多少天的假期。要解决这个问题，您需要建立一个知识库，其中可以包含这类 HR 政策文档，并需要访问员工个人数据以便提供个性化的回应。

“牵一发而动全身。”

Baha Azarmi，
Elastic 全球客户工程副总裁

第 0 步是识别限制工作效率的流程。通过削减繁琐任务，您可以给员工更多空间来发挥创意。一个执行得当的实施方案是负责任的体现。**投资于员工，根据需要提升他们的技能，并规划调整后的工作流和流程，这些都是与生成式 AI 成功结合的关键要素。**

在探寻实施“理由”时，请牢记 **K.I.S.S. 原则 — 保持简单而具体。** 确定您想要首先解决的生成式 AI 用例，是将生成式 AI 融入运营的关键第一步。之后，小型项目将为您的有效实施铺平道路。

例如，在之前的 HR 场景中，生成式 AI 可以应用在多种用例中：

1

发现

员工在界面上查询 — 我今年还剩多少天带薪休假？为了回应这个问题，AI 需要进行搜索，并通过 HR 政策文档和员工记录来提供与查询相关的文档。

2

总结

更进一步，生成式 AI 可能会分析这些文档并以对话方式为员工总结相关信息。“你今年还有 10 天带薪假和 4 天浮动假可用。请查看内部网页上关于带薪休假政策的内容。”

3

创建与自动化

聊天机器人可以帮助经理节省时间，比如通过创建回复来批准或拒绝休假申请，并给出相应的理由。此外，它还能够创建日历邀请并在系统中记录带薪休假请求。

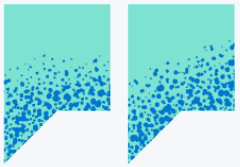
适应不同行业需求

辅助工具、助手、机器人 — 生成式 AI 以各种 “形态”，在多个领域和行业提供宝贵的生产力提升服务。无论是作为安全性和可观测性辅助工具，或是与内部和外部应用协同工作，生成式 AI 都可帮助公司提高效率，加强安全措施，改善客户体验，并加速形成竞争优势。

通过利用数据解锁最佳 AI 回应，并充分发挥底层搜索技术的强大功能，**公司能够减少在重复性任务上花费的时间，缩短回应时间，从而全面提升工作效率。**添加 RAG 后，可让您利用专有数据生成既安全又符合文档和用户级别权限要求的生成式 AI 回应。

突然间，您会发现速度和相关性都有了显著提升 — 这正是越来越精通科技的客户所期望的服务标准。确保您的服务能够符合这些期望至关重要。同样地，将生成式 AI 应用到与用户体验最紧密相关的场景中也非常关键。最糟糕的情况莫过于，投入大量资源构建的项目却无人问津。

生成式 AI 通常会被整合到公司的 IT 基础架构中，作为 AI 助手或者安全性和/或可观测性的辅助工具。



每个企业都有利用生成式 AI 的机会，因为生成式 AI 从信息系统中获取信息的方式在本质上更加符合人类直觉和需求。

Ash Kulkarni

Elastic 首席执行官

AI 助手

借助内部和外部应用，员工和顾客可以最大限度地利用生成式 AI 的对话技能。AI 助手能够作为随时待命的专家、个人购物助手，甚至是日程安排助手，以灵活、适应性强且个性化的方式，为每位用户提供帮助。

安全性和可观测性辅助工具

通过生成式 AI 辅助工具，您可以提升可观测性和安全能力。这些辅助工具专为与 IT 团队协同工作而设计，可充当专业的问题解决伙伴。例如，您可以向辅助工具发出请求，让其详细描述安全告警被触发的原因，并（根据组织以往遇到的类似攻击）获取推荐的攻击分类和修复步骤。这种类型的提示可以为组织生成动态运行手册。

借助这些集成，各行各业的公司能够提升他们的个性化、自动化和潜在的工作效率水平，进而衍生出**三个主要的生成式 AI 用例**：

改善运营弹性

运营弹性对于保持系统平稳运行至关重要。借助生成式 AI，IT 团队可以加速根本原因分析，跨各种环境关联更多数据，以更快地找出问题所在，并且拥有专门的发现工具来加快团队响应速度 — 这一切都是为了保障业务连续性。

提升客户体验

客户满意度是每个企业的核心。生成式 AI 不仅能为您的团队提供快速解决问题的工具和获取所需信息的途径，同时还可给予客户个性化的关注，确保他们能够快速访问到所需信息。那么结果如何呢？不仅改善了客户体验，还直接促进了业务成果的提升。

降低安全风险

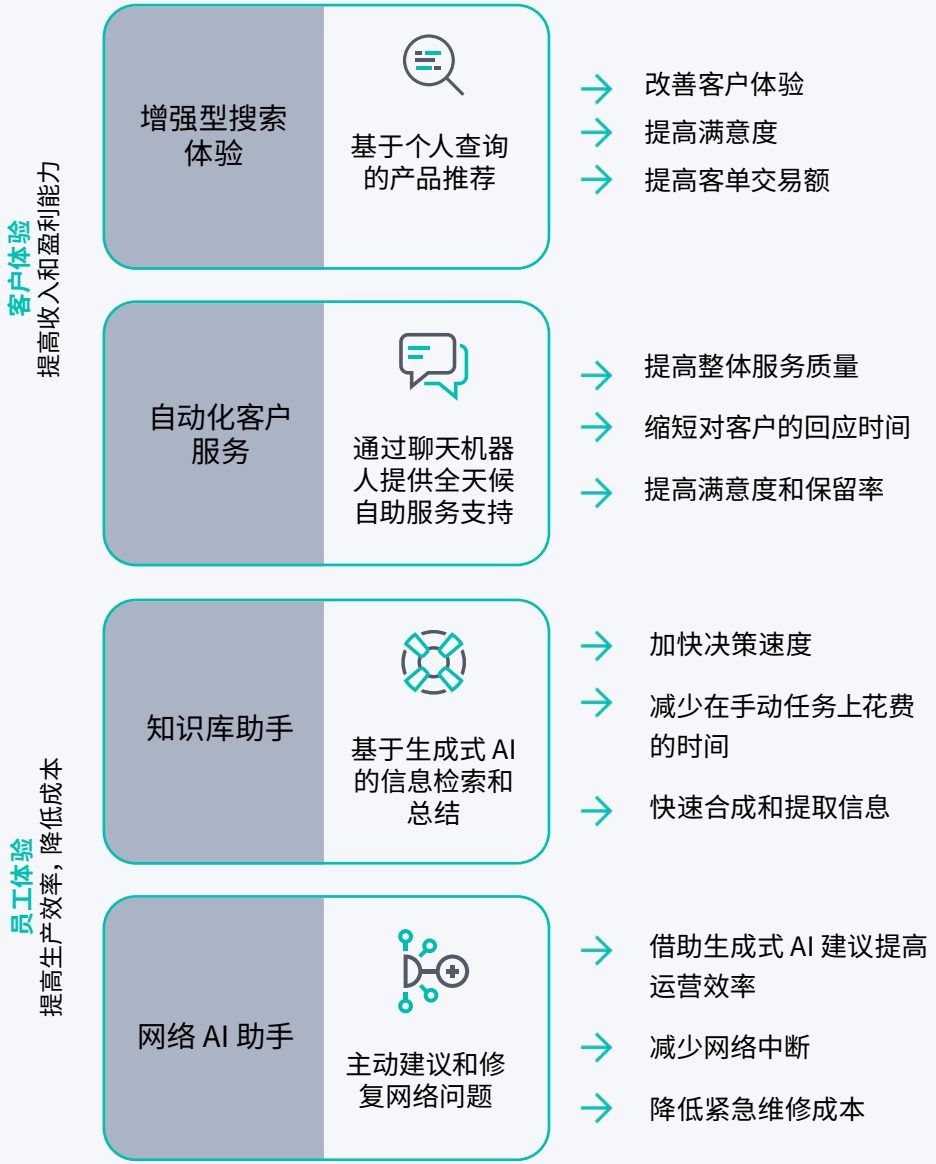
随着数字世界飞速发展，新型复杂的安全威胁层出不穷。要应对这些威胁，不仅需要灵活且主动的措施，还需要具备应对和管理这些威胁的专业知识。生成式 AI 不仅可以增强您的安全团队和运维能力，还能实现告警自动化并保持主动防御态势。

在各行各业中，生成式 AI 都能通过提供个性化、相关且具有指导性的回应来增强现有的员工和客户体验。无论您从事哪个行业，都可以找到一种方法将生成式 AI 融入您的搜索功能，从而解锁数据的新功能。

电信

对于电信公司来说，预计生成式 AI 将创造超过 600 亿美元的经济价值[4]。利用生成式 AI，电信公司能够让员工和客户在查询其网站或内部知识库时，快速获得个性化且相关的回应。那么结果如何呢？客户服务体验更好，工作效率更高。

[4] 来源：麦肯锡；“拨开炒作迷雾：挖掘 AI 和 GenAI 在科技、媒体和电信领域的潜力”(2024)。



金融服务

借助生成式 AI，金融服务公司能够进一步为客户和员工提供个性化体验。预计在客户体验、预计通过改善客户体验、加强欺诈预防和推动自动化，金融服务行业将创造超过 2500 亿美元的经济价值[5]。

[5]来源：麦肯锡；“生成式 AI 的经济潜力：下一个生产力前沿”(2023)。

客户体验
提高收入和盈利能力

零售银行助手	基于生成式 AI 的信息检索和总结	→ 扩大个人财务的可见性 → 提供量身定制的优惠以提高转化率 → 提高满意度和保留率
增强客户服务	主动建议和修复网络问题	→ 提高整体服务质量 → 缩短对客户的回应时间 → 提高保留率

员工体验
提高工作效率，降低成本和风险

欺诈检测总结	异常检测/交易摘要和后续最佳行动	→ 提升欺诈检测的准确性和速度 → 通过实现任务自动化降低成本 → 减少财务损失
虚拟助手	基于 NLP 的信息检索和总结	→ 加快决策速度 → 减少在手动任务上花费的时间 → 快速合成和提取信息

零售

对零售业而言，生成式 AI 的吸引力在于其有望通过提高搜索结果的相关性、推荐更多产品，以及跨渠道发送个性化跟进信息来增加客户保留率。您有没有收到过“您的购物车里还有未购买的商品！”这样的邮件？AI 可以自动化并改进这些功能，提供更优质的建议和更个性化的产品发现体验。

无论是打造下一代客户体验以促进电子商务销售，还是利用最新技术赋能员工以提高工作效率，生成式 AI 预计将为零售商创造超过 2,400 亿美元的经济价值[5]。

[5]来源：麦肯锡；“生成式 AI 的经济潜力：下一个生产力前沿”(2023)。

客户体验
提高收入和盈利能力

个性化的产品搜索和发现	问题解答，量身定制的搜索体验	→ 提高网站访客转化率 → 提高客单交易额 → 提高满意度
增强客户服务	通过聊天机器人实现自助互动	→ 缩短对客户的回应时间 → 提升服务质量，减少客户流失 → 提高保留率

员工体验
提高生产效率，降低成本

增强客户服务	优化代理人员体验和互动	→ 首次联系即解决问题 → 更快的入职流程 → 减少代理人员流动率
预测性维护	评估关键系统的健康状况，以确定关键维护任务的优先级	→ 减少设备和系统中断 → 降低紧急维修成本 → 提升运营效率

案例研究：HSE

HSE 是欧洲直播电商领域的领军品牌之一[6]。

“对 Home Shopping Europe [HSE] 而言，要想取得商业成功，首先要关注的是网站个性化和相关性。”

Peter Strasser
HSE 软件开发人员

机会

与所有电商业务一样，搜索功能对于客户体验和销售业绩至关重要。HSE 必须满足来自多种渠道的客户购物需求，因此会产生反映客户如何接触到产品的多样化搜索词。

HSE 利用生成式 AI 和 LLM 来提取客户查询的语义含义，并生成与传统关键词匹配相补充的搜索结果。

结果

由于搜索结果更准确、更相关，HSE 的**点击率提高了 4%，客户满意度提高了 8%**。

见解

专注于一个您已经想要改进的领域，比如客户搜索体验。看看如何整合生成式 AI，将体验提升到新的水平，实现更高的个性化和相关性。

[6] 来源：Elastic；“通过使用 Elasticsearch on AWS，HSE 不仅提高了客户满意度，还将维护时间缩短了 42%”(2024)。

汽车与制造

借助 AI 技术，汽车和制造业流程的每一个环节都可以得到优化和精简，预计将可为该行业带来超过 1,700 亿美元的经济价值[5]。从产品研发创新到实施个性化的客户维系战略，生成式 AI 具备彻底改变整个行业的潜力。飞行汽车？也许未来可期！

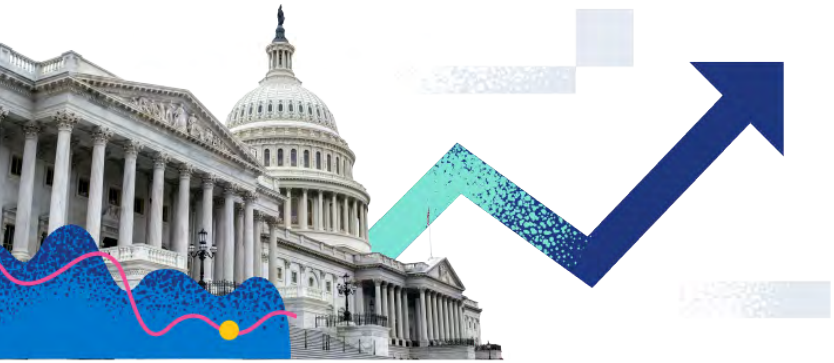
[5]来源：麦肯锡；“生成式 AI 的经济潜力：下一个生产力前沿”(2023)。

客户体验
提高收入和盈利能力

交互式数字手册	虚拟产品助手	→ 实时回答有关产品特性、维护和故障排查方面的问题 → 减少客服咨询 → 提高满意度
增强客户服务	通过聊天机器人实现自助互动	→ 缩短对客户的回应时间 → 提升服务质量，减少客户流失 → 提高保留率

员工体验
提高生产效率，降低成本

运营技术优化	预测性维护：问题及解决方案概要总结	→ 快速发现并解决问题提高运营 → 效率和决策速度 → 降低制造成本
产品情感分析	总结产品占比情况并提出改进建议	→ 根据客户视角改进产品供应 → 缩短新产品实现价值的时间周期



公共领域

生成式 AI 通过安全地将自身与机构数据连接起来，可以显著加速任务成果的产出，改善民众服务，并帮助政府分析师和安全专业人员在正确的时间获取到正确的数据。

面向民众的应用包括：

- 以个性化方式获取公共服务
- 优化的线上民众体验
- 提高可访问性和增加自助服务选项

面向员工的应用包括：

- 更准确的调查和情报
- 通过自动化手动流程和工作流来提高工作效率
- 更高效的采购流程

减轻工作负担
自动化手动流程和工作流

合规性
启用基于角色的数据访问权限

实时态势感知
做出更准确的决策

员工工作效率
在合适的时间获取所需的信息

民众体验
通过个性化的数字互动建立信任

公共服务
提高可访问性和增加自助服务选项

动态情报
加快任务搜索和见解的生成

网络安全
进行实时风险评估和分析

案例研究：Relativity

Relativity 致力于帮助企业、律师事务所和机构存储和利用数据进行电子取证和法律领域的信息搜索[7]。

“Relativity 客户目前面临的最大挑战是来自异构数据源的数据量激增。而这些数据又因通信模式不同所产生的差异，使得这一挑战变得更加复杂。”

Brittany Roush
高级产品经理

机会

Relativity 需要在整合数据的同时坚持安全第一的原则。面对数据量、来源和复杂性的激增，传统的关键字搜索方法变得不再奏效。于是，RAG 上场了。

结果

结合 RAG 和矢量数据库，Relativity 打造了基于专有数据的搜索体验，为用户提供了快速、相关且准确的搜索服务。这一生成式 AI 解决方案符合 PCI、DSS、SOC2 和 HIPAA 等合规性标准。

见解

在构建之初，请务必考虑未来的扩展性。从小规模开始，有助于更好地发掘生成式 AI 的潜力，并专注于那些最相关的应用场景。一旦找到最佳的契合点，未来发展将充满无限可能。

[7]来源：Elastic；“Relativity 现已使用 Elasticsearch 和 Azure OpenAI 打造前沿的搜索体验”(2024)。

您已经了解到生成式 AI 在各行业中蕴含的巨大经济潜力。或许您还已经构想了某些潜在的用例。希望您也已经明确了实施这一技术的初衷和动力。

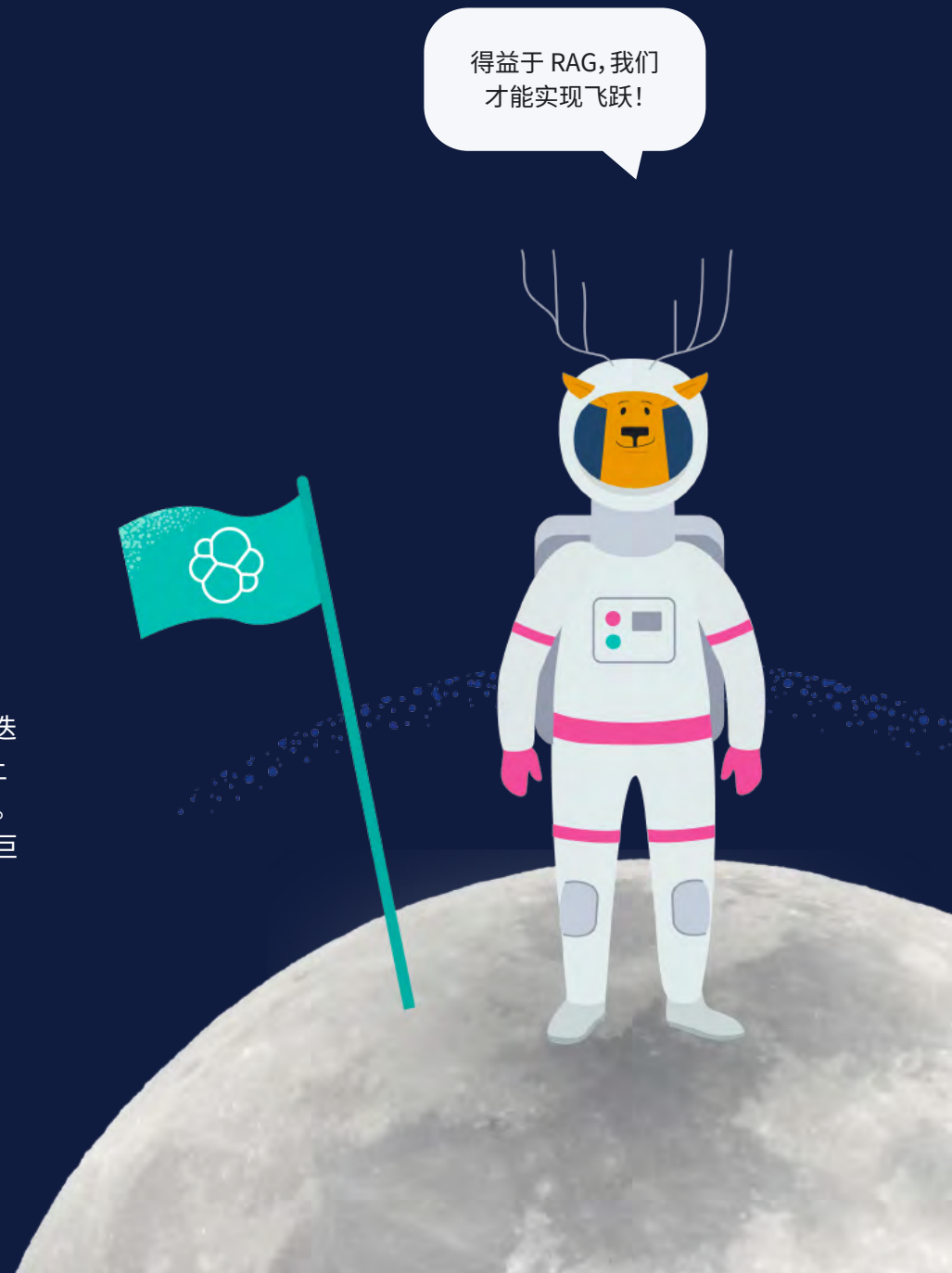
然而，实施生成式 AI 可能看起来是一个艰巨且具有颠覆性的过程。这涉及隐私问题，需要在合规方面进行一些基础工作，并且还会改变员工的工作方式。负责任的实施需要对员工进行培训、提升技能，以及对部分员工团队进行重组。

尽管面临重重挑战，但生成式 AI 能够为公司带来的价值是无可否认的。为保持竞争力，实施生成式 AI 是必然之举。好消息是，在测试阶段，您就可以开始快速实现价值回报，而不必等到完全准备好投入生产后再看到成效。换句话说，现在就是开始行动的时候了。

第 2 部分：

将生成式 AI 融入运营

对机器而言只是微不足道的一小步，但对组织来说却是意义非凡的一次跨越



将生成式 AI 融入运营并非一蹴而就，而是一个需要计划和明确目标的迭代过程。先从一个简单的生成式 AI 项目开始（即迈出这一小步），可以让团队逐渐适应学习曲线，使他们有机会完善流程、微调技术并解决问题。这样一来，您的团队和公司就能够迈出成功的第一步，为实现期待中的巨大飞跃奠定坚实基础。

现在，我们来看看如何实现这一目标吧。

步骤 1 确定您所期望的理想结果

您已经锁定了一个问题。您知道自己想要优化一个低效的流程。现在，您需要思考用户将如何与您的解决方案进行交互。您是要增强搜索应用程序还是聊天机器人？或者您是寻找一种与团队或客户互动的新方式？

您的思考过程可能如下：

- 您希望提高客户保留率。
- 于是，您决定实施个性化的产品搜索和发现应用。
- 接着，您设定了成功的衡量标准。这可以视为您的“次级目标”。

  您想要利用生成式AI。为什么？为了个性化产品搜索和发现。为什么？**以下是您所期望的理想结果：**通过与我们的数据以全新的方式互动，客户将能轻松找到他们所需的产品，并根据他们的搜索历史和位置，自动向他们推荐可能感兴趣的产品。这样一来，客户保留率将得到提升。
- 现在，您要开始实施第一个生成式 AI 项目了，这可是一个庞大的任务。

问问自己：

通过这种新的数据交互方式，我们可以采取哪些行动并达成哪些成果？

这个问题的答案将有助于我们设定目标。通过确定您所期望的理想结果，这将决定您项目的“优秀”标准，并在更广泛的层面上，指导您公司的发展方向。

步骤 2

评估影响。衡量成效。

要成功地将生成式 AI 融入运营，您需要制定一套关键绩效指标 (KPI)，帮助衡量对您而言的 “优秀” 意味着什么。了解生成式 AI 如何提升组织的生产力只是众多绩效指标中的一个。

其他指标还可能包括通过客户支持环境中的评价来衡量的客户满意度的提升、支持工单量的减少，或问题解决时间的缩短。根据您正在测试的用例，您需要设定相应的绩效指标。将这些指标融入测试过程的每个步骤中至关重要，有助于您和团队了解所取得的进展情况。

基本绩效指标

1

工作效率影响

评估您的用例对工作效率的影响。比较在使用和不使用生成式 AI 完成特定任务所需时间的变化。

2

可扩展性

评估模型在使用量和需求增加时的扩展能力。它是否仍然能够可靠且准确地运行？

3

利润

评估实施生成式 AI 对业务成本的影响。在这个评估中，您可能需要考虑一些业务指标，比如记录的客户投诉数量或销售业绩的变化情况等。

4

合规性

持续监测生成式 AI 是否遵守数据隐私法规。

5

客户满意度

审查业务指标，如客户流失率、销售业绩增长情况和品牌忠诚度维持情况，并分析客户反馈。

利用这些指标来评估项目的可行性、可操作性、可扩展性和成本效益。这些指标将有助于您确定投资回报率，并且随着您拓展用例，这些指标还可以进一步丰富和扩展。

步骤 3

选择一个模型（未来行动计划）

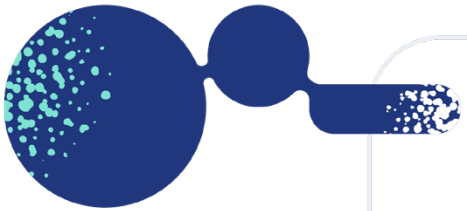
如何构建能满足业务需求的生成式 AI 架构？许多因素会影响您的选择：成本、语言、IT 生态系统、部署能力和时间表、数据隐私法规以及管治。因此，采取一种审慎的态度 — 从简单而具体的用例开始 — 是至关重要的。

要让生成式 AI 运融入运营，您需要以下几个组件：

- **一个完全托管的云基础架构**：有助于提高系统的敏捷性，优化成本效益，并减少资源浪费。鉴于芯片和硬件正在以惊人的速度发展，如果您选择投资建设自己的 AI 数据中心，那么它很可能在几个月内就会变得过时。
- **一个 LLM**：使生成式 AI 能够用自然语言进行沟通和理解的基础。
- **一个数据平台**：具备矢量搜索、混合搜索和传统关键字搜索功能，可用于从您的专有数据中为 LLM 提供合适的上下文，从而丰富其数据。
- **广泛的 API**：可让您扩充数据并将其传递给 LLM 和搜索引擎。

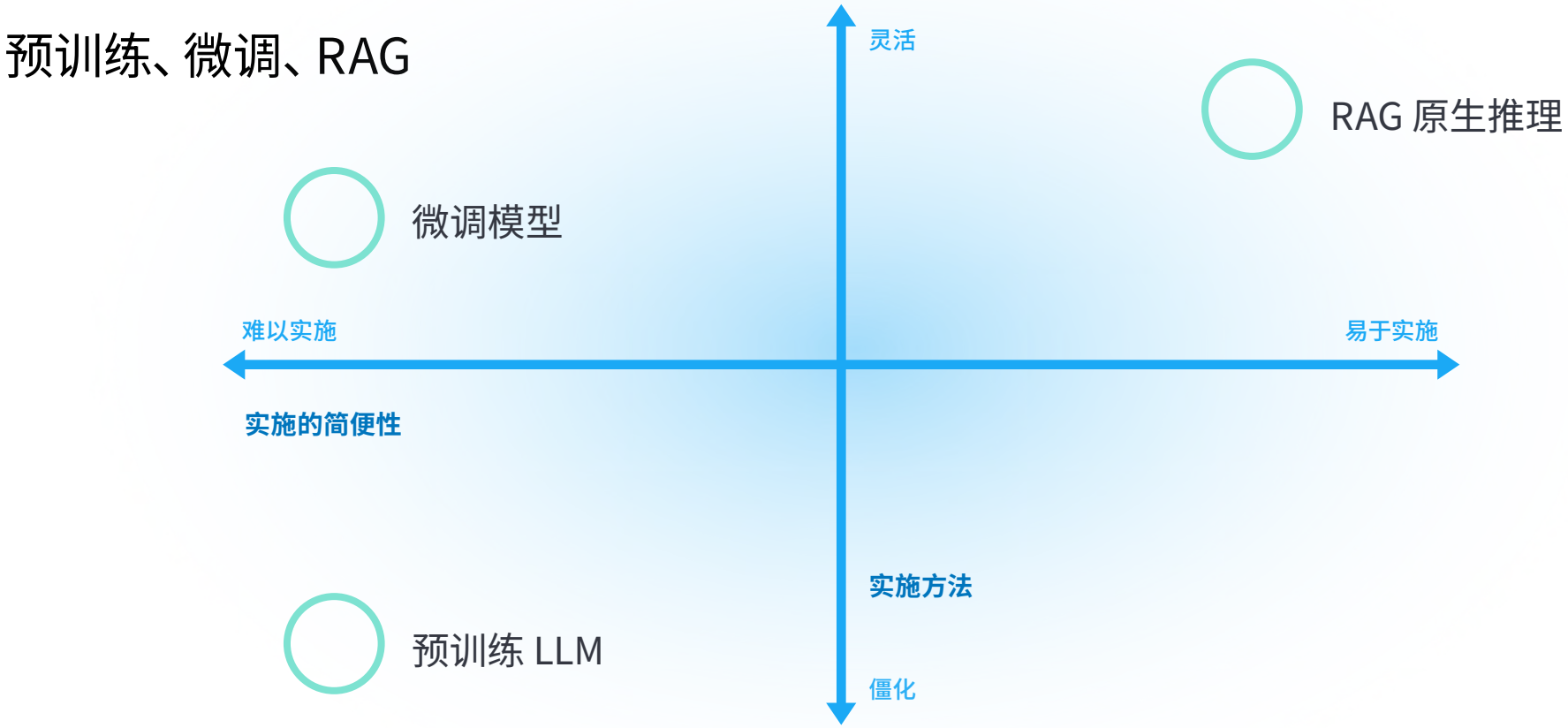
企业实现 AI 搜索所需的要素

如何组合这些组件呢？无论是微调自己的模型、自带矢量数据库、自带模型，还是它们之间的任何组合，都将决定您的实施方案，进而影响时间表、测试的复杂性，以及是否需要补充团队能力。



	预训练 LLM	微调模型	RAG
	这种资源密集型方法需要从零开始，在大量数据集上训练大型语言模型。	这种方法需要利用现有的 LLM 与您的搜索引擎和矢量数据库相结合，为您的专有数据提供上下文信息	这个过程需要使用现有的预训练 LLM 和一系列技术来微调模型，以满足您的需求。
成本	$$$$	$$$	$$
部署时间	长（数月）	中等（数周）	短（数天）
数据隐私性	您是否拥有足够大的数据集，以便为 LLM 提供充足的学习材料？如果没有，可能需要使用公共数据。您是否会将公共数据和私有数据结合起来？	您是否拥有足够大的数据集，以便为 LLM 提供充足的学习材料？如果没有，可能需要使用公共数据。您是否会将公共数据和私有数据结合起来？	这种方法能够确保您的私有数据保持私密。
准确性和相关性	很难一贯保持	对于已经进行调优的模型，针对其特定的任务，准确性和相关性更容易得到保证。	RAG 的优势尤其在于它能够通过引用来源或让用户知道它无法回答问题来 “消除” 幻觉。

请考虑这些选项：



选择正确的前进方向

对于每个人来说，并没有一条绝对正确的道路，因此请确保您的决策与在之前步骤中设定的目标相符，并且这些目标与相关方保持一致。最终，您需要找到一条能清晰地向业务利益相关者和团队阐明的路径。

除非您打算进行更全面的改革和更大规模的项目，否则从头开始构建 LLM 并进行微调会非常耗费资源。您将面临各种问题：比如，是否需要一个矢量数据库来补充您的搜索引擎？能否升级搜索引擎，在其中创建和存储嵌入式数据，并构建逻辑以持续推动搜索功能？如何将其扩展为一个推荐系统？

对于具有混合搜索和语义搜索功能的基于云的解决方案而言，操作可能会相对简单。通过连接到现有 LLM，您可以使用 RAG 为客户打造更贴近实际需求的搜索体验。

以下是需要考虑的事项：

- 首先，需要对您当前的 IT 环境进行全面评估。通常情况下，并不一定需要重新构建基础架构。
- 可以考虑采用开箱即用型的 (OOTB) 解决方案，例如开箱即用型的 LLM、矢量数据库和全套软件包等。

  赞成： 黑盒技术可能会让您更快地完成部署和运行。

  反对： 但是它们的可定制性极低，您无法以同样的方式进行扩展。
- 找到一款既具备灵活性又尽可能减少干扰的补充产品。您需要对搜索相关性和性能进行基准测试，并可能需要更换模型，以确定最适合您的方案。

步骤 4

大胆试错，快速迭代

快速发展的数字生态系统，意味着生成式 AI 项目中包含许多动态变化的部分。对于预训练的 LLM，您的控制范围有限，同时调整架构的灵活性也有限。

现在是时候采取迭代方法了：确定用例、设定期望结果、制定关键绩效指标 (KPI)，并考虑如何实施生成式 AI 项目。

记住

从根本上说，将生成式 AI 融入运营就是要从您的数据中获取答案。请务必时刻牢记合规性问题。您设置的测试是否可能会违反隐私政策？这是一个低风险的测试吗？

在这个阶段，您需要做到以下几点：

- **建立反馈循环：** 确定谁向谁报告，以及报告的内容，并确定项目中的关键利益相关者。
- **扩充您的 LLM：** 确保您的 LLM 能够访问存储在矢量数据库中的正确信息。矢量数据库将使您能够快速提供最为相关的信息来扩充您的 LLM。
- **优化用户体验：** 设计一个方便用户使用的界面并持续进行测试。最终，生成式 AI 的目的是为员工和客户服务。构建一个符合应用程序和用户需求的界面，对于成功实施生成式 AI 项目至关重要，并且能够确保其可扩展性。
- **建立一个可扩展的参考架构：** 在测试生成式 AI 项目时，要着眼大局。当您扩展项目规模或进一步拓展用例时，如何调整架构？

举例来说，如果从零开始构建矢量数据库让人感到任务繁重，您可以考虑使用可下载的现成数据库 — 没错，这样的数据库确实存在。有了这个矢量数据库，您就能够迈向更高的层次：混合搜索。在搜索应用程序中使用语义搜索，可以帮助您测试下一代 AI 项目的原型。这就是从小规模开始并不断迭代所展现出的力量。

步骤 5

管治和运维

生成式 AI 项目会带来一系列的挑战，例如：数据隐私和合规性问题、伦理考量、质量控制，以及风险管理等。您需要提前预见潜在的障碍，并确保您的项目与业务目标保持一致。

在进行管治和运维审查时，您需要考虑下面的一系列因素：

- **成本管理：** 按照每千个词元收费；提示和回应分别计费。
- **日志：** 您需要记录每个回应，以查看模型与客户之间的沟通情况，以进行质量控制。
- **确定回应情感：** 确定 LLM 回应的情感倾向，以使其符合公司的品牌调性（这是另一个重要的质量控制步骤）。
- **监测幻觉：** 幻觉包括错误或误导性信息，但也可能包括聊天机器人输出的仇恨言论和反社会行为。
- **标记不确定的答案：** 监测回应的质量和相关性对于质量控制至关重要。这有助于了解哪些应用程序需要更多的人工介入，从而在需要扩展时提前做出规划。

关于 AI 中的偏见问题

生成式 AI 模型依赖于它们训练所使用的数据。如果训练数据存在偏见或局限性，这些问题就会在输出结果中反映出来。

组织可以通过仔细考虑和限制训练模型所使用的数据，或使用符合自身需求的定制专用模型来降低这些风险。然而，编写这项技术或负责整理训练模型所需数据的人员也可能存在偏见。

偏见在任何情况下都难以根除。但这并不意味着组织可以就此放弃，而是应该引导用户在解决问题时进行一些批判性思考。

此外，还需要让您的法律团队参与进来，并确保将他们的工作纳入概念验证的环节。尽管他们的参与可能会减缓测试阶段进展，为了实现负责任、合乎伦理且合规的实施，建立全面而高效的审查流程是至关重要的。

关于数据安全的问题

由于安全威胁每天都在影响着各个组织，因此数据安全至关重要。您的客户信任您保护他们的数据，这正是许多公司采用“零信任”框架的原因。该框架秉持的原则是，无论用户和设备位于组织网络边界之内还是之外，都不应自动或默认地给予其信任。

为了优化安全性，您可以考虑以下措施：

1. **采用 RAG 方法：** RAG 模型利用检索机制来更好地理解输入提示的上下文，从而生成更贴合情境且剔除了敏感细节的回应。当与具有文档级和基于角色的安全性的数据平台结合使用时，RAG 能够确保权限得到尊重。

2. **投资或扩展您的可观测性解决方案：** 解决信任问题。利用监测功能追踪数据轨迹，并监测生成式 AI 项目产生的回应。了解您的数据流向了哪里，以及生成式 AI 对客户传达的信息。

最终，要将生成式 AI 引入您的生态系统，就需要建立新的操作规程，并据此制定新政策。有了更高效的流程和更高的收入，从执行琐碎任务中节省下来的时间可以转投到这些更重要的工作中。

步骤 6

设定时间表。给出基准点。

列出一个时间框架，比如以一个季度为例。在这个时间框架内，设立 30 天和 90 天的目标。通过这个季度来验证基于生成式 AI 增强的用例所能带来的价值。

30 到第 30 天，您需要启动首个测试了。那么，情况应该是怎样的呢？

- [ ] 您已经选定一个用例
- [ ] 您已安排一个小团队来完成这个任务
- [ ] 您已经根据需要组织过培训课程
- [ ] 您已经确定了想要的结果
- [ ] 您已构建原型界面

90 到第 90 天，您将准备好推出首个用例了。那么，情况应该是怎样的呢？

- [ ] 您已向一些内部人员开放测试
- [ ] 您已对生成的输出进行过测试、调优和衡量
- [ ] 您一直在持续监测用户与界面的互动方式
- [ ] 您已经制定一套标准，用于明确高质量输出的关键要素
- [ ] 您已经收集了一些关键绩效指标的数据
- [ ] 您已经评估了这个计划的价值

这些任务应作为大致的基准。您公司具体的需求（比如团队构成和他们所使用或添加到技术堆栈中的技术）将影响您能够部署第一个用例并收集见解的速度。

在这个阶段，需要考虑以下几点：

1. **错误率：**衡量错误率。生成式 AI 是否能生成正确且相关的输出？这对于微调生成式 AI 至关重要。
2. **训练时间和成本：**衡量训练模型所需的时间和资源。这样做有助于确保测试阶段的高效性，从而加快投入运营的进程。
3. **人工干预：**生成式 AI 是否仅能在人工干预的情况下才能运行？为了保持可靠性和准确性，**需要多大程度的监督？**
4. **回应时间和输出质量：**衡量生成式 AI 提供输出的速度，并将输出质量与一套既定规则或指南进行比较。

就像这样，您就可以做好准备将生成式 AI 成功融入运营并扩展规模了。

新时代大幕已经开启

许多行业领导者已经领略到生成式 AI 带来的益处，越来越多的企业也在努力复制这种成功，并紧跟不断变化的客户期望。生成式 AI 的创新正在飞速发展。但无论如何，成功都离不开坚实的基础。

制定最符合自身需求的生成式 AI 实施策略，可以帮助您充分发挥数据的价值，避免被那些令人兴奋但无关紧要的创新所分散注意力。为了最有效地将生成式 AI 融入运营，请分阶段投入时间和资源。有针对性地将新技术融入业务，是实现最高投资回报的秘诀。此外，根据运营需求定制和调整 AI 工具，可以确保其相关性和有效性，而这正是生成式 AI 的初衷所在。

请记住，在实施生成式 AI 时，必须负责任地处理数据隐私保护、安全性以及敏感性和伦理问题。除了为全球经济贡献数万亿美元的价值外，生成式 AI 还提供了一个机会，让公司内的所有人都能使用这项技术并提升技能。因此，您不仅要将自己定位为生成式 AI 的开拓者，还要成为公司开发新业务流程的先驱者。

让我们立即开始吧。

确定您的第一个生成式 AI 用例需要各团队通力合作。您需要让安全团队、IT 团队、开发团队和业务团队从一开始就齐心协力。**下面是 Elastic 如何为您提供帮助：**

与您的安全团队合作

您可以提高从业人员的工作效率，降低风险。要将生成式 AI 应用于您的安全用例，首先需要在开放平台上采用统一的方法。您可以借助 Elastic Security，利用生成式 AI 强大功能，为您的安全团队打造一个符合其需求的体验。

认识 Elastic Security

与您的 SRE 和 IT 运维团队合作

为您的 SRE 和工程师赋能，让他们能够使用交互式自然语言聊天界面，更快地找到最相关的信息。了解如何将对话式 AI 与 Elastic 可观测性和高级 Machine Learning 相结合，基于您的专有数据和运行手册，打造能感知上下文的交互式聊天体验。

认识 Elastic 可观测性

与您的开发团队合作

您可以升级开发团队的工具包，以帮助他们通过自助选项（如使用 Elastic Search 的高度个性化的聊天机器人）提供客户支持。同时，让客服代表也能用同样强大的搜索工具快速解决问题，并且还有生成式 AI 体验帮助他们从各种数据源中找到答案。了解如何为您的知识库实施强大的搜索功能。

认识 Elastic Search

elastic | The Search AI Company