

构建 Agentic AI 系统： 打造能推理、可规划、 自适应的 AI 智能体

[美] 安贾纳瓦·比斯瓦斯(Anjanava Biswas) 著
里克·塔鲁克达尔(Wrick Talukdar)
茹炳晟 殷海英 译

清华大学出版社

北 京

北京市版权局著作权合同登记号 图字：01-2025-3792

Copyright ©Packt Publishing 2025. First published in the English under the title Building
Agentic AI Systems – (9781803238753)

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。

版权所有，侵权必究。举报：010-62782989, beiqinquan@tup.tsinghua.edu.cn。

图书在版编目(CIP)数据

构建 Agentic AI 系统：打造能推理、可规划、自适应
的 AI 智能体 / (美) 安贾纳瓦·比斯瓦斯 (Anjanava Biswas),
(美) 里克·塔鲁克达尔 (Wrick Talukdar) 著；茹炳晟，殷海英译。
北京：清华大学出版社, 2025. 12. -- ISBN 978-7-302-70398-3

I . TP18

中国国家版本馆 CIP 数据核字第 20253KV496 号

责任编辑：王 军

封面设计：孔祥峰

版式设计：思创景点

责任校对：成凤进

责任印制：刘海龙

出版发行：清华大学出版社

网 址：<https://www.tup.com.cn>, <https://www.wqxuetang.com>

地 址：北京清华大学学研大厦 A 座 邮 编：100084

社 总 机：010-83470000 邮 购：010-62786544

投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈：010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者：三河市铭诚印务有限公司

经 销：全国新华书店

开 本：170mm×240mm 印 张：14.75 字 数：289 千字

版 次：2025 年 12 月第 1 版 印 次：2025 年 12 月第 1 次印刷

定 价：79.80 元

产品编号：113549-01

中文版推荐序

当前，人工智能正从“工具赋能”迈向“原生智能”的新阶段，以大模型、Agent、RAG 等为代表的 AI 技术不再局限于为传统系统叠加功能，而是通过重构底层架构与运行逻辑，催生全新智能化应用形态。2025 年 8 月国务院发布的《关于深入实施“人工智能+”行动的意见》更明确要求“培育智能原生新模式新业态，大力发展智能原生技术、产品和服务体系，加快培育底层架构和运行逻辑基于人工智能的智能原生企业”，凸显了这一领域的战略地位。

本书所介绍的 **Agentic AI** 系统正是一种典型的新型智能化软件系统。这种系统以基于大模型的智能化和自治性的 **Agent** 为中心，能推理、会规划、自适应。除此之外，这里的“系统”二字又突出了面向真实场景和现实需求的工程化要求：从外部用户需求看，这类系统需要具有高度的可信性，满足可靠性、安全性、透明性、可解释性以及伦理道德等多个方面的要求；从内部构造方式看，这类系统包含大模型、Agent、知识库、工具等多种不同类型的组成部分，并且包含一系列最佳实践和设计模式。简而言之，就是要通过使用工程化的方法和手段实现高度可信的工程化系统。

本书围绕这一主题，从基本原理、关键技术、设计模式、构造方法、可信保障等多个方面展开介绍，并提供了具体的案例分析和实现参考，为 **Agentic AI** 系统方面的初学者提供了一个全面和专业性的介绍。本书特别突出了智能体系统设计的重要性，包括相关设计模式和设计技巧，这对于工程化系统无疑是特别重要的。同时，本书还特别强调了 **Agentic AI** 系统的信任、伦理和安全问题。在人工智能特别是 **Agentic AI** 技术加速融入每个人的日常生活以及各行各业的当下，强调这些方面的潜在风险以及从工程化设计角度的应对指导无疑是非常重要的。

相信大家读完这本书之后，一定会在了解 **Agentic AI** 系统基本原理的同时，初步体验到如何以工程化目标为导向推动 **Agentic AI** 技术在各行各业的落地应用。

彭 鑫

复旦大学计算与智能创新学院副院长

行业名家力荐

作为 AI 领域的深度探索者,《构建 Agentic AI 系统》一书令我震撼且印象深刻。

本书系统性地剖析了智能体推理、规划与自适应的核心能力,不仅夯实了从 VAE 到 Transformer 的理论基础,更通过多智能体协作、工具调用等前沿实战案例展现了应用蓝图。尤为可贵的是,作者始终将伦理安全与可信赖设计贯穿全书,为开发者提供了兼顾创新与责任的实践指南。强烈推荐给所有致力于 AI 前沿探索的技术架构师与算法工程师!这是一本通往下一代自主智能系统的关键指南!

司晓 腾讯研究院院长

在大语言模型从“对话工具”向“自主智能体”跃升的关键时刻,本书的出版极具时效性。作为首批系统阐述 Agentic AI 构建方法论的著作,本书填补了从理论到工程实践的关键空白。全书围绕智能体的核心能力——推理、规划、反思与多智能体协作展开,深入剖析了工具使用、HTN 规划、CWD 协调模式等前沿技术。特别感谢茹炳晟老师的精心翻译,凭借其在 AI 工程实践领域的深厚积淀,译文不仅准确流畅,更在技术术语把握和工程细节呈现上展现出专业水准,为中国读者提供了 CrewAI、AutoGen、LangGraph 等主流框架的高质量实战指南。本书兼顾技术深度与应用实践,是推动我国智能体技术研发与产业落地的重要参考文献。

王昊奋 OpenKG 轮值主席

《构建 Agentic AI 系统》是一本很全面的入门介绍。二位译者把复杂的技术背景和抽象的理论都翻译得很得体,读起来不像是一本翻译的书。这本书的核心在于介绍了多智能体协作的协调者、工作者和委托者的模型。这个模型在处理简单事务上是足够好的,它能在安全的环境中快速完成任务。在并不安全的实际网络环境中,我们还需要“验证者”“领域专家”等角色协同作战,才能安全完成用户的愿望。我给读者的建议是,快速阅读全书并实践“learning by doing”的精神:从本书的 GitHub 项目开始,学习它的代码并快速扩展。把实用的智能体落地到中国实际的互联网或企业内网环境中,争取在实践中开发和训练各种智能体技术,给用户带来实际的收益。

邹欣 北京中关村学院工程中心负责人

当“Agent 时代”加速来临，这本著作恰逢其时地搭建起生成式 AI 与智能体技术的完整知识框架。从生成式 AI 的核心模型与安全挑战，到智能体的系统原理、组成架构，再到反思内省、工具规划与多智能体协作的模式，既拆解技术底层逻辑，又聚焦实战设计。

无论你是 AI 领域的研究者还是开发者，都能从中获得启发。

毛剑 哔哩哔哩技术委员会主席

三十年的技术沉淀让我见证了 AI 从理论逐步走进现实，作为架构师，我深知构建真正智能的 AI 系统有多么挑战。茹炳晟的译作《构建 Agentic AI 系统》精准捕捉了 AI 发展的关键转折点——从被动响应到主动推理、从简单执行到自主规划。这本书不仅提供了理论框架，更给出了实用工具和方法，帮助开发者跨越从概念到实现的鸿沟。这本译作既保留了原作的技术深度，又增添了东方思维的细腻，让复杂概念变得平易近人。感谢老友为中文技术社区带来如此珍贵的礼物！

曹洪伟 某 AI 公司首席架构师，《MCP 极简入门》作者

我们正从“会回答问题的模型”迈向具备目标驱动、规划执行与可控自治能力的 Agentic AI 系统。未来的领先企业，不在于资源堆叠，而在于能否构建稳定、可演化的 AI 生产系统——能够持续监控、自动执行、自主迭代。这本书的价值，不在于描绘愿景，而在于拆解机制、提供路径：从系统架构到角色分工，从任务规划、工具编排到安全与治理，让智能体从概念走向可落地、可规模化、可审计的工程体系。

如果你希望 AI 不只是“辅助工具”，而是成为组织的第二操作系统，支撑真实业务的持续运行与自我成长，这本书值得从头读到尾。

李佳芮 句子互动创始人&CEO

塑造未来无限可能的 Agentic AI

在人工智能发展的漫长画卷中，我们似乎正站在一个前所未有的转折点上。曾几何时，人工智能系统大多是被动的工具，它们遵循预设的规则，在划定好的边界内执行确定的任务。然而，随着生成式人工智能的爆发式进展，一种新的范式——Agentic AI——正迅速崛起。它不再满足于被动响应，而是追求主动思考、规划乃至协作；它不再仅仅处理信息，而是试图理解情境、做出决策并影响环境。这本《构建 Agentic AI 系统：打造能推理、可规划、自适应的 AI 智能体》的翻译工作，正是在这股浪潮奔涌而至之际，试图为中文读者架起的一座理解与探索的桥梁。

本书的独特价值在于，它没有停留在对现象的描述或对未来的空泛畅想，而是系统性地构建了一条从生成式 AI 基础到智能体系统核心原理，再到复杂设计与伦理考量的完整路径。生成式 AI 模型，如 VAE、GAN 及各类自回归模型，赋予了机器内容创造的能力，这是智能体与世界交互、表达其“认知”的基础。但本书深刻指出，生成能力远非“智能”的全部。真正的突破在于如何将这些能力嵌入到一个具有自主性、反应性、甚至主动性的“智能体”架构之中，使其能够感知环境、进行推理、规划行动序列，并能通过反思和学习不断优化自身行为。这标志着 AI 的研究重心正从“模型中心”转向“智能体中心”，从追求单一任务的卓越性能转向构建能适应开放环境的综合性认知系统。

在翻译过程中，书中关于智能体“反思与内省”的论述尤其令我深思。当 AI 开始审视自己的推理过程，尝试进行元推理和自我解释时，我们触及的已不仅是技术难题，更是哲学层面关于意识与自我认知的古老议题。本书以实践为导向，将这类看似抽象的概念转化为具体的技术思路与设计模式，例如如何实现智能体的自我建模，如何让其从成功与失败的经验中学习。这不仅是为了提升智能体的性能，更是迈向可解释、可信赖 AI 的关键一步。一个能够向人类解释其决策逻辑的智能体，才是可能被真正纳入关键决策循环的合作伙伴。

此外，本书对多智能体协作的深入探讨极具前瞻性。单个智能体的能力再强，其效用也是有限的。未来的智能系统，更可能呈现为多种“异质”智能体动态协作的生态系统。“协调者-工作者-委派者”等设计模式的引入，为我们勾勒了这幅错综复杂又壮观的图景。在这些模式中，智能体之间通过通信、协商、竞争与合作，共同解决单个智能体无法应对的复杂问题。这不仅是分布式人工智能的延伸，

更蕴含着对人类社会协作模式的深刻隐喻，其成功实践将极大地推动自动化决策、资源优化等领域的发展。

当然，能力的提升必然伴随着责任的加重。本书并未回避随之而来的严峻挑战。在“构建生成式 AI 系统的信任”与“安全管理与伦理考量”等章节中，作者以大量篇幅讨论了透明度、偏见缓解、可控性以及伦理一致性等至关重要的问题。赋予 AI 系统以能动性，意味着我们必须确保这种能动性被约束在符合人类价值观和利益的轨道上。翻译这些内容时，我深感其分量之重。它提醒每一位从业者，我们手中的代码不仅是功能实现的工具，更是未来智能生态的基石，关乎安全、公平与信任。技术的狂飙突进必须与审慎的伦理框架和稳健的治理机制并行，否则我们精心构建的智能体可能会偏离我们预设的航向。

本书的实践性同样值得称道。大量的代码示例、操作练习和实施策略，确保了高屋建瓴的理论能够落地为切实可行的项目。无论你是致力于将最新模型应用于实际产品的开发者，还是负责制定技术战略的架构师或管理者，都能从中获得直接的启发和实用的工具。

翻译此书的过程，于我而言，是一次深入的学习与反思。它让我更清晰地看到，人工智能领域正在发生的，并非仅仅是算法的迭代或模型的扩大，而是一场范式的迁移——从创建孤立的、专门化的工具，转向培育具有上下文理解力、战略思维和协作精神的数字智能体。这条路充满荆棘，但也无比激动人心。

希望这本译作能够成功传递原著的精髓与洞见，帮助中文世界的读者更好地把握 Agentic AI 的核心脉络，共同参与到塑造未来智能系统的伟大进程中。让我们携手，以负责任的态度和创新的精神，一步步走进那个智能体与人类协同共进的新时代。书中难免存在疏漏与不足之处，恳请广大读者不吝指正。

茹炳晟

2025 年 9 月 22 日

这本书献给我已故的父亲，他的谆谆教诲指引我勇敢直面每一次挑战。也献给我的母亲、妻子和儿子，是你们的爱与支持成就了这本书。感谢我的导师和同事们，是你们在人工智能这个前沿领域为我照亮前路。能够向你们学习并一起走过这段旅程，我感到无比荣幸。

——安贾纳瓦·比斯瓦斯

我将这本书献给我的父母、儿子、妻子和所有支持我的人——你们的坚定支持和不断鼓励，成为我前行的动力。你们对我的信任，让每一个挑战都变成了机会，让梦想逐渐变成了现实。正因如此，旅程中的每一个成就都显得格外珍贵。

——里克·塔鲁克达尔

推荐序一

人工智能的发展始终由一种深切的愿景所驱动：不仅要创造能够完成预定任务的系统，还要让这些系统具备推理、适应和自主行动的能力。随着生成式 AI 和智能体系统的兴起，我们正处于一个变革的关键时刻。现在，人工智能系统能够反思自身经验、制定有效的战略，并与其他系统进行高效合作。本书正是在这个重要时刻问世，全面探讨了这一新兴领域及其深远的影响。

我在人工智能领域有几十年的从业经验，曾在微软研究院和风险投资支持的 AI 初创公司担任技术领导，深度参与了大规模 AI 产品的研发和部署，服务了数百万用户。此外，我在 IEEE 担任多个人工智能领域领导职务，也让我从另一个角度看到了符合伦理和负责任的 AI 发展有多么重要。

我最初是在 IEEE 全球 AI 项目的合作中结识了作者安贾纳瓦·比斯瓦斯和里克·塔鲁克达尔。他们在亚马逊云科技(AWS)公司担任重要技术职务，专注于 AI 和机器学习的工作，同时在学术界进行深入研究。安贾纳瓦和里克都是生成式 AI 领域的专家，他们拥有世界级的技术能力和深刻的洞察力，参与了全球范围内的前沿 AI 平台设计、开发和部署。他们对技术难题和伦理问题的深刻理解，使得他们非常适合撰写这本书。

本书分为三部分。

第 1 部分帮助读者打下了坚实的理论基础，清晰地介绍了生成式 AI 和智能体系统的定义及核心特点。无论你是刚接触这一领域的新手，还是已有一定经验的专家，都能从中学到从变分自编码器和生成对抗网络到复杂智能体架构的核心概念，帮助你更好地理解这些复杂的技术。

第 2 部分，重点讲解了开发具备自省能力、适应能力和协作决策能力的智能体所需的实践技术。其中的反思、内省、规划和多智能体协作的内容，提供了重要的方法和见解，帮助开发者设计出能够不断学习和进步的系統。

第 3 部分深入探讨了将生成式智能体系统应用到现实世界时可能带来的影响。通过强调透明性、可解释性、安全性和伦理治理，书中明确指出了与先进 AI 技术相关的责任。它提供了一个关键框架，既有助于建立公众对该技术的信任，也能确保这些强大工具与社会的价值观和伦理标准保持一致。

这本书不仅是一本技术参考书，更是塑造生成式 AI 和智能体未来的必备指南。它邀请研究人员、行业领袖和创新者积极参与，共同谱写人工智能的下一个

篇章，在这个过程中，我们既要负责地突破技术边界，也要让技术持续为社会发展赋能。

我全力推荐这本及时且富有洞察力的著作。我鼓励你深入研讨接下来的章节，积极投身于人工智能这一激动人心、蓬勃发展的前沿领域，成为其中的一员。

马修·R. 斯科特
Minset.ai 首席技术官
2025 年春

推荐序二

当 ChatGPT 在 2022 年 11 月首次发布时，在全球范围内引发轰动。这一时刻，大家仿佛都顿悟了——突然间，你感觉自己在与一个真人对话。回答不仅流畅，而且信息丰富且实用。从那时起，AI 技术发展迅速，数亿人开始使用 ChatGPT 和类似的系统作为写作助手，或者用它们来获取那些通过传统网络搜索难以找到的答案。

然而，当你让一个人工助手规划旅行时，他不仅会告诉你该怎么做，还会亲自为你预订航班和酒店。这正是智能体系统的精髓所在。像 ChatGPT 中使用的大语言模型，不仅可以用来生成文本，还可以用于执行任务——通过生成代码、调用 API 和与网络服务交互来完成任务。就像万维网从提供静态内容转变为支持互动式应用一样，智能体系统也将彻底改变人类与 AI 互动的方式。

不同于那些主要关注智能体系统理论的书籍，本书重点讨论了技术的落地实现。作者提供了实际的代码示例、应用案例，分步讲解如何将智能体工作流程整合到组织中以解决实际业务问题。本书既是易于理解的入门资料，也是内容全面的参考资料，适合软件开发人员、机器学习工程师、技术负责人、主管以及其他希望跟上智能体系统发展的决策者阅读。

使用过 ChatGPT 等 AI 助手的人都知道这类工具在写作和编程方面的价值，用户可以在采取行动前验证输出结果。然而，AI 有时会非常自信地输出错误答案，导致用户误信它们。人类在表达不确定时通常会加上“我不确定，但我认为……”这样的语句，而 AI 系统则缺乏内置的不确定性指标。我们期待的是，AI 智能体会采取正确的行动，并在有疑问时主动向人类请教。

本书深入探讨了信任、透明性、可解释性和可靠性等关键问题，同时讨论了伦理问题，如偏见和安全性。书中描述了如何实现智能体系统的个性化与情境感知能力，以及这类系统如何规划行动，并根据行动结果做出动态调整。

作为一名在人工智能领域耕耘数十载的专家，我相信本书将成为学生、研究人员和专业人士的宝贵资源，为我们在被智能系统深刻影响的时代指引前进的道路。我与里克·塔鲁克达尔的初次邂逅是在 IEEE 会议上，他在理论与实践之间

架起了桥梁，给我留下了深刻的印象。在《构建 Agentic AI 系统：打造能推理、可规划、自适应的 AI 智能体》一书中，他和安贾纳瓦·比斯瓦斯正是做到了这一点——为 AI 驱动的自动化未来提供了一个实用的指南。

亚历克斯·阿塞罗博士
美国国家工程院院士，IEEE 会士

前言

本书旨在为生成式 AI 和智能体系统提供理论基础与实践指导。生成式 AI 和智能体系统正掀起下一波 AI 技术的浪潮，它们正以前所未有的方式驱动自动化、创造力和决策能力革新。这些技术能让机器在自主推理和适应的同时，生成文本、图像，甚至战略规划，正在改变医疗、金融、机器人等多个行业。

本书首先介绍了生成式 AI，讲解了**变分自编码器、生成对抗网络 and 自回归模型等重要模型**。在探讨这些模型在内容创作、设计和科学研究中的应用的时，讨论了它们的局限性和挑战。

接下来，深入分析智能体系统的相关概念，如智能体的能动性、自主性和多智能体协作。我们将不同类型的智能体架构进行对比，包括慎思型、响应型和混合型，并探讨了多个智能体如何相互协作、互动，以达成共同目标。

打好理论基础后，本书将进入实践落地部分。书中将介绍智能体如何反思自己的推理过程、进行规划，并有效利用外部工具，具体涵盖元推理、自我解释、战略规划和多智能体协作等技术。此外，本书讨论了如何设计既智能又值得信赖的 AI 智能体，平衡自主性与可控性，并确保 AI 技术在伦理框架下负责任地发展。

最后，本书将探讨生成式智能体系统在多个领域的实际应用案例，涵盖了自然语言处理、机器人技术、决策支持和优化等。书中还将重点讨论信任、透明度、偏见缓解和 AI 安全性等问题，这些因素对确保 AI 系统的可靠性至关重要。

在整本书中，你将看到很多代码示例、实际操作练习和实施策略，帮助你将理论知识与实际应用结合起来。不论你是 AI 从业者、研究人员、工程师，还是技术领导者，本书都将为你提供构建能够推理、协作和自我进化的智能 AI 智能体所需的技能和知识。

让我们一起开始这段旅程，共同塑造智能系统的未来——一步步推动智能体不断进化。

目标读者

本书适合那些希望理解并构建具有自主性、适应性和智能性的 AI 驱动型智能体的 AI 从业人员、开发者、研究人员、工程师和技术管理者。不论你是希望

将生成式模型融入智能系统的开发者，还是探索高级智能体功能的 AI 架构师，本书都将为你提供扎实的理论基础和可直接上手的实施策略。

主要内容

第 1 章 生成式 AI 基础，介绍了生成式 AI 的核心概念、各种模型类型(包括变分自编码器、生成对抗网络 and 自回归模型)以及其在现实中的应用，讨论了偏见、局限性和伦理问题等挑战。

第 2 章 智能体系统的原理，定义了智能体系统，涵盖了智能体的能动性、自主性及其核心特征(如响应性、主动性和社交能力)等内容。同时，探讨了不同的智能体架构和多智能体协作。

第 3 章 智能体的基本组成部分，详细介绍了智能体的关键元素，包括知识表示、推理、学习机制、决策制定，以及生成式 AI 在增强智能体能力方面的作用。

第 4 章 智能体的反思与自省能力，探讨了智能体如何分析自己的推理过程，如何从经验中学习并通过元推理、自我解释和自我建模等技术来改进决策。

第 5 章 赋予智能体使用工具与进行规划的能力，讨论了智能体如何利用外部工具实现规划算法，并将工具使用与战略决策相结合，以提高效率和达成目标。

第 6 章 探索“协调者-工作者-委派者”设计模式，介绍了多智能体协作中的 CWD 模型，解释了智能体如何承担协调者、工作者或委派者等角色，以优化任务执行和资源分配。

第 7 章 高效的智能体系统设计技术，讲解了设计智能体的最佳实践，包括聚焦式指令、设定保护措施与约束条件、平衡自治与控制、确保透明度和可追溯性等。

第 8 章 构建生成式 AI 系统的信任，探讨了建立 AI 信任的技术，包括透明性、可解释性、处理不确定性与偏见，以及设计可靠且可解释的 AI 系统。

第 9 章 安全管理与伦理考量，讨论了生成式 AI 的风险与挑战，确保 AI 负责任发展的策略、伦理指南，以及 AI 部署中的隐私和安全问题。

第 10 章 常见用例与应用场景，展示了生成式 AI 在创意内容生成、对话式 AI、机器人技术和决策支持系统等领域的实际应用。

第 11 章 结论与未来展望，总结了本书的关键概念，探讨了生成式 AI 和智能体系统的最新趋势，讨论了通用人工智能，并展望了未来的挑战与机遇。

如何最大化本书的价值

如果具备以下条件，阅读本书时将更轻松：

- **熟悉 AI 和机器学习的相关概念**：虽然本书涵盖了基础原理，但如果你事先掌握 AI、机器学习、深度学习和 Python 编程等知识，将有助于理解更高级的内容。
- **具备实践能力**：尝试使用本书提供的代码示例和框架，进行生成式 AI 和智能体系统的构建。搭建本地或云端开发环境，将提升你的学习体验。
- **能够对人工智能伦理与安全性进行批判性思考**：在探索生成式 AI 和自主智能体时，思考信任、偏见以及负责任的 AI 设计的影响，旨在构建符合伦理规范的智能系统。

本书涵盖的软件和硬件	操作系统需求
Python、Jupyter Notebooks 和 CrewAI	Windows、macOS、Linux

下载示例代码文件

本书的代码包托管在 GitHub 上，网址为 <https://github.com/PacktPublishing/Building-Agentive-AI-Systems>，也可扫描右侧二维码下载。



本书遵循的约定

本书遵循若干文本约定，具体内容如下。

加粗：表示新术语、重要词汇或屏幕上显示的词语。例如，菜单或对话框中的词语会加粗显示。举个例子：“**个性化的入职计划**：根据已识别的目标和需求，创建一个量身定制的入职计划，概述实现设定目标的步骤、里程碑和时间表。”

提示或重要说明

会以这种形式出现。

关于作者

安贾纳瓦·比斯瓦斯是一位屡获殊荣的高级 AI 专家和解决方案架构师，拥有超过 17 年的行业经验。他专注于机器学习、生成式 AI、自然语言处理、深度学习、数据分析和云架构，与大型企业合作，在云端构建和扩展先进的 AI 系统。安贾纳瓦因其在应用 AI 领域的贡献而广受认可。他曾在多个科学期刊上发表研究成果，并积极参与 AI、ML 开源项目。他的职业荣誉包括英国计算机协会(BCS)、英国工程技术学会(IET)、印度电子与电信工程师学会(IETE)等学会的会士资格，还是 IEEE 高级会员。安贾纳瓦常以嘉宾身份参与公开演讲，曾在 IBM 和 Oracle 等行业巨头企业中担任重要职务。安贾纳瓦来自印度，目前与妻子和儿子居住在加利福尼亚州的圣地亚哥，在科技领域持续创新并激励他人。

我想向所有陪伴我走过生成式 AI 之旅的人表达由衷的感谢。感谢曾经与我共同工作的同事们，你们挑战了我的思维，分享了这个快速发展的领域的活力——你们的合作是无价的。感谢我的导师们，你们看到了我想法中的潜力，并以智慧和耐心引导我——你们对我的信任使这本书得以问世。

感谢我那了不起的儿子，年仅 11 岁的你，总能用天马行空的想象力提醒我，未来属于那些敢于梦想的人。你的问题迫使我将复杂的内容简化为清晰、易懂的逻辑——这或许是理解生成式 AI 最宝贵的技能。

感谢我的妻子，每一个深夜和突破性时刻里始终陪伴着我的坚定伴侣——你的耐心是这些想法得以实现的无声基石。

感谢我的母亲，她的智慧超越了技术，她的鼓励从未停歇。还要感谢我的父亲，虽然他已不在身边，但我所选择迎接的每一个挑战，都离不开他的影响。我知道，你一定会带着骄傲翻阅这本书的每一页。

正如艾达·洛夫莱斯曾经说过的：“分析机编织代数图案，就像织布机编织花卉和叶子图案。”今天的生成式 AI 延续了人类创造力与计算可能性之间的美妙互动——这段“互动之舞”，我有幸记录下来。

致所有笃信科技当以升华人性而非削弱人性为使命的读者——这本书献给你们。

—— 安贾纳瓦·比斯瓦斯

里克·塔鲁克达尔是亚马逊公司在生成式人工智能领域的富有远见的技术领袖，拥有超过二十年的全球 AI、云计算和产品管理经验。作为 AI 驱动转型的先锋，他领导了大规模的现代化项目，推动了企业增长，并影响了全球数百万人的生活。他主导将多项获奖的 AI、ML 技术产品化，这些技术现在已大规模部署在《财富》500 强企业中，塑造了现实世界中的 AI 应用场景。

作为 AI 研究领域的知名人物，里克在生成式 AI、多模态、自然语言处理和计算机视觉等方面的研究成果被广泛引用和参考。作为 IEEE 高级会员、多个行业论坛的主席和小组成员，他为 CTSoc 行业论坛、网络智能委员会(Network Intelligence Committee, NIC)等全球委员会提供专业建议，参与制定行业标准，并推动人工智能未来的角色定位。他常受邀在世界科技峰会(World Technology Summit)、IEEE 荣誉学会(IEEE HKN)、国际消费电子大会(ICCE)、剑桥能源周(CERAWeek)、阿布扎比国际石油博览会(ADIPEC)等顶级会议上分享创新成果，将前沿研究与现实世界中的 AI 应用结合起来，加速行业创新。

他凭借扎实的计算机科学背景，共同主持 IEEE NIC，助力年轻专业人才成长。作为作者和思想领袖，他不断推动 AI 技术边界的突破，激励着未来的创新者。里克与家人现定居于加利福尼亚。

正如玛丽安·赖特·艾德elman所说：“教育是为了改善他人的生活，并使你的社区和世界变得更加美好。”

我由衷地感谢所有在这段旅程中支持、指导和激励过我的人。

感谢我的家人——我的父母、儿子和妻子——你们对我的坚定信念是我力量的源泉。

感谢我的儿子，他对“爸爸在做什么”的无尽好奇，每天都在提醒我寻找探索的意义。你的惊奇和问题在我最需要时带给我创意，而你的想象力让我看到了这项工作能解锁的无限可能。你教会了我，灵感常常来自一个相信你能解释宇宙的人最单纯的提问。

感谢我的父母，你们始终如一的支持、对我的信心和对理解这个快速发展的世界的渴望提醒我，知识是用来分享的，而不是储存的。你们对终身学习的信念是我的道德准则，你们对我工作的认可始终是我最重要的动力源泉。

感谢我的妻子，你沉稳的力量、持久的耐心与持续的激励，不仅使这段征程成为可能，更赋予其深刻意义。无数长夜相伴的身影、克服重重障碍时的坚定信念，以及那份抚慰心灵的笃定，最终指引我完成了所开创的事业。你始终是我人生的锚点与灵感的源泉。

感谢人工智能研究社区的各位同仁，你们对创新的不懈追求始终激励着我，也让我心怀谦卑。过去这些年，我撰写过研究论文、技术文章与博客，而创作人

生中的第一本书，无疑是一次极具变革意义的经历：它促使我必须将复杂的技术概念转化为清晰的表达，将抽象的想法转化为能产生实际价值的内容。

感谢我的朋友和支持者，在每一次挑战中，有你们相伴，那些阻碍才得以变成前行路上的里程碑。无论是一句温暖的鼓励、一次思想的交流，还是在最需要的时候默默出现，你们都让我明白：没有哪一段旅程是真正孤独的。

致每一位陪伴我的人。是你们与我并肩同行，坚信我的愿景，也一同书写了这本书里的每一个字。

——里克·塔鲁克达尔

关于评审员

曼吉特·查克拉博蒂是一位资深的技术领袖，在金融服务领域推动数字化转型方面积累了丰富的经验。作为亚马逊网络服务的企业解决方案架构师，他带领团队负责遗留系统的现代化改造，并为全球最大金融机构设计创新的云原生解决方案。曼吉特在商业和技术架构方面有着扎实的经验，他擅长通过数据分析提供切实可行的见解。他的专业领域包括云迁移、大型主机现代化、系统集成、混合部署、数据分析和商业智能等。作为备受欢迎的演讲者，曼吉特曾在多个知名技术论坛上发表演讲。他还是一位资深的作者，为多个技术出版物贡献了自己的见解，积极与技术社区分享经验。此外，曼吉特也是一位优秀的审稿人，定期对行业投稿和学术研究论文进行评审，极大地推动了技术进步和学术交流。他目前居住在美国佛罗里达州。

马丁·亚内夫是一位非常出色的软件工程师，拥有近十年的行业经验，涉及领域包括航空航天和医疗技术。在马丁卓越的职业生涯中，他在为关键领域(如空中交通管制和色谱分析系统)开发和集成先进软件解决方案方面取得了卓越成就。作为菲奇堡州立大学的计算机科学教授，马丁深入了解 OpenAI API 的各项技术，并在构建、训练和微调 AI 系统方面表现出了极高的水平。马丁还是一位广受认可的作者，致力于帮助他人解决 AI 开发中的复杂问题。凭借丰富的经验和多领域的技能储备，马丁始终在推动软件工程领域的创新和变革。

关于译者

茹炳晟，腾讯 Tech Lead，腾讯研究院特约研究员，腾讯集团技术委员会委员，中国计算机学会 TF 研发效能 SIG 主席，腾讯云架构师技术同盟入会成长主席，中国通信标准化协会 TC608 云计算标准和开源推进委员会云上软件工程工作组副组长，中国商业联合会互联网应用技术委员会智库专家，“软件研发效能度量规范”团体标准核心编写专家，复旦大学 CodeWisdom 团队成员，国内外各大技术峰会的联席主席、出品人和 Keynote 演讲嘉宾，公众号“茹炳晟聊软件开发”主理人，二十余本计算机专业图书的作者和译者。

殷海英，计算机科学技术博士、硕士学位，工商管理硕士(MBA)学位，联合国教科文组织认证教师；在美国多所大学的医学院研究和进修人体解剖学、心理学、人类生理学、医学神经科学、公共卫生基础等；已出版(含翻译)三十余部专业著作，内容涵盖 MySQL、AWS、Airflow、数据科学、Python、Kubernetes、机器学习、深度学习、LLM、AI Agent 及图算法等众多前沿技术领域；在 AI 数据治理、智能计算中心效能优化、分布式调度系统、大模型微调与部署、知识图谱构建等关键技术方向拥有多项专利及核心项目成果。

目 录

第 1 部分：生成式 AI 和智能体系统的基础

第 1 章 生成式 AI 基础	3
1.1 生成式 AI 简介	3
1.2 生成式 AI 模型的类型	4
1.2.1 变分自编码器	5
1.2.2 生成对抗网络	6
1.2.3 自回归模型和 Transformer 架构	7
1.2.4 基于 LLM 的 AI 智能体	9
1.3 生成式 AI 的应用	14
1.4 生成式 AI 面临的挑战与局限性	17
1.4.1 数据质量与偏见	17
1.4.2 数据隐私	18
1.4.3 计算资源	18
1.4.4 伦理和社会影响	19
1.4.5 泛化能力与创造力	19
本章小结	20
习题	20
习题答案	20
扩展阅读	20
参考文献	21
第 2 章 智能体系统的原理	23
2.1 技术要求	23
2.2 理解自我管理、能动性和自主性	23
2.2.1 自我管理	24
2.2.2 能动性	25
2.2.3 自主性	25
2.2.4 智能体的能动性和自主性的示例	26
2.3 回顾智能体及其特征	29
2.4 探讨智能体系统的架构	30
2.4.1 慎思型架构	30
2.4.2 响应型架构	32
2.4.3 混合架构	33
2.5 理解多智能体系统	34
2.5.1 多智能体系统的定义和特征	35
2.5.2 多智能体系统中的互动机制	36
本章小结	41
习题	41
习题答案	42

第 3 章 智能体的基本组成部分	43
3.1 技术要求	43
3.2 智能体中的知识表示	44
3.2.1 语义网络	44
3.2.2 框架	45
3.2.3 基于逻辑的表示	46
3.3 智能体中的推理	48
3.3.1 演绎推理	48
3.3.2 归纳推理	49
3.3.3 溯因推理	50
3.4 自适应智能体的学习机制	52
3.5 智能体系统中的决策与规划	53
3.5.1 效用函数	53
3.5.2 规划算法	55
3.6 利用生成式人工智能增强智能体能力	58
本章小结	61
习题	61
习题答案	61
第 2 部分：设计与实现基于生成式人工智能的智能体	
第 4 章 智能体的反思与自省能力	65
4.1 技术要求	66
4.2 智能体中反思的重要性	66
4.2.1 增强决策能力	66
4.2.2 提高适应性	67
4.2.3 增加伦理考量	67
4.2.4 优化人机交互体验	68
4.3 智能体的自省能力	69
4.4 实现反思能力	70
4.4.1 传统推理	70
4.4.2 元推理	70
4.4.3 自我解释	78
4.4.4 自我建模	81
4.5 应用场景和示例	83
4.5.1 客户服务聊天机器人	83
4.5.2 个性化营销智能体	84
4.5.3 金融交易系统	85
4.5.4 预测智能体	86
4.5.5 电子商务中的定价策略	87
本章小结	88
习题	88
习题答案	89
第 5 章 赋予智能体使用工具与进行规划的能力	91
5.1 技术要求	92
5.2 理解智能体中工具使用的概念	92
5.2.1 工具调用与函数调用	93
5.2.2 为智能体定义工具	94
5.2.3 工具的类型	96
5.2.4 工具在智能体系统中的重要性	99
5.3 智能体的规划算法	99
5.3.1 实用性有限的规划算法	100
5.3.2 具备一定实用性的规划算法——FF	101

5.3.3 实用性最强的规划 算法·····	102	5.5.1 CrewAI 示例·····	110
5.4 工具使用与规划的整合·····	108	5.5.2 AutoGen 示例·····	112
5.4.1 对工具进行推理·····	108	5.5.3 LangGraph 示例·····	113
5.4.2 制订工具使用计划·····	109	本章小结·····	115
5.5 探索实际代码实现·····	110	习题·····	115
		习题答案·····	115
第 6 章 探索“协调者-工作者-委派者”设计模式	117		
6.1 技术要求·····	117	6.4.3 协商与冲突解决·····	130
6.2 理解 CWD 模型·····	118	6.4.4 知识共享·····	130
6.2.1 CWD 模型的关键原则·····	119	6.5 在生成式 AI 系统中实施 CWD 模式·····	131
6.2.2 CWD 模型在旅行 系统中的应用·····	120	6.5.1 系统提示词与智能体 行为·····	132
6.3 为智能体设计并分配 角色·····	122	6.5.2 指令格式化·····	132
6.4 智能体之间的沟通与 协作·····	129	6.5.3 交互模式·····	133
6.4.1 沟通·····	129	本章小结·····	134
6.4.2 协调机制·····	130	习题·····	134
		习题答案·····	135
第 7 章 高效的智能体系统设计技术	137		
7.1 技术要求·····	137	7.4.1 短期记忆·····	146
7.2 聚焦式系统提示和智能体 指令·····	138	7.4.2 长期记忆·····	147
7.2.1 定义目标·····	138	7.4.3 情景记忆·····	148
7.2.2 任务规范·····	139	7.4.4 上下文管理·····	149
7.2.3 上下文感知·····	141	7.4.5 决策整合·····	150
7.3 状态空间与环境建模·····	142	7.5 智能体 workflows 中的顺序 处理和并行处理·····	151
7.3.1 状态空间表示·····	142	7.5.1 顺序处理·····	151
7.3.2 环境建模·····	143	7.5.2 并行处理·····	152
7.3.3 集成与交互模式·····	144	7.5.3 工作流优化·····	152
7.3.4 监控与适应·····	146	本章小结·····	154
7.4 智能体记忆架构与上下文 管理·····	146	习题·····	155
		习题答案·····	155

第 3 部分：信任、安全、伦理和应用

第 8 章 构建生成式 AI 系统的信任	159
8.1 技术要求	160
8.2 AI 中信任的重要性	160
8.3 建立信任的技术	161
8.3.1 透明度和可解释性	161
8.3.2 不确定性和偏见的处理	165
8.3.3 高效的输出沟通	165
8.3.4 用户控制权与同意权	166
8.3.5 伦理化开发与责任	167
8.4 透明度和可解释性的实现	168
8.5 不确定性和偏见的应对	169
本章小结	170
习题	171
习题答案	171
第 9 章 安全管理与伦理考量	173
9.1 了解潜在风险和挑战	173
9.1.1 对抗攻击	174
9.1.2 偏见和歧视	175
9.1.3 错误信息和幻觉	176
9.1.4 数据隐私侵犯	176
9.1.5 知识产权风险	178
9.2 确保安全且负责任的 AI	179
9.3 探索伦理准则与框架	182
9.3.1 以人为本的设计	182
9.3.2 可追溯性与责任	182
9.3.3 隐私与数据保护	182
9.3.4 多元化利益相关者的参与	183
9.4 解决隐私与安全问题	183
本章小结	184
习题	185
习题答案	185
第 10 章 常见用例与应用场景	187
10.1 创意和艺术应用	188
10.1.1 创意和艺术智能体的演进	188
10.1.2 实际应用	188
10.2 自然语言处理和对话式智能体	190
10.2.1 语言智能体的演进	190
10.2.2 实际应用	191
10.3 机器人技术和自主系统	193
10.3.1 机器人智能体的演进	193
10.3.2 实际应用	193
10.4 决策支持与优化	195
10.4.1 决策支持智能体的演进	196
10.4.2 实际应用	196
本章小结	199
习题	199
习题答案	200

第 11 章 结论与未来展望	201
11.1 核心要点总回顾	202
11.2 前沿趋势与新动向	202
11.2.1 多模态智能：整合 多样化输入	203
11.2.2 高级语言理解	203
11.2.3 经验式学习：强化 学习创新	203
11.2.4 对各行各业的实际 影响	204
11.3 通用人工智能	204
11.3.1 AGI 到底牛在哪里	204
11.3.2 核心挑战	205
11.4 挑战与机遇	206
本章小结	207

第1部分： 生成式AI和智能体系统的基础

本部分介绍了生成式 AI 和智能体系统的基本概念，为读者理解和开发智能的、具有自我决策能力的智能体打下坚实的理论基础。

本部分包括以下章节：

- 第 1 章 生成式 AI 基础
- 第 2 章 智能体系统的原理
- 第 3 章 智能体的基本组成部分

第 1 章

生成式 AI 基础

生成式 AI 迅速成为人工智能和机器学习领域的关键技术，正在彻底改变各行各业中的创作过程和问题解决方式，并推动基于智能体的系统朝着更高的自主性迈进。本章将介绍生成式 AI 的基础内容。首先，解释生成式 AI 是什么，并简要回顾其发展历程。接着，介绍几种常见的生成模型，包括变分自编码器、生成对抗网络、自回归模型和 Transformer 模型。最后，探讨生成式 AI 的实际应用，并讨论一些与生成式 AI 相关的局限性和挑战。

本章主要介绍以下内容，为进一步探索由生成式 AI 驱动的智能体奠定基础。

- 生成式 AI 简介。
- 生成式 AI 模型的类型。
- 生成式 AI 的应用。
- 生成式 AI 面临的挑战与局限性。

通过本章的学习，你将全面理解生成式 AI，包括其基本概念、应用场景以及当前面临的挑战。你还将了解该技术的潜力和局限，特别是在推动智能体和智能体系统发展过程中的关键作用。

1.1 生成式 AI 简介

生成式 AI 指的是一类能够生成各种形式内容的 AI 技术，这些内容包括但不限于文本、图像、音频和视频。这类 AI 系统可以根据其训练数据和输入参数生成新的内容，通常这些输入参数是文本提示词，也可以是图像等其他形式。近年来，生成式 AI 备受关注，正是因为这种技术驱动的新型用户界面操作简便，能够轻松地在几秒钟内生成高质量的文本、图像和视频。

简单来说，生成式 AI 就是通过学习已有数据的规律、结构和分布，生成与

这些数据的特征相似的新内容。也就是说，生成式 AI 通过学习输入数据背后隐藏的模式、结构和分布规律，在模型内部形成一套生成机制，以类似的方式创造新数据。例如，如果训练数据是人脸图像，生成式 AI 模型就可以生成一些并不存在的，但非常逼真的全新人脸图像。这些模型的工作原理是学习数据的概率分布，并从中随机选取数据点来生成新的实例。这与判别式模型不同，后者的目标是区分不同类别的数据。例如，判别式模型可以被训练用于区分猫和狗的图片，而生成式模型则被训练用于创造全新的猫或狗的图片。

生成式 AI 的概念最早可以追溯到机器学习的初期，实际上，它诞生于机器学习刚刚起步的时候。然而，直到最近，随着神经网络技术的发展和计算能力的大幅提升，这一领域才开始受到广泛关注。在早期的生成式模型中，如高斯混合模型和隐马尔可夫模型，使用的是一些较简单的统计方法。而对于复杂数据分布的建模方法，之前的技术虽然兼容性较好，但随着深度学习技术的崛起，这些方法已经逐渐被取代。因此，深度学习的出现是生成式 AI 领域的一个重大突破。比如，21 世纪第二个十年初期出现的变分自编码器就能够利用深度神经网络建模更加复杂的数据分布。而与此同时，生成对抗网络也应运而生，它首次提出了通过博弈论原理，让两个神经网络合作生成数据的方式。这些技术突破为生成真实、高质量的数据提供了更高的可能性。

生成式 AI 有望为许多行业带来创新，比如医疗、金融、教育、旅游、制造业等。在创意产业中，生成式 AI 帮助艺术家和设计师突破传统思维，激发新的创作灵感。在医疗领域，生成式 AI 为药物研发的路径探索和个性化治疗提供了新的可能。在商业领域，生成式 AI 通过提供个性化的内容和体验来提高与客户互动的效率，增强企业与客户互动的效果。另外，生成式 AI 的广泛应用也引发了对创造力和原创性的思考：当机器能够批量生成高度相似的内容时，人们不禁要思考创造力究竟是什么、谁才是真正的艺术家，以及 AI 创作的伦理界限应如何界定。

了解了生成式 AI 的基本概念和发展历程后，接下来将深入探讨生成式 AI 模型的不同类型。

1.2 生成式 AI 模型的类型

生成式 AI 是人工智能领域中极为活跃的一个分支，旨在通过学习现有数据集的规律，生成新的合成数据，目标是产生与训练数据在统计特性和结构上相似的输出。几种最具代表性的生成式 AI 模型包括变分自编码器、生成对抗网络、自回归模型和 Transformer 架构。

1.2.1 变分自编码器

变分自编码器(VAE)是最受欢迎的生成式模型之一。VAE 的核心思想是学习数据与潜在空间之间的概率映射关系。简单来说,既学习如何将真实数据转换为简化的表示形式(例如压缩形式),又学习如何将其还原为看起来像真实数据的形式。VAE 的设计旨在确保生成的数据与真实数据的似然性足够高,同时保持一个结构良好的潜在空间,以便生成与训练数据相似的新数据样本。VAE 的几种常见类型如下。

- **变分自编码器**: 基础模型,用于压缩和重构数据。它提供了一个框架,可以从已学习的潜在空间中生成新样本。VAE 是一个生成式模型,它在学习压缩和重构数据的同时,学习潜在空间的概率分布。简单来说,VAE 就像一个聪明的算法,能够理解和重现数据。可以将它比作一位技艺精湛的艺术师,能够将一幅细节丰富的画作压缩成简单的草图(编码),然后根据这幅草图还原完整的画作(解码)。VAE 之所以特别,是因为它不仅仅是复制数据,还能学习数据的本质,从而创造出新的、相似的数据。这就像艺术家不仅学会复制特定的画作,还能理解画作的风格,以至于可以创作出全新的符合该风格的原创作品。在实际应用中,VAE 被应用于许多领域。例如,在药物研发中,VAE 被用来生成新的分子结构。阿斯利康的研究人员使用 VAE 探索化学空间,并提出具有理想属性的新型药物分子,这有可能加速药物开发过程^[1]。
- **贝塔变分自编码器(Beta-VAE)**: VAE 的扩展版本,引入了一个超参数,用来明确控制数据重构质量和潜在空间解耦之间的权衡关系。该模型旨在创造更具可解释性的解耦表示。简单来说,Beta-VAE 就像 VAE 的一个灵活升级版,它允许研究人员调整模型在重建精确细节和理解数据底层特征之间的侧重点。这就像教导一名艺术生,使其不仅学会复制画作,还要学会理解并分离画作的元素,如颜色、形状和风格。Beta-VAE 所具有的解耦特征的能力,使其特别适合计算机视觉和机器人学等领域。例如,研究人员使用 Beta-VAE 来教机器人更好地理解物体。通过学习分离大小、颜色和位置等特征,机器人可以更轻松地识别和操作不同的物体,从而在各种任务中具备更强的适应性和更高的效率^[2]。
- **条件变分自编码器(CVAE)**: VAE 的一种变体,通过额外的信息(通常是类别标签)来控制生成过程,不仅生成与训练集相似的数据,还能生成符合要求的特定类别的样本。可以把 CVAE 看作 VAE 的一个更加受控的版本,就像一位艺术家可以根据要求绘画不同风格的画作。这个艺术家不仅能随意创作新的画作,还能根据“画一幅风景画”或“画一幅肖像画”等要求创作特定类别的艺术作品。这种额外的可控性使得 CVAE 在实际应

用中非常有用。例如，在计算机游戏领域，CVAE 被用于游戏开发和程序化内容生成，如角色设计、关卡布局、音乐和音效等。通过提供不同的条件，如“创建一个森林关卡”或“创建一个沙漠关卡”，CVAE 可以生成各种各样的游戏场景，从而节省设计师的时间，并通过更丰富和有趣的游戏世界提升玩家的体验^[3]。

1.2.2 生成对抗网络

生成对抗网络(GAN)基本上由两个神经网络组成：生成器和判别器。生成器负责生成合成数据样本；而另一个经过训练的神经网络则应该能够区分真实样本和生成的样本。在训练这两个网络时，它们是以对抗的方式共同训练的：生成器试图欺骗判别器，而判别器则试图正确区分真实数据和伪造数据。在这场“较量”中，生成器变得越来越擅长伪造数据。以下是一些不同类型的 GAN。

- **生成对抗网络**：基础模型，其中生成器和判别器以对抗的方式进行训练，也是大多数生成式模型创新的基础模型。如前所述，你可以把 GAN 看作两位玩家之间的博弈——一位伪造者(生成器)试图制造假币，而一位侦探(判别器)试图识别假币。在较量过程中，它们都变得越来越擅长各自的工作，这意味着伪造者在制造假币方面越来越出色。这种巧妙的设计使得 GAN 能够生成极其逼真的伪造数据，如图像或声音^[4]。
- **深度卷积 GAN(DCGAN)**：这是对基础 GAN 模型的一个改进，使用了深度卷积神经网络；目前，它是生成高质量图像的最佳架构之一。可以把 DCGAN 看作一位比 GAN 技艺更精湛的艺术家的升级版。就像从简单的素描工具升级到一个完整的数字艺术工作室，使得可以创作出细节更丰富、真实感更强的图像。DCGAN 特别擅长理解和重现图像中的复杂模式^[5]。
- **Wasserstein GAN(WGAN)**：该模型提出了一个全新的损失函数，使用了 Wasserstein 距离。它消除了训练 GAN 时存在的一些问题，从而使训练更加稳定，并提高了样本质量，也能更精准地衡量两个概率分布之间的距离。WGAN 相当于为 GAN 中的艺术家(生成器)和评论家(判别器)提供了一种更好地沟通和评估彼此工作的方式。它们不再只是简单地说“好”或“坏”，而是可以给出更细致的反馈，如“你越来越接近了”或“你偏离方向太远了”。这种反馈机制带来了更稳定的改进和更高质量的结果。在实际应用中，WGAN 已被用于医学成像，生成合成医学图像供训练使用。这有助于创建更大、更具多样性的数据集，用于训练诊断 AI 系统，提高其从扫描图像和 X 光片中识别疾病的准确性^[6]。
- **StyleGAN**：该模型能够生成高质量的真实图像，尤其擅长将风格和内容分开处理。后续推出的 StyleGAN2 和 StyleGAN3 仍然旨在提高图像的真实感和细节表现。可以把 StyleGAN 看作一位高级数字艺术家，不仅能够创作

出逼真的图像，还能混合和匹配不同的风格与内容，就像有一位画家可以将凡·高的风格应用于现代城市景观的创作中。这种灵活性赋予 StyleGAN 较高的通用性^[7]。

1.2.3 自回归模型和 Transformer 架构

自回归模型一次生成一个数据点，每个数据点都依赖于前一个数据点。令人惊讶的是，这种方式在数据序列或结构具有关联性的任务中非常有用，例如文本生成和图像生成。Transformer 架构源自“Attention Is All You Need”论文^[8]，是一种彻底改变许多序列数据处理任务(特别是自然语言处理领域的任务)的模型架构。其关键组件如图 1.1 所示。

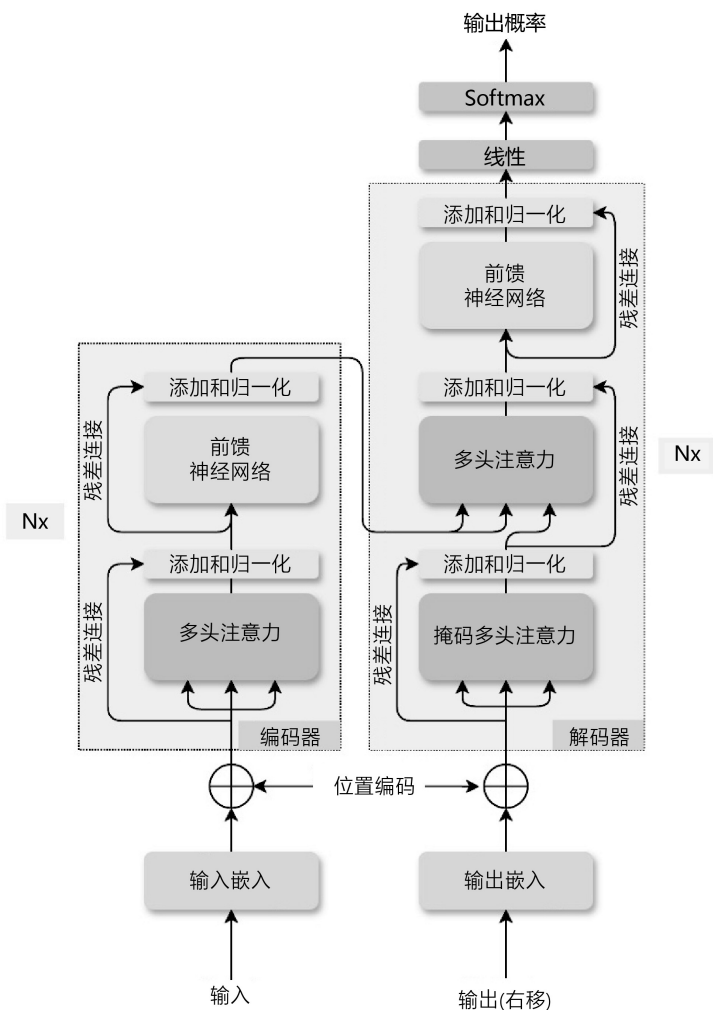


图 1.1 Transformer 架构的关键组件

这些关键组件如下。

- **自注意力机制**：一种计算技术，使模型在处理每个元素时，动态地关注输入的不同部分。
- **多头注意力**：多个注意力机制并行运行，使模型能够同时关注输入的不同方面。
- **位置编码**：为序列中每个元素添加位置信息。
- **前馈神经网络**：处理注意力层的输出。
- **层归一化与残差连接**：增强训练的稳定性 and 优化信息传递效率。

Transformer 可以在自回归和非自回归配置下使用，使其能够适应多种任务，以下是一些示例。

- **像素卷积神经网络(PixelCNN)**：自回归地逐像素生成图像，每个像素都依赖之前生成的像素。该模型在生成高质量、细节丰富的图像时表现良好^[9]。
- **PixelSNAIL**：是对 PixelCNN 的改进，加入了注意力机制，以捕捉像素之间的强依赖关系，从而提高图像的质量^[10]。
- **生成预训练 Transformer(GPT)**：是一类专注于文本生成的模型，即预测下一个应该出现的词汇，包括 GPT-2、GPT-3 和 GPT-4 等，它们在生成连贯且上下文相关的文本方面取得了巨大进步^[11-12]。
- **BERT(双向编码器表示的 Transformer)**：与 GPT 不同，BERT 旨在从文本的两个方向理解上下文。它使用 Transformer 的编码器部分，通常用于需要一次性理解全部输入内容的任务，而不是自回归生成文本^[13]。
- **文本到文本迁移 Transformer(Text-To-Text Transfer Transformer, T5)**：该模型将所有自然语言处理(NLP)任务都框定为文本到文本的形式。它使用完整的 Transformer 架构(包括编码器和解码器)，可以处理各种文本生成任务^[14]。

这些模型的表现会根据任务的不同有所变化，但它们展示了 Transformer 架构可以同时适应自回归任务(如 GPT)和非自回归任务(如 BERT)的适配能力，并展现出其在处理序列数据方面的强大能力。不同的生成式模型有各自的优势。这些模型的应用范围覆盖生成逼真的图像、流畅的文本和完全创新的数据样本等。基于 GPT 等模型的成功，研究人员将这些架构进行了扩展，开发出了大语言模型(LLM)。这些模型通常是在大量公开的网络文本和其他可免费获取的文本数据上进行训练的，展示了它们在理解和生成类似人类的文本方面的出色能力，适用于各种应用，展现出其卓越的适应能力和先进的语言处理能力。常见的大语言模型包括 GPT-3、GPT-4、PaLM 和 BERT-large。这些模型不断突破自然语言处理领域的边界，在问答、文本总结、创意写作、代码生成等任务中表现优异。以下是常见的大语言模型类型。

- **自回归 LLM**: 这种语言模型通常能够逐字生成文本, 即一次生成一个词元, 常用于文本生成、文本补全和创意写作等任务。常见的自回归 LLM 包括 GPT 系列(GPT-3、GPT-4)和 PaLM。
- **仅编码器 LLM**: 这些模型专注于自然语言理解(NLU)任务, 即分析和理解输入的文本, 而不是生成新文本。像 BERT 及其变种(如 RoBERTa)这样的仅编码器模型, 着重于理解文本的上下文和含义。它们能够同时处理整个输入序列, 从而更好地捕捉上下文信息, 帮助完成文本分类、命名实体识别(NER)、情感分析等任务。
- **编码器-解码器 LLM**: 这种模型融合了编码器和解码器, 不仅能够理解输入文本的上下文, 还可以生成相应的文本输出。例如, T5 模型就是一个典型的编码器-解码器 LLM, 它能完成语言翻译、摘要生成和问答任务。
- **多模态 LLM**: 多模态模型能够处理的不仅限于文本, 还包括图像、视频、音频等其他数据形式。多模态 LLM 可以处理并生成多种形式的内容, 包括文本、图像、音频、视频, 甚至它们的组合。随着计算能力的提升, 训练多模态 LLM 已经变得可能。这类模型包括 DALL-E、稳态扩散模型、Flamingo、GPT-4 和 LLaVa 等。
- **指令微调 LLM**: LLM 通常是在大规模公开文本数据上训练的, 但有时候这些模型可能无法高效处理一些它们未接触过的数据, 从而无法准确理解用户的意图。例如, 某些专用设备的使用手册可能包含模型在训练时没有接触过的专业术语。在这种情况下, 可以通过领域特定的数据对模型进行微调, 使其能够按照特定指令执行任务。这种微调可以让模型不仅在原有任务上表现优秀, 还能在特定领域任务中发挥作用。比如, InstructGPT 就是通过对 GPT-3 进行指令微调来实现的。这类模型常用于聊天机器人、虚拟助手和其他特定任务应用中。
- **领域特定 LLM**: 如前所述, 虽然 LLM 通常是在大量公开的、广泛可用的网络数据上训练的, 但在特定领域的应用中, 它们的表现可能不尽如人意。虽然指令微调 LLM 是一种让这些通用模型适应领域特定任务的方式, 但研究人员也创新性地从零开始训练领域特定的 LLM。这意味着这类模型是基于高度专业化的领域知识数据进行预训练的。例如, BioBERT 模型是在专业化的医学数据上训练的, 而 LegalBERT 则是在法律文档上训练的。这类模型在医学、法律、金融等特定领域的专业任务中非常有用。

1.2.4 基于 LLM 的 AI 智能体

前文所讨论的各类大语言模型均有其独特优势与不足, 且各自适配不同的应用场景。本书将聚焦于深入剖析 LLM 智能体的能力, 并探讨如何利用这些能力

构建先进的智能系统。需要重点说明的是，LLM 智能体并不能简单归入某一个单一类别，因为它本质上是 LLM 技术的高级应用形式，而非某一特定类型的模型架构。通常而言，智能体以指令微调 LLM 为基础构建，后续章节还将介绍，它往往会融合其他额外组件与功能模块，以实现更复杂的任务。

我们可以通过一个简单案例来直观展示基于 LLM 的智能体系统具备的能力。某旅游网站推出了一款航班预订助手聊天机器人，这是一个基于 LLM 构建的对话式交互界面，能够通过类人对话帮助用户预订机票。在对话过程中，该聊天机器人可在后台执行一系列任务。航班预订助手为一位用户预订机票的示例对话流程如下。

	你好，我想预订一张机票。
	你好！我很乐意帮你预订机票。首先，请告诉我您的出发城市和目的地好吗？
	我想从圣地亚哥去旧金山国际机场。
	太好了，感谢您提供这些信息。您要从圣地亚哥飞往旧金山国际机场。为了帮您选择最佳的航班，我还需要一些更详细的信息。 1. 您的首选出行日期是哪一天？ 2. 是单程还是往返？ 3. 有多少位乘客？
	我想下周五去，周日回来。只有我一个人旅行。
	感谢您提供更多信息。我来总结一下目前的信息。 • 出发地和目的地：圣地亚哥(SAN)至旧金山(SFO) • 出发日期：下周五，2024 年 8 月 23 日 • 返程日期：下周日，2024 年 8 月 25 日 • 乘客：1 位成人 在搜索航班之前，我还有几个问题。 1. 您对航班时间有偏好吗(上午、下午或晚上)？ 2. 您有偏好的航空公司吗，还是所有航空公司都可以？ 3. 您偏好哪种舱位等级(经济舱、高级经济舱、商务舱或头等舱)？
	我偏好早上的航班。没有航空公司偏好。经济舱就可以。
	感谢您提供这些信息。我将为您查找所有航空公司经济舱的早班机。请稍等，我为您寻找最佳选择。(正在搜索航班，请稍候……)
	我为您的旅程找到了几个选择。以下是最佳的早班机。 去程(圣地亚哥到旧金山)，下周五。 方案 1，美国联合航空 UA1234：上午 7: 15 出发，上午 8: 45 到达 方案 2，阿拉斯加航空 AS5678：上午 8: 30 出发，上午 10: 00 到达

续表

	返程(旧金山到圣地亚哥), 下周日。 方案 1, 达美航空 DL9876: 上午 9: 00 出发, 上午 10: 30 到达 方案 2, 西南航空 WN5432: 上午 10: 15 出发, 上午 11: 45 到达 往返总价(含税费)如下。 • 方案 1 (美国联合航空/达美航空): \$325 • 方案 2 (阿拉斯加航空/西南航空): \$310 您偏好哪个方案? 或者您需要我为您寻找其他航班吗?
	第二个方案看起来不错。可以帮我预订吗?
	当然! 我很乐意为您预订第二个方案。 去程: 阿拉斯加航空 AS5678。 • 日期: 2024 年 8 月 23 日 • 上午 8: 30 从圣地亚哥出发, 上午 10: 00 到达旧金山国际机场
	返程: 西南航空 WN5432。 • 日期: 2024 年 8 月 25 日 • 上午 10: 15 从旧金山国际机场出发, 上午 11: 45 到达圣地亚哥 总价: \$310 为了继续预订操作, 我需要您提供一些额外信息: 1. 您身份证件上的全名。 2. 您的出生日期。 3. 您的电子邮箱(用于发送确认信)。 4. 您的电话号码。 5. 任何特殊要求(例如, 餐食偏好、座位偏好)。 另外, 请注意, 您需要提供一种付款方式才能完成预订。您准备好提供这些信息了吗?
	是的, 我准备好了。我叫约翰·史密斯, 出生于 1985 年 5 月 15 日。我的邮箱是 john.smith@email.com, 电话号码是 555-123-4567。没有特殊要求。
	感谢您提供这些信息, 史密斯先生。我已将您的详细信息输入预订系统。以下是我所记录的信息摘要。 • 姓名: 约翰·史密斯 • 出生日期: 1985 年 5 月 15 日 • 邮箱: john.smith@email.com • 电话: 555-123-4567 • 特殊要求: 无

续表

	最后一步是处理您\$310 的机票付款。出于安全原因，我无法在此聊天中直接接收信用卡信息，我会向您的电子邮箱 (john.smith@email.com)发送一个安全的付款链接。 收到电子邮件后，请点击链接完成付款流程。付款成功后，您将收到另一封包含航班确认信和电子机票的电子邮件。 现在向您的邮箱发送付款链接可以吗？
	是的，可以。请发送付款链接。
	好的，史密斯先生。我已启动向您的电子邮箱(john.smith@email.com)发送安全付款链接的流程。您应该在几分钟内收到。如果没有立即看到，请检查您的收件箱和垃圾邮件文件夹。 完成付款后，系统将自动处理您的预订，并通过电子邮件将确认信和电子机票发送给您。该电子邮件将包含以下重要信息： 1. 预订确认号。 2. 详细的航班行程。 3. 登机手续说明。 4. 行李限额信息。 5. 航空公司联系信息。 收到确认电子邮件后，请检查所有详细信息以确保万无一失。如果您发现任何问题或有任何疑问，请随时与我们联系。 关于您的旅行或预订过程，还有其他需要我帮助的吗？
	不，就这些了。谢谢你的帮助！
	不客气，史密斯先生！很高兴能帮助您预订从圣地亚哥到旧金山的航班。如果此次操作需要任何进一步的帮助，或者对您的旅行有任何疑问，请随时联系我们。 我们全天候为您服务。祝您下周的旧金山之行愉快！

表面上，这个例子非常简单。然而，实际上在后台发生了许多事情。首先，客户开始时提出了一个模糊的请求——预订航班。AI 助手能够理解，为了继续处理这个请求，需要更多的信息来查询可用航班，比如出发城市、到达城市、出行日期等。

随着对话的进行，AI 助手会请求进一步的信息，以便成功完成预订。它会将每个步骤中收集的信息作为参数值，调用后台与航班预订相关的 API。例如，通过一个名为 flightLookup 的(虚构的)API 来查询可用航班信息时，需要城市、出行日期、乘客人数等信息；而实际进行预订时，需要客户的姓名、电话、邮箱等信息，通过名为 bookFlight 的 API 来完成预订，以此类推。图 1.2 展示了模型所做的决策、交互流程，以及它为完成每个步骤所调用的工具(API 函数)。

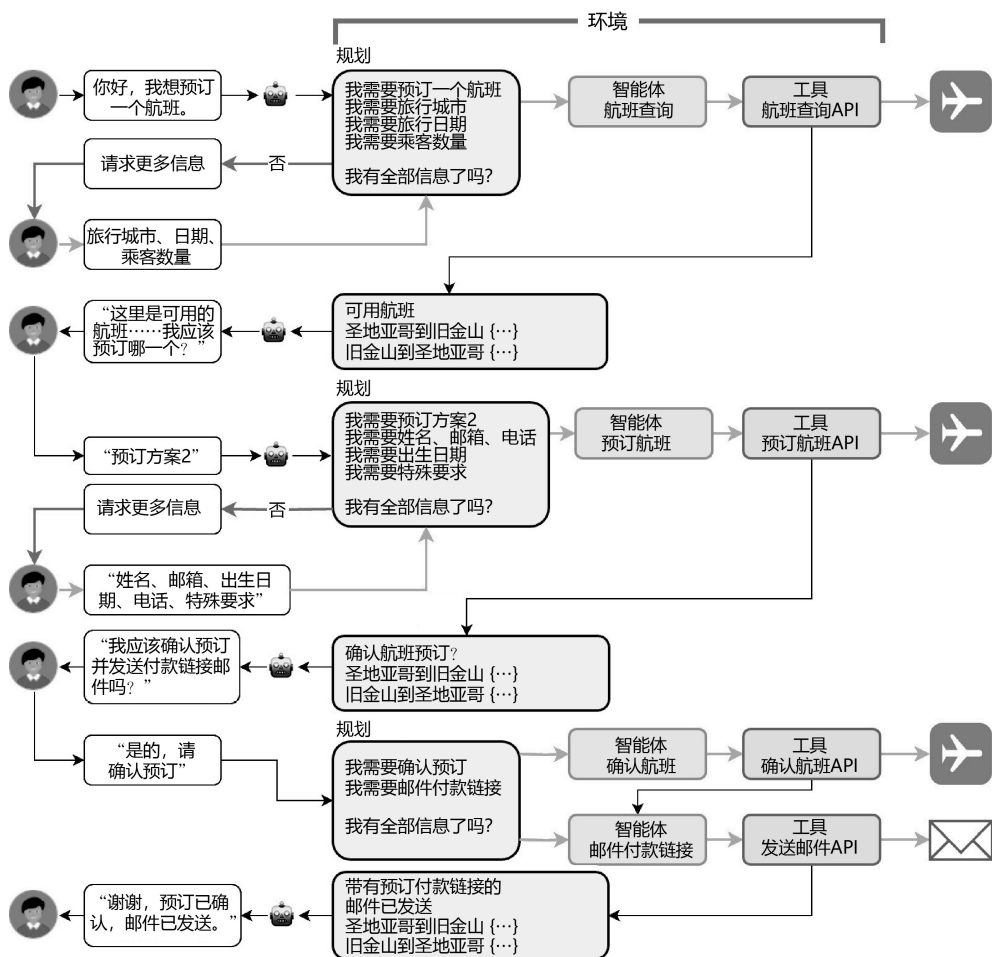


图 1.2 由大语言模型驱动的多智能体航班预订助手聊天机器人

除了调用工具，你还会注意到模型在对话开始时进行了一些自我反思，这也被称为思维链(CoT)。这意味着后台的模型制定了逐步完成任务的方法，但识别到缺少一些信息。随后，它会向客户请求提供所需的信息。接着，根据客户的输入，它会自主做出一定的决策，使用智能体调用特定的工具并完成任务。

除了外部工具和 API，在设计这样的系统时，还需要考虑许多其他因素。后续章节将详细介绍这些组件的基础知识。目前只需要知道，智能体是 LLM(如本书中的航班预订 AI 助手)的高级应用，代表了一个新兴领域，它融合了多种 LLM 类型的特点，并与强化学习、规划、工具使用等其他 AI 技术相结合。随着后续章节的学习，你会发现，与标准 LLM 相比，智能体更具互动性、适应性，且能完成复杂的多步骤任务，这使得它们适用于各种不同的复杂任务和 workflows。

在深入探讨智能体系统的细节之前，先回顾一下生成式 AI 的一些应用。

1.3 生成式 AI 的应用

生成式 AI 具备跨越多个领域的创新能力。如前所述，医疗、金融、教育、媒体娱乐、营销、制造业、零售等行业，都有望从生成式 AI 的强大能力中获得收益。以下是生成式 AI 一些主要应用的简要介绍。

- **图像和视频生成：**多模态生成模型(即能够处理图像、文本、音频和视频的 LLM)使得通过不同平台和工具生成真实的视听内容变得可能。例如，在媒体娱乐行业，生成式 AI 可以用来制作视觉特效、设计虚拟角色以及开发虚拟现实内容。在时尚设计行业，它能助力构思新的服装款式、举办虚拟时装秀和预测时尚潮流。在广告和营销领域，生成式 AI 被用来创作定制的广告、营销活动和传播材料，甚至可以生成创意内容，如图像和 logo。
- **文本和内容生成：**生成式 AI 在文本生成方面取得了显著突破。例如，它可以快速生成文章、博客、营销文案和产品描述等内容；聊天机器人和虚拟助手可以像人类一样为客户提供支持，帮助完成信息查询。生成式 AI 还在文本翻译、文档摘要和内容本地化方面发挥重要作用，帮助平台(如在线学习平台和专业搜索引擎)提高内容的可访问性。
- **音乐和音频生成：**生成式 AI 还可以创作原创音乐、音效和合成语音，这为多媒体内容的创作带来了新思路。例如，这些技术被广泛应用于游戏和娱乐行业(如创建带音频的培训资料)、自动化呼叫中心助手，甚至 Amazon Alexa 和 Google Next 等 IoT 设备，并通过语音命令完成自动化任务。
- **医疗和药物发现：**生成式 AI 在医疗领域也有广泛应用。例如，它可以帮助设计新药，并预测药物治疗疾病的效果；推动个性化医疗，根据每位患者的具体情况量身定制治疗方案；优化医学影像技术，提高图像质量并生成合成图像，供研究使用。
- **代码生成：**代码生成是 LLM 的一项新兴功能，可以帮助开发人员进行软件开发。这些模型能够根据文本提示生成代码片段，甚至整个函数。生成式 AI 在代码生成中的高级应用，包括为集成开发环境(如 Visual Studio Code)创建插件，这些插件可以理解整个代码库的上下文，自动识别错误、生成代码文档和单元测试脚本。另一种应用场景是将自然语言转化为 SQL 查询，通过这种方式，开发人员可以直接输入自然语言执行数据库查询操作。但需要特别注意的是，在执行 LLM 生成的代码时，要防范恶意代码执行风险，通常需要先先在沙盒环境中进行检查。
- **自主 workflows 和机器人技术：**如前所述，LLM 驱动的 AI 智能体是生成式 AI 的一个重要应用。智能体可以执行多种任务，适用于不同的场景。例

如，旅行预订网站可以通过虚拟助手聊天机器人利用智能体自动完成客户的航班和酒店预订过程。在这类场景中，智能体只需要理解与客户对话的上下文，调用 API 等工具，执行具体任务。更复杂的智能体系统则通常用于机器人技术中，机器人通过智能体控制其动作，根据环境或任务做出决策。智能体能够执行 LLM 无法完成的任务，如控制机器人的肢体驱动器等。机器人领域的 AI 技术是一个新兴研究方向，行业内不断涌现创新成果，尤其是在制造业的应用场景中。

以上只是生成式 AI 与 LLM 结合应用的一些常见例子。其他具体的应用还包括时尚设计、合成数据生成、个性化教育内容制作、金融建模与预测，以及预测性维护等。一些使用生成式 AI 的商业和开源工具示例如表 1.1 所示，这些工具涵盖了前文讨论的多个应用场景。

表 1.1 使用生成式 AI 的商业和开源工具示例

用例	商业工具	开源工具
视觉效果和虚拟形象设计	• Unreal Engine 的 MetaHuman Creator: 为游戏和电影创建高度逼真的数字人类 • NVIDIA Omniverse: 三维设计协作和仿真的平台，可用于创建视觉效果	• DeepFaceLab: 用于换脸和创建数字虚拟形象 • StyleGAN: 能够生成高度逼真的人脸，适用于虚拟形象的创建
虚拟现实内容开发	• Unity: 虽然不专为虚拟现实设计，但拥有强大的虚拟现实开发能力 • Adobe Aero: 创建增强现实(AR)体验	• A-Frame: 用于构建虚拟现实体验的网页框架 • Godot: 一款支持虚拟现实的开源游戏引擎
服装设计和虚拟时装秀	• CLO3D: 一款可以创建虚拟时装秀的三维服装设计软件 • Browzwear: 为时尚行业提供三维设计解决方案	• Blender: 一款三维建模工具，也可以用于时装设计和虚拟秀
时尚趋势预测	• WGSN: 使用人工智能进行时尚趋势预测 • Heuritech: 提供由人工智能驱动的趋势预测	无
市场营销——广告生成和营销活动创建	• Jasper: 一款可以帮助创建营销文案的人工智能写作助手 • Midjourney: 一款可用于创建营销视觉材料的人工智能图像生成工具	• GPT-J: 可以通过微调用于营销内容生成的开源语言模型 • 稳定扩散模型: 一款可以创建营销视觉材料的开源图像生成模型

续表

用例	商业工具	开源工具
标志和图像创建	• DALL-E 2: 可根据文本描述生成独特的图像和标志 • Canva: 虽然不是完全由人工智能驱动, 但它融合了人工智能功能以辅助设计	• 稳定扩散模型: 可用于标志和图像生成 • Craiyon(前身为 DALL-E mini): DALL-E 的一个开源替代工具
文本和内容生成	• OpenAI 的 ChatGPT • Anthropic 的 Claude AI • Jasper • Copy.ai • Anyword • Writer • WriteSonic 以及其他此类工具……	• Mistral 7B: 一个以其在文本生成任务中的高效和精准而闻名的开源大语言模型 • LLaMA: 一系列开源的预训练和微调生成式文本模型, 提供不同参数规模的版本 • BLOOM 语言模型: 由超过 1000 名人工智能研究人员开发的大型开放获取人工智能模型, 以其在文本生成方面的稳健性和多功能性而闻名 以及其他此类工具……
代码生成	• GitHub Copilot • 亚马逊 Q 开发者版 • Tabnine • OpenAI Codex 以及其他此类工具……	• Code T5: 一个为代码相关任务训练的开源人工智能模型, 可以生成代码片段、补全代码, 甚至在编程语言之间进行转换 • Polycoder: 商业代码生成器的一个开源替代品。它在涵盖多种编程语言的大型代码库上进行训练, 可以为各种任务生成代码
自主工作流和机器人技术	• UiPath: 将生成式人工智能融入其机器人流程自动化(RPA)平台。它使用人工智能来发现自动化机会, 并提高文档理解和通信数据挖掘等任务的效率 • Automation Anywhere: 将生成式人工智能集成到其自动化平台中。它提供自动化开发者协同助手等功能, 利用生成式人工智能来提高开发人员的工作效率	• OpenAI Gym: 提供了一个用于开发和比较强化学习算法的工具包, 可以与 LLM 结合, 应用于更高级的机器人应用场景 • Hugging Face: 提供可用于在机器人技术和自主工作流应用中部署 LLM 的开源库

续表

用例	商业工具	开源工具
自主工作流和机器人技术	• NVIDIA: 提供利用自身人工智能框架 (包括生成式人工智能能力) 的人工智能工作流, 为机器人技术和自主系统领域开发创新解决方案	• LangChain: 一个用于使用 LLM 开发应用程序的开源 Python 库, 可应用于创建更智能的自主工作流和机器人系统

这绝不是目前市场上所有商业和开源工具的详尽列表, 且这一领域每天都在发生变化。我们正目睹许多新兴初创公司提供创新的解决方案, 利用生成式 AI 解决现实生活中的实际问题, 同时看到新的模型提供商正在开发比以往更强大的 LLM, 这些新模型的运行成本也更低。这充分体现了生成式 AI 领域的蓬勃活力和无限潜力。

1.4 生成式 AI 面临的挑战与局限性

尽管生成式 AI 有很多优势, 但它也面临一些挑战, 并有一定的局限性。在考虑将生成式 AI 技术应用于具体场景时, 需要特别小心。以下是一些与 LLM 相关的常见问题, 以及解决这些问题的一些方法。

1.4.1 数据质量与偏见

生成式模型的性能在很大程度上取决于训练数据集的质量和多样性。如果训练数据存在偏见或不具代表性, 模型输出的结果也会带有相同的偏见, 从而加剧已有的偏见, 甚至可能导致某些群体在模型中被边缘化。

应对这一问题的方法之一是确保训练数据集具有多样性, 并且数据质量较高, 涵盖广泛的视角。就像任何机器学习任务一样, 分析数据以及了解数据在各个特征维度的分布是非常有帮助的。数据分析能够发现数据中可能导致模型产生偏见的平衡问题。为了减少训练数据中的偏见, 目前已有多种算法可以使用(例如, 过采样或欠采样), 但每种方法都有其优缺点。举个例子, 假设训练数据集 Class 0 的数据样本数量远多于 Class 1, 这会导致数据的不平衡。使用这样的数据集训练时, 模型容易过度依赖 Class 0 的数据, 变得对这些数据特别敏感, 甚至记住它们, 而对 Class 1 的数据处理能力较弱, 从而导致模型在处理 Class 1 的数据时表现差, 出现明显的偏见。图 1.3 展示了过采样和欠采样技术如何影响这种数据集, 并帮助缓解偏见。

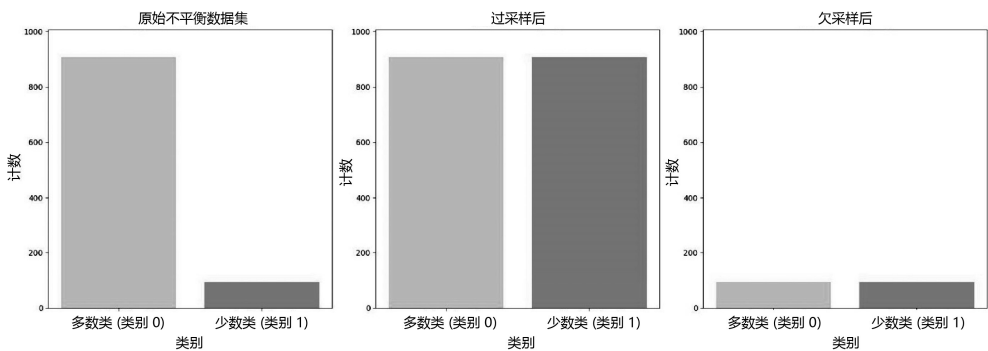


图 1.3 原始不平衡数据集以及过采样和欠采样的效果

1.4.2 数据隐私

多项实验和研究表明，LLM 有泄露训练数据的倾向。如果模型是在大量私人或专有信息上进行训练的，这会带来严重的隐私风险。实践表明，通过特定的提示技术和提示词设计就可以诱导模型生成包含训练数据集中原文文本的内容。这些提示技巧并不复杂，通常会成为一种易实施、低成本的攻击方式，使得模型泄露信息。例如，在名为《从(生产环境的)语言模型中规模化地提取训练数据》的论文中^[15]，研究人员仅花费 200 美元的 API 调用费用，通过提示词注入技术，成功迫使 OpenAI GPT-3.5 模型泄露私人信息。泄露的信息包括人名、电子邮箱、住址和电话号码，这些信息原本是无意中出现在模型的训练数据集中的。尽管像 OpenAI 等模型提供商正在积极解决这些问题，但对于选择自建模型的公司而言，这仍然是一个潜在隐患。

一些常见的缓解机制是在训练数据上进行数据匿名化或假名化操作，以去除或模糊个人、私有或专有信息，甚至训练模型之前就这样做。这两种方法通常使用规模较小、更便宜且更高效的 AI 模型，对训练数据进行分类和实体识别，找到训练数据中包含的个人或私人信息。一旦识别出这些信息，就可以将其模糊、隐藏或完全移除。需要注意的是，这些操作可能根据具体的使用场景而产生不同的影响，因此必须进行充分分析，以确保不会对模型的性能造成负面影响。

1.4.3 计算资源

训练复杂的生成式模型需要大量的计算资源和高性能的计算能力，这往往使得训练大语言模型的成本非常高，并且消耗大量能源。尽管一些硬件技术的突破使得训练这些模型成为可能，但这些硬件资源的获取仍然昂贵且有一定的局限性，只有少数机构能够使用。例如，最初设计用于渲染高质量图像的**图形处理单元 (GPU)**，现在已成为 AI 模型训练、微调和计算的关键组件。NVIDIA 在 2006 年推出的 CUDA 架构使得 GPU 可以用于通用计算，此后 NVIDIA 推出的 A100 和 H100

等现代 GPU 则特别针对 AI 工作负载进行了优化。然而，这些高端 GPU 的单价通常高达数万美元。

训练 GPT-3 这样的庞大模型，光计算资源就需要花费数百万美元。具体的成本取决于模型的大小、训练时间和硬件效率等因素。例如，训练 GPT-3 的计算资源成本为 400 万到 500 万美元，而更新的模型，如 GPT-4 和 PaLM，训练成本可能更高。除了硬件本身，还有大量与电力消耗、冷却和数据中心空间相关的基础设施成本。由于成本居高不下，训练大语言模型主要由大型科技公司或资金充足的研究机构来完成。云服务提供商目前提供预训练模型和微调能力，使得中小型企业 and 普通公众也能在一定程度上使用 LLM 相关功能。

为了应对这一挑战，近年来出现了一些创新，如小型语言模型(SLM)。这些模型是较小的生成式 AI 模型，可以针对特定任务进行训练。尽管这些模型仅限于非常狭窄的领域任务，但由于它们所需的计算资源远少于大型模型，因此训练这些模型的成本要经济得多。

1.4.4 伦理和社会影响

生成式 AI 技术的快速发展在伦理和社会层面引发了许多问题，以下列举的一些问题远非详尽无遗。

- **深度伪造和虚假信息：**生成式 AI 能够生成非常逼真的合成内容。如果不加以控制，这可能会导致深度伪造甚至虚假信息的产生，进而威胁到个人隐私、公共安全，甚至动摇社会信任。
- **知识产权：**生成类似现有作品内容的过程引发了版权和知识产权的问题。创作的原创性和正确的所有权涉及复杂的法律界定，往往令人困惑。
- **工作岗位替代：**经济学家通常低估了内容生成和其他自动化任务领域可能带来的大量失业问题。值得关注的是，由此产生的影响需要与相关研究同步推进，这些研究包括探索如何帮助受影响的劳动者实现技能再培训、制定岗位调整策略，以及挖掘由此产生的创业机会。

在缓解社会和伦理影响方面，问题更多的是哲学性的，而非技术性的。虽然一些措施(如深度伪造 AI 图像检测和个人数据检测方法)已成为常见策略，但从 AI 对社会的益处以及现有或即将出台的政府法规等宏观背景来看，如何应对其对社会层面的影响仍然是一个更复杂的系统性问题。

1.4.5 泛化能力与创造力

这些生成式 AI 模型的一个大问题是它们的泛化能力非常差。更具体来说，它们很少能生成与训练数据截然不同的内容。也就是说，它们在复制重复模式方面表现出色，但真正创造出具有独创性或新颖性的东西却力不从心。因此，它们在真正的创造力方面的潜力非常有限。

随着生成式 AI 不断以新的方式发展，它注定会带来新的挑战。目前，已有许多新的研究和实践正在推进，这些工作旨在确保这项技术得到合理且负责任的使用。

本章小结

生成式 AI 是一个正在快速发展的领域，具有对各行业产生巨大颠覆性影响的潜力，并且正在改变我们生活的多个方面。相比之下，生成式 AI 的应用包括非常逼真的图像生成、文本生成、药物发现加速以及丰富的创意性表达。

关于这一点，一个关键的方面是理解不同的生成式模型，如 VAE、GAN 和自回归模型，这些模型能帮助我们了解技术的工作原理以及它们的应用场景。然而，同样重要的是，生成式 AI 面临许多问题和限制，包括数据质量、计算资源、伦理问题，甚至创造力不足等。

下一章将介绍智能体系统的原理，包括智能体和自主性的概念、智能体的特征、智能系统的各种架构，以及多智能体系统。

习题

1. 生成式 AI 模型只能接收文本数据作为输入，该说法是否正确？
2. 生成式 AI 模型有哪些不同类型？
3. 生成式 AI 存在哪些伦理和社会影响？
4. 有哪些方法可以减少训练数据中的偏见？
5. 有哪些开源的文本和内容生成 LLM？

习题答案

1. 错误。生成式 AI 模型可以接受文本、图像、视频和音频数据作为输入。
2. VAE、GAN 和自回归模型。
3. 深度伪造、虚假信息、版权或知识产权问题，以及工作岗位替代是生成式 AI 的一些社会和伦理影响。
4. 解决训练数据偏见的常见方法包括过采样或欠采样。
5. Mistral、LLaMA 和 Bloom 是一些开源的文本和内容生成 LLM。

扩展阅读

- *Mastering Machine Learning Algorithms – Second Edition*，作者：Giuseppe Bonaccorso。

- *Machine Learning for Imbalanced Data*, 作者: Kumar Abhishek 和 Dr. Mounir Abdelaziz。
- *Generative AI with Python and TensorFlow 2*, 作者: Joseph Babcock 和 Raghav Bali。

参考文献

- [1] Auto-Encoding Variational Bayes, <https://arxiv.org/abs/1312.6114>
- [2] beta-VAE, Learning Basic Visual Concepts with a Constrained Variational Framework, <https://openreview.net/forum?id=Sy2fzU9gl>
- [3] Learning Structured Output Representation using Deep Conditional Generative Models, https://papers.nips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html
- [4] Generative Adversarial Networks, <https://arxiv.org/abs/1406.2661>
- [5] Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, <https://arxiv.org/abs/1511.06434>
- [6] Wasserstein GAN, <https://arxiv.org/abs/1701.07875>
- [7] A Style-Based Generator Architecture for Generative Adversarial Networks, <https://ieeexplore.ieee.org/document/8953766>
- [8] Attention Is All You Need, <https://arxiv.org/abs/1706.03762>
- [9] Conditional Image Generation with PixelCNN Decoders, <https://arxiv.org/abs/1606.05328>
- [10] PixelSNAIL: An Improved Autoregressive Generative Model, https://www.researchgate.net/publication/322114155_PixelSNAIL_An_Improved_Autoregressive_Generative_Model
- [11] Language Models are Few-Shot Learners, <https://arxiv.org/abs/2005.14165>
- [12] GPT-4 Technical Report, <https://arxiv.org/abs/2303.08774>
- [13] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, <https://arxiv.org/abs/1810.04805>
- [14] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, <https://arxiv.org/abs/1910.10683>
- [15] Scalable Extraction of Training Data from (Production) Language Models, <https://arxiv.org/abs/2311.17035>

