LangChain
大模型应用开发

[英] 本 • 奥法斯(Ben Auffarth) 著
郭涛 译

清華大學出版社
北 京

北京市版权局著作权合同登记号　图字：01-2024-3384

图书在版编目（CIP）数据

LangChain 大模型应用开发 / (英) 本·奥法斯 (Ben Auffarth) 著 ; 郭涛译. -- 北京 : 清华大学出版社, 2025. 1. -- ISBN 978-7-302-67729-1

Ⅰ. TP311.561

中国国家版本馆 CIP 数据核字第 2024AQ0586 号

责任编辑：王　军
封面设计：高娟妮
版式设计：恒复文化
责任校对：成凤进
责任印制：沈　露

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：北京同文印刷有限责任公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　　张：**16.25　　**字　　数：**347 千字
版　　次：2025 年 1 月第 1 版　　**印　　次：**2025 年 1 月第 1 次印刷
定　　价：79.80 元

产品编号：105021-01

译 者 序

大模型的出现和落地开启了人工智能(AI)新一轮的信息技术革命，改变了人们的生活方式、工作方式和思维方式。大模型的落地需要数据、算力和算法三大要素。经过几年发展，大模型的数据集(包括多模态数据集)制作已经形成了规约，Meta、Google 和百度等人工智能公司都有自己的一套数据集标准制作流程。算力方面主要依托 GPU、TPU 等硬件资源进行集群计算(即并行计算)。在算法方面，主要以 Transformer 架构为主流框架，出现了 OpenAI 的 GPT 系列大模型、Meta 的 Llama 系列大模型以及清华大学的 ChatGLM 系列大模型。目前虽然已经有几千个甚至更多的大模型，但要从零开始训练一个大模型，需要很大的财力、物力和资源，个人和一些小团队很难做到。从数据集制作来说，需要几十位数据科学家和工程师构建数据预处理工作流，要安排大量标注人员实现数据标注。从硬件资源来看，训练一个大模型需要几千张甚至更多的 GPU 卡，需要训练几个月甚至更久。从人力资源来看，需要投入十几位计算机科学家、算法工程师和软件工程师，这是很多企业及个人望尘莫及的。

通过“七七四十九天炼丹”，大模型出炉了。但也并不意味着可以直接拿来使用，横扫市场，挣得盆满钵满。有了基础大模型，还要根据应用领域、场景和需求，在基础上做微调，使大模型既具备通用知识，又有“专业技能”。接下来，急需一个平台来实现大模型微调、部署、管理、接口设计和应用开发，打造开启大模型时代的钥匙。

在这种背景下，在基础大模型基础上形成了微调和提示工程等新的技术范式。同时也出现了大模型应用落地的软件产品，如 LangChain、Ollama、Chatbox、LM Studio、AnythingLLM、LocalAI 和 MaxKB 等，主要用于大模型微调、部署、管理和应用服务开发。这些产品各有特色，要根据自己的业务场景、业务需求和特色选择。本书主要讨论 LangChain，它提供了一个完整的生态系统，为开发者带来了一系列核心模块和工具。另外值得一提的是，针对大模型，各个企业研发出了向量数据库，向量数据库可以很好地组织和管理半结构化、非结构化数据，将所有数据以嵌入方式形成向量表示。

本书围绕大模型、生成式人工智能、LangChain 等主题，以理论、案例和近几年的技术前沿为主线展开，以代码实现为途径，适合大模型应用开发、人工智能和大数据等领域的学者和工程师阅读，也可以作为非计算机背景人员作为入门大模型应用实战的读物。

在翻译本书的过程中，我查阅了大量的经典著(译)作，也得到了很多人的帮助。感谢本书的审校者——电子科技大学外国语学院研究生相思思。最后，感谢清华大学出版社的编辑，他们做了大量的编辑与校对工作，保证了本书的质量，使得本书符合出版要求。在此深表谢意。

由于本书涉及的内容广度和深度较大，加上译者翻译水平有限，在翻译过程中难免有不足之处，若各位读者在阅读过程中发现问题，欢迎批评指正。

译者

译 者 简 介

郭涛，主要从事人工智能、智能计算、概率与统计学、现代软件工程等前沿技术的交叉研究。出版多部译作，包括《OpenAI API 编程实践 III(Java 版)》《Python 预训练视觉和大语言模型》《概率图模型原理与应用(第 2 版)》。

作 者 简 介

Ben Auffarth 是一位经验丰富的数据科学领导者，拥有计算神经科学博士学位。Ben 分析过 TB 级数据，在核数多达 64k 的超级计算机上模拟过大脑活动，设计并开展过湿法实验室实验，构建过处理承保应用的生产系统，并在数百万文档上训练过神经网络。他著有 *Machine Learning for Time Series* 和 *Artificial Intelligence with Python Cookbook* 两本书，现于 Hastings Direct 从事保险工作。

撰稿人简介

Leonid Ganeline 是一名机器学习工程师，在自然语言处理方面拥有丰富的经验。他曾在多家初创公司工作，创建模型和生产系统。他是 LangChain 和其他几个开源项目的积极贡献者。他的兴趣是模型评估，特别是大规模语言模型评估。

审校者简介

Ruchi Bhatia 是一名计算机工程师，拥有卡内基梅隆大学信息系统管理硕士学位。目前，她在惠普公司担任产品营销经理，在快速发展的数据科学和人工智能领域发挥自己的一技之长。作为 Kaggle 的 Notebooks，Datasets and Discussion(笔记、数据集和讨论)类别中最年轻的三连冠顶级大师，她为此感到自豪。她曾在 OpenMined 担任数据科学负责人，带领数据科学家团队创造出创新而有影响力的解决方案。

致　　谢

创作本书是一段漫长(甚至是艰难)的旅程，但也是一段激动人心的旅程。我非常感谢几位关键人物的贡献，他们的贡献让本书更加丰富多彩。首先，我衷心感谢 Leo，他富有洞察力的反馈意见极大地完善了本书。同样让我感到高兴的还有我睿智的编辑们——Tanya、Elliot 和 Kushal。他们的努力超出了我的预期。尤其是 Tanya，她在指导我写作的过程中发挥了重要作用，不断地督促我理清思路，对成书产生了重要影响。

Ben Auffarth

我要感谢我的父母，感谢他们教会我如何理性思考；还要感谢我的妻子，感谢她支持我的工作。

Leonid Ganeline

我想借此机会向我的父母表示衷心的感谢。在我的人生旅途中，他们给予我坚定的支持和鼓励，这对我来说弥足珍贵。如果没有他们对我能力的信任和长期的指导，我不可能取得今天的成就。爸爸妈妈，谢谢你们一直陪在我身边。

Ruchi Bhatia

前　言

在充满活力、飞速发展的人工智能领域，生成式人工智能作为一股颠覆性力量脱颖而出，它将改变我们与技术的交互方式。本书是对**大规模语言模型(Large Language Model，LLM)**(推动这一变革的强大引擎)这一错综复杂的世界的一次探险，旨在让开发人员、研究人员和人工智能爱好者掌握利用这些工具所需要的知识。

探究深度学习的奥秘，让非结构化数据焕发生机，了解 GPT-4 等大规模语言模型如何为人工智能影响企业、社会和个人开辟道路。随着科技行业和媒体对这些模型的能力和潜力的热切关注，现在正是探索它们如何发挥作用、蓬勃发展并推动我们迈向未来的大好时机。

这本书就像你的指南针，指引你理解支撑大规模语言模型的技术框架。本书提供了一个引子，让你了解它们的广泛应用、基础架构的精巧以及其存在的强大意义。本书面向不同的读者，从初涉人工智能领域的人到经验丰富的开发人员。我们将理论概念与实用的、代码丰富的示例融为一体，让你不仅从知识上掌握大规模语言模型，还能创造性地、负责任地应用大规模语言模型。

当我们一起踏上这段旅程时，让我们做好准备，塑造此时此刻正在展开的人工智能生成叙事，同时也被它所塑造——在这一叙事中，你将用知识和远见武装自己，站在这一令人振奋的技术演进的最前沿。

读者对象

本书面向开发人员、研究人员以及其他任何有兴趣进一步了解大规模语言模型的人。本书的编写简洁明了，包含大量代码示例，你可以边做边学。

无论是初学者还是经验丰富的开发人员，对于任何想要充分利用大规模语言模型并在大规模语言模型和 LangChain 方面保持领先的人来说，这本书都将是宝贵的资源。

主要内容

第 1 章介绍了以深度学习为核心的生成式人工智能如何彻底改变了文本、图像和视频的处理方式。该章介绍了大规模语言模型等生成式模型，详细介绍了它们的技术基础和在各个领域的变革潜力；涵盖了这些模型背后的理论，重点介绍了神经网络和训练方法，以及类人内容的创建。该章概述了人工智能的演变、Transformer 架构、文本到图像模型(如 Stable Diffusion)，并涉及声音和视频应用。

第 2 章揭示了超越大规模语言模型随机鹦鹉(模仿语言但无法真正理解语言的模型)的必要性。针对过时的知识、行动限制和幻觉风险等局限性，该章重点介绍了 LangChain 如何集成外部数据和干预措施，以实现更连贯的人工智能应用。该章批判性地探讨了随机鹦鹉的概念，揭示了产生流畅但无意义语言的模型缺陷，并阐述了提示、思维链推理和检索增强生成是如何改善大规模语言模型以解决语境、偏差和不透明等问题。

第 3 章介绍了基础知识，帮助设置运行书中所有示例的环境。该章首先介绍了 Docker、Conda、Pip 和 Poetry 的安装指南，然后详细介绍了如何集成 OpenAI 的 ChatGPT 和 Hugging Face 等不同提供商的模型，包括获取必要的 API 密钥。该章还讨论了在本地运行开源模型的问题。最后，该章将构建一个大规模语言模型应用程序来协助客户服务智能体，以实例说明 LangChain 如何简化操作并提高响应的准确性。

第 4 章通过加入事实核查来减少虚假信息，采用复杂的提示策略进行总结，整合外部工具来增强知识，从而将大规模语言模型转变为可靠的助手。该章探讨了用于信息提取的密度链(Chain of Density)，以及用于自定义行为的 LangChain 装饰器和表达语言。该章还介绍了 LangChain 中处理长文档的 map-reduce，并讨论了管理 API 使用成本的词元数监控。

该章着眼于实现一个 Streamlit 应用程序来创建交互式大规模语言模型应用程序，并利用函数调用和工具的使用来超越基本的文本生成。该章介绍了两种不同的智能体范式——规划-求解(plan-and-solve)和零样本(zero-shot)——以演示决策策略。

第 5 章深入探讨了利用**检索增强生成**(Retrieval-Augmented Generation，RAG)来增强聊天机器人的能力，这种方法可让大规模语言模型获取外部知识，提高其准确性和特定领域的熟练程度。该章讨论了文档向量化、高效索引以及使用 Milvus 和 Pinecone 等向量数据库进行语义搜索。我们实现了一个聊天机器人，结合调节链来确保负责任的沟通。这个聊天机器人可以在 GitHub 上找到，它是探索对话记忆和上下文管理等高级主题的基础。

第 6 章探讨了大规模语言模型在软件开发中的新兴作用，强调了人工智能在自动编码任务和充当动态编码助手方面的潜力。该章探讨了人工智能驱动的软件开发现状，用模型进行实验以生成代码片段，并介绍了使用 LangChain 智能体来自动化软件设计开发。对智能体性能的批判性反思强调了人类监督对减少错误和复杂设计的重要性，为人工智能和人类开发人员共生合作的未来奠定了基础。

第 7 章探讨了生成式人工智能与数据科学的交集，强调了大规模语言模型在提高生产力和推动科学发现方面的潜力。该章概述了当前通过 AutoML 实现数据科学自动化的范围，并将大规模语言模型与增强数据集和生成可执行代码等高级任务相结合，扩展了 AutoML 数据学科自动化概念。该章还介绍了使用大规模语言模型进行探索性数据分析、运行 SQL 查询和可视化统计数据的实用方法。最后，智能体和工具的使用展示了大规模语言模型如何解决以数据为中心的复杂问题。

第 8 章深入探讨了微调和提示工程等调节技术，这些技术对于定制大规模语言模型性能以适应复杂的推理和专业任务至关重要。该章对微调(fine-tuning)和提示工程(prompting)进行了解读。微调是指根据特定任务的数据对大规模语言模型进行进一步训练，而提示工程则是指战略性地引导大规模语言模型生成所需要的输出。该章还实施了高级提示策略，如少样本学习(few-shot learning)和思维链(chain-of-thought)，以增强大规模语言模型的推理能力。该章不仅提供了微调和提示工程的具体实例，还讨论了大规模语言模型的未来发展及领域应用。

第 9 章探讨了在实际应用中部署大规模语言模型的复杂性，涵盖了确保性能、满足监管要求、规模稳健性和有效监控的最佳实践。该章强调了评估、可观察性和系统化操作的重要性，以使生成式人工智能在客户参与和具有财务后果的决策中受益。该章还概述了利用 Fast API、Ray 等工具以及 LangServe 和 LangSmith 等新工具部署和持续监控大规模语言模型应用程序的实用策略。这些工具可以提供自动评估和衡量标准，支持各部门负责任地采用生成式人工智能。

第 10 章探讨了生成式人工智能的潜在进步和社会技术挑战，探讨了这些技术对经济和社会的影响，讨论了工作取代、虚假信息及人类价值一致性等伦理问题。在各行各业为人工智能引发的颠覆性变革做好准备之际，本章反思了企业、立法者和技术专家建立有效治理框架的责任。该章还强调了引导人工智能发展以增强人类潜能的重要性，同时解决了深度伪造、偏见和人工智能武器化等风险；还强调了透明度、道德部署和公平使用的紧迫性，以积极引导生成式人工智能革命。

充分利用本书

要充分受益于本书所提供的价值，至少需要对 Python 的基本知识有一个了解。此外，掌握一些机器学习或神经网络的基础知识也会有所帮助，但这并不是必需的。请务必遵守第 3 章 3.1 节中设置的 Python 环境说明，并获取 OpenAI 和其他提供商的访问密钥。

notebook 和项目使用方式

本书的代码托管在 GitHub 上，网址是 https://github.com/benman1/generative_ai_with_langchain。在该仓库中，你可以找到每一章目录，其中包含了本书中需要的 notebook 和项目。也可扫描封底二维码下载本书源代码。如前所述，在使用代码之前，请确保按照第 3 章所述的说明安装必要的依赖项。如果你有任何问题或疑虑，请在 Discord 上提问或在 GitHub 上提交问题。

下载彩色图片

我们还提供了一个 PDF 文件，其中包含本书所用截图/图表的彩色图片。可扫描封底的二维码下载。

文本约定

本书中使用了一些文本约定。

CodeInText(文本代码)：表示文本、数据库表名、文件夹名、文件名、文件扩展名、路径名、虚拟 URL、用户输入和 Twitter 句柄中的代码词。例如“将下载的 WebStorm-10*.dmg 磁盘镜像文件挂载到系统中的另一个磁盘”。

代码块设置如下：

```
from langchain.chains import LLMCheckerChain
from langchain.llms import OpenAI

llm = OpenAI(temperature=0.7)

text = "What type of mammal lays the biggest eggs?"
```

当我们希望你注意代码块中的特定部分时，相关的行或项会以粗体显示：

```
from pandasai.llm.openai import OpenAI
llm = OpenAI(api_token="YOUR_API_TOKEN")

pandas_ai = PandasAI(llm)
```

任何命令行输入或输出的写法如下：

```
pip install -r requirements.txt
```

目　录

第 1 章　什么是生成式人工智能 …… 1
1.1　生成式人工智能简介 …… 1
1.1.1　什么是生成式模型 …… 4
1.1.2　为什么是现在 …… 5
1.2　了解大规模语言模型 …… 6
1.2.1　GPT 模型是如何工作的 …… 7
1.2.2　GPT 模型是如何发展的 …… 12
1.2.3　如何使用大规模语言模型 …… 17
1.3　什么是文本到图像模型 …… 18
1.4　人工智能在其他领域的作用 …… 22
1.5　小结 …… 23
1.6　问题 …… 23

第 2 章　面向大规模语言模型应用程序：LangChain …… 25
2.1　超越随机鹦鹉 …… 25
2.1.1　大规模语言模型的局限性 …… 27
2.1.2　如何减少大规模语言模型的局限性 …… 27
2.1.3　什么是大规模语言模型应用程序 …… 28
2.2　LangChain 简介 …… 30
2.3　探索 LangChain 的关键组件 …… 33
2.3.1　链 …… 33
2.3.2　智能体 …… 34
2.3.3　记忆 …… 35
2.3.4　工具 …… 36
2.4　LangChain 如何工作 …… 38
2.5　LangChain 软件包结构 …… 40
2.6　LangChain 与其他框架的比较 …… 41
2.7　小结 …… 43
2.8　问题 …… 44

第 3 章　LangChain 入门 …… 45
3.1　如何为本书设置依赖 …… 46
3.2　探索 API 模型集成 …… 49
3.2.1　环境设置和 API 密钥 …… 50
3.2.2　OpenAI …… 51
3.2.3　Hugging Face …… 52
3.2.4　谷歌云平台 …… 53
3.3　大规模语言模型交互基石 …… 54
3.3.1　大规模语言模型 …… 54
3.3.2　模拟大规模语言模型 …… 55
3.3.3　聊天模型 …… 56
3.3.4　提示 …… 57
3.3.5　链 …… 59
3.3.6　LangChain 表达式语言 …… 60
3.3.7　文本到图像 …… 61
3.3.8　Dall-E …… 61
3.3.9　Replicate …… 63
3.3.10　图像理解 …… 64
3.4　运行本地模型 …… 65
3.4.1　Hugging Face transformers …… 66
3.4.2　llama.cpp …… 68
3.4.3　GPT4All …… 69
3.5　构建客户服务应用程序 …… 70
3.5.1　情感分析 …… 70
3.5.2　文本分类 …… 71

3.5.3 生成摘要 72
3.5.4 应用 map-reduce 73
3.5.5 监控词元使用情况 76
3.6 小结 77
3.7 问题 77

第 4 章 构建得力助手 79
4.1 使用工具回答问题 80
4.1.1 工具使用 80
4.1.2 定义自定义工具 81
4.1.3 工具装饰器 82
4.1.4 子类化 BaseTool 82
4.1.5 StructuredTool 数据类 83
4.1.6 错误处理 84
4.2 使用工具实现研究助手 85
4.3 探索推理策略 89
4.4 从文件中提取结构化信息 95
4.5 通过事实核查减少幻觉 100
4.6 小结 102
4.7 问题 102

第 5 章 构建类似 ChatGPT 的聊天机器人 103
5.1 什么是聊天机器人 104
5.2 从向量到 RAG 105
5.2.1 向量嵌入 106
5.2.2 在 LangChain 中的嵌入 107
5.2.3 向量存储 109
5.2.4 向量索引 110
5.2.5 向量库 111
5.2.6 向量数据库 112
5.2.7 文档加载器 117
5.2.8 LangChain 中的检索器 118
5.3 使用检索器实现聊天机器人 120
5.3.1 文档加载器 121
5.3.2 向量存储 122
5.3.3 对话记忆：保留上下文 125
5.4 调节响应 130
5.5 防护 131
5.6 小结 132
5.7 问题 132

第 6 章 利用生成式人工智能开发软件 133
6.1 软件开发与人工智能 134
6.2 使用大规模语言模型编写代码 138
6.2.1 Vertex AI 138
6.2.2 StarCoder 139
6.2.3 StarChat 143
6.2.4 Llama 2 144
6.2.5 小型本地模型 145
6.3 自动化软件开发 147
6.3.1 实现反馈回路 149
6.3.2 使用工具 152
6.3.3 错误处理 154
6.3.4 为开发人员做最后的润色 155
6.4 小结 157
6.5 问题 157

第 7 章 用于数据科学的大规模语言模型 159
7.1 生成式模型对数据科学的影响 160
7.2 自动化数据科学 162
7.2.1 数据收集 163
7.2.2 可视化和 EDA 164
7.2.3 预处理和特征提取 164
7.2.4 AutoML 164
7.3 使用智能体回答数据科学的问题 166

7.4 使用大规模语言模型进行数据探索 …… 169
7.5 小结 …… 173
7.6 问题 …… 173

第 8 章 定制大规模语言模型及其输出 …… 175
8.1 调节大规模语言模型 …… 176
8.2 微调 …… 180
8.2.1 微调设置 …… 181
8.2.2 开源模型 …… 184
8.2.3 商业模型 …… 187
8.3 提示工程 …… 188
8.3.1 提示技术 …… 190
8.3.2 思维链提示 …… 192
8.3.3 自一致性 …… 193
8.3.4 思维树 …… 195
8.4 小结 …… 198
8.5 问题 …… 198

第 9 章 生产中的生成式人工智能 …… 199
9.1 如何让大规模语言模型应用程序做好生产准备 …… 200
9.2 如何评估大规模语言模型应用程序 …… 202
9.2.1 比较两个输出 …… 204
9.2.2 根据标准进行比较 …… 205
9.2.3 字符串和语义比较 …… 206
9.2.4 根据数据集进行评估 …… 207
9.3 如何部署大规模语言模型应用程序 …… 211
9.3.1 FastAPI Web 服务 …… 213
9.3.2 Ray …… 216
9.4 如何观察大规模语言模型应用程序 …… 219
9.4.1 跟踪响应 …… 221
9.4.2 可观察性工具 …… 223
9.4.3 LangSmith …… 224
9.4.4 PromptWatch …… 225
9.5 小结 …… 227
9.6 问题 …… 227

第 10 章 生成式模型的未来 …… 229
10.1 生成式人工智能的现状 …… 229
10.1.1 挑战 …… 230
10.1.2 模型开发的趋势 …… 231
10.1.3 大科技公司与小企业 …… 234
10.1.4 通用人工智能 …… 235
10.2 经济后果 …… 236
10.2.1 创意产业 …… 238
10.2.2 教育 …… 239
10.2.3 法律 …… 239
10.2.4 制造业 …… 239
10.2.5 医学 …… 240
10.2.6 军事 …… 240
10.3 社会影响 …… 240
10.3.1 虚假信息与网络安全 …… 241
10.3.2 法规和实施挑战 …… 241
10.4 未来之路 …… 243

第1章 什么是生成式人工智能

在过去 10 年中，深度学习在处理和生成文本、图像和视频等非结构化数据方面取得了巨大发展。这些先进的人工智能模型在各行各业都备受欢迎，其中包括**大规模语言模型(Large Language Models，LLM)**。目前，媒体和业界都在大张旗鼓地宣传人工智能，有理由相信，随着这些进步，**人工智能(Artificial Intelligence，AI)**即将对企业、社会和个人产生广泛而重大的影响。推动人工智能发展的因素有很多，包括技术进步、高知名度的应用程序以及在多个领域产生变革性影响的潜力。

本章将探讨生成式模型及其基础知识。将对技术概念和训练方法进行概述，这些技术概念和训练方法是模型生成新颖内容的驱动力。本章不会深入探讨声音或视频的生成式模型，我们的目标是传达一种高层次的理解，即神经网络、大型数据集和计算规模等技术如何使生成式模型在文本和图像生成中实现新功能。我们要揭开这些模型的神秘面纱，看看它们有什么魔力，能在不同领域生成与人类相似的内容。在此基础上，读者将能更好地思考这一快速发展的技术所带来的机遇和挑战。

本章主要内容：

- 生成式人工智能简介
- 了解大规模语言模型
- 模型开发
- 什么是文本到图像模型
- 人工智能在其他领域能做些什么

从介绍术语开始吧！

1.1 生成式人工智能简介

媒体对人工智能相关突破及其潜在影响进行了大量报道，其中包括**自然语言处理(Natural Language Processing，NLP)**和计算机视觉的进步，以及 GPT-4 等复杂语言模型的开发。特别值得一提的是生成式模型，该模型能够生成文本、图像和其他创造性

内容，且生成的内容往往与人类生成的内容无异，因此受到了广泛关注。这些模型还具有语义搜索、内容处理和分类等广泛的功能。这些功能可以实现自动化，从而节省成本，使人类的创造力发挥到前所未有的水平。

注意：

生成式人工智能是指能够生成新内容的算法，与传统的预测性机器学习或人工智能系统不同，它并非仅限于分析或操作现有数据。

基准测试能捕获模型在不同领域的任务性能，是模型开发的主要驱动因素。**大规模多任务语言理解(Massive Multitask Language Understanding，MMLU)**基准是一个包含 57 项任务的综合套件，横跨数学、历史、计算机科学和法律等多个领域。它是一种标准化的方法，用于评估大规模语言模型在零样本和少样本任务设置下的多任务性能和泛化能力。大规模多任务语言理解基准的重要性在于它提供了一个具有挑战性的、多方面的测试，以检验模型对各种主题的理解和解决问题的能力。它允许对不同的大规模语言模型进行系统比较，并跟踪在开发具有鲁棒性语言理解和推理能力的模型方面所取得的进展，而不仅仅局限于狭窄的领域。

图 1.1 的灵感来自 Stephen McAleese 在 LessWrong 上发表的一篇名为“GPT-4 Predictions”的博文，文章显示了大规模语言模型在**大规模多任务语言理解**基准测试中的进步。

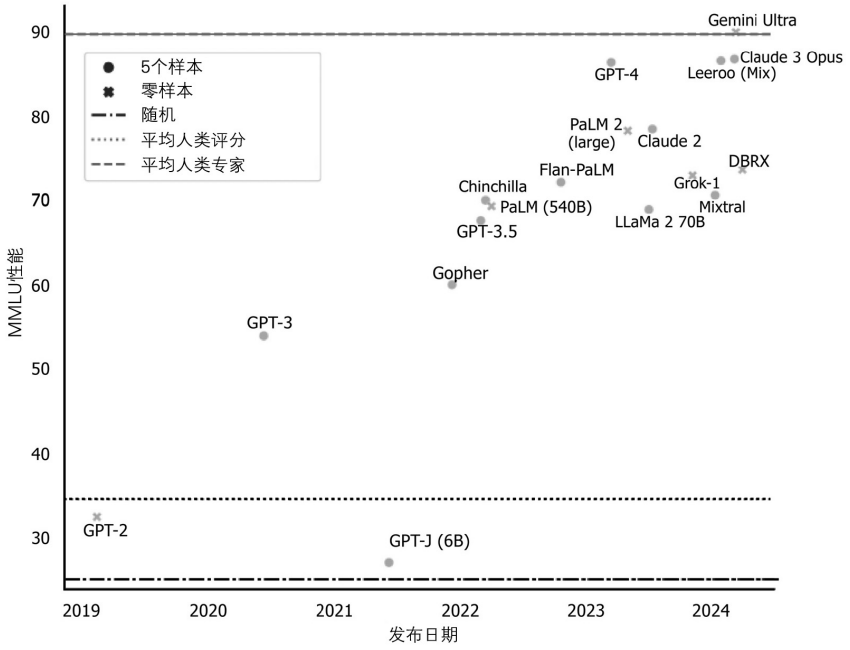

图 1.1　大规模语言模型在大规模多任务语言理解基准测试中的平均性能

注意：

请注意，由于这些结果都是自我报告的，并且是通过 5 个样本或 0 个样本条件下获得的，因此应谨慎对待。大多数基准结果来自 5 个样本(用“o”表示)。对于一部分结果，如 GPT-2、PaLM 和 PaLM-2 的结果，则是指零样本(“x”)。

从图 1.1 中可以看出，近年来大规模多任务语言理解基准有了显著提高。特别是它强调了 OpenAI 通过公共用户界面提供的模型的提升，尤其是从 GTP-2 到 GPT-3，从 GPT-3.5 到 GPT-4 版本之间的改进。

图 1.1 中显示的是模型的大规模多任务语言理解基准性能。零样本是指只给模型提示问题，而 5 个样本是指给模型额外的 5 个问答示例。根据“Measuring Massive Multitask Language Understanding”(Hendrycks 等，2023 年修订)，这些额外的示例占到性能的 20% 左右。

在 Claude3、GPT-4 和 Gemini 中，很难明确宣布最强的大规模语言模型，因为它们的性能似乎非常匹配，而且在不同任务中也各不相同。最终，最强大规模语言模型的选择可能取决于具体的用例和要求，包括其成本。

这些模型之间存在一些差异，它们的训练方式也可能导致性能差异，例如规模、指令微调、注意力机制的调整以及训练数据的选择。首先，从 15 亿(GPT-2)到 1 750 亿(GPT-3)再到超过 1 万亿(GPT-4)的大规模参数增长使模型能够学习更复杂的模式。然而，2022 年初的另一个重大变化是根据人类指令对模型进行训练后微调，即通过提供示范和反馈来教模型如何执行任务。

在基准测试中，一些模型最近开始比一般人类评分者表现更好，但总体上仍未达到人类专家的表现。人类工程学的这些成就令人印象深刻，但应该注意的是，这些模型的性能取决于领域，在小学数学单词问题的 GSM8K 基准上，大多数模型的性能仍然很差。随着像 OpenAI 的 GPT 这样的人工智能模型不断改进，它们可能成为需要各种知识和技能的团队不可或缺的资产。

你可以把像 GPT 4 或 Claude 3 这样强大的大规模语言模型视为一个多面手，它不辞辛劳地工作，不要求报酬(除了订阅费或 API 费用)，在数学和统计学、宏观经济学、生物学和法律(该模型在统一律师资格考试中表现出色)等学科能提供合格的帮助。随着这些人工智能模型更加熟练和易于使用，它们很可能在塑造未来的工作和学习方面发挥重要作用。

通过提高知识获取和适应的便利性，这些模型有可能为各行各业的人们带来新机遇，创造公平的竞争环境。这些模型在需要高水平的推理和理解的领域已显示出潜力，不过进展取决于所涉任务的复杂程度。

至于图像生成式模型，在视觉内容创作及计算机视觉任务(如目标检测、分割、字幕等)方面已经突破了界限。

下面厘清术语，详细解释生成式模型、人工智能、深度学习和机器学习的含义。

1.1.1 什么是生成式模型

在大众媒体中，谈到这些新模型时，常常使用“人工智能”这一术语。然而，在理论和应用研究领域里，人们经常开玩笑说，人工智能只不过是机器学习的一种花哨的说法，或者说人工智能就是穿着西装的机器学习，如图 1.2 所示。

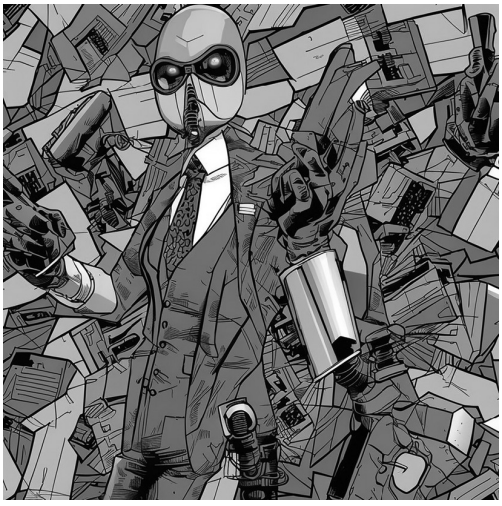

图 1.2　穿着西装的机器学习。由 replicate.com 上的一个模型生成(基于 Stable Diffusion v2.1)

明确区分生成式模型、人工智能、机器学习、深度学习和语言模型等术语是值得的。

- **人工智能(Artificial Intelligence，AI)**是一个广泛的计算机科学领域，其重点是创建能够自主推理、学习和动作的智能体。
- **机器学习(Machine Learning，ML)**是人工智能的一个子集，侧重于开发能够从数据中学习的算法。
- **深度学习(Deep Learning，DL)**是一种使用多层的深度神经网络，从数据中学习复杂模式机制的机器学习算法。
- **生成式模型(Generative Model)**是一种机器学习模型，可以根据从输入数据中学到的模式生成新数据。
- **语言模型(Language Model，LM)**是用于预测自然语言序列中单词的统计模型。一些语言模型利用深度学习，在海量数据集上进行训练，成为**大规模语言模型**(LLM)。

图 1.3 说明了大规模语言模型如何将神经网络等深度学习技术与语言建模中的序列建模目标相结合，并实现超大规模的应用。

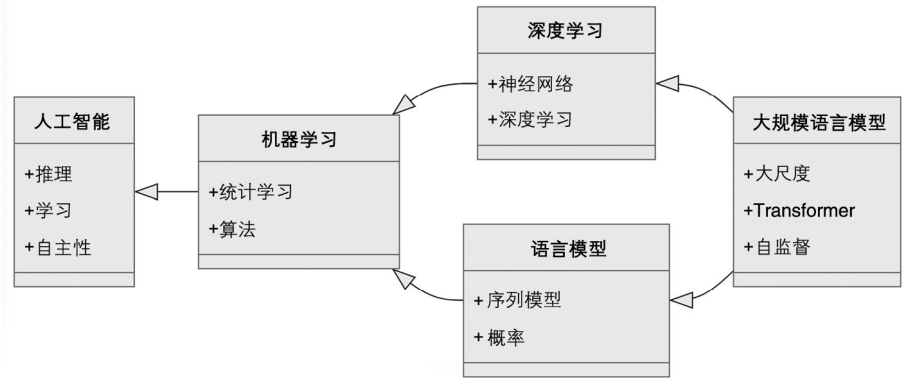

图 1.3　不同模型的类图。大规模语言模型代表深度学习技术与语言建模目标的交叉点

生成式模型是一种强大的人工智能，可以生成与训练数据相似的新数据。生成式人工智能模型已经取得了长足的进步，能够利用数据中的模式从零开始生成新的示例。这些模型可以处理不同的数据模式，并可用于不同的领域，包括文本、图像、音乐和视频。主要的区别在于，生成式模型可以合成新数据，而不仅仅是做出预测或决策。这使得生成文本、图像、音乐和视频等的应用成为可能。

在真实数据稀缺或受到限制的情况下，生成式模型有助于生成合成数据来训练人工智能模型。这种数据生成方式可以降低打标签的成本，提高训练效率。微软研究院在训练其 phi-1 模型时采用了这种方法(*Textbooks Are All You Need*，2023 年 6 月)，他们使用 GPT-3.5 创建了合成 Python 教科书和练习。

不同领域的快速进步显示了生成式人工智能的潜力。在行业内，人们对人工智能的能力及其对业务操作的潜在影响越来越感到兴奋。但是，在第 10 章中将讨论一些关键挑战，如数据可用性、计算要求、数据偏差、评估困难、潜在滥用和其他社会影响，这些都是未来需要解决的问题。

生成式人工智能广泛应用于生成三维图像、头像、视频、图表和插图，用于虚拟现实或增强现实、视频游戏图形设计、徽标创建、图像编辑或增强。这里最流行的模型类别是以文本为条件的图像合成，特别是文本到图像的生成。如前所述，本书将重点讨论大规模语言模型，因为它们具有最广泛的实际应用，但也会涉及图像模型，因为它们有时也非常有用。

下面深入探讨一下这一进展，并提出这样一个问题：为什么现在会出现这种情况？是什么条件使这一进展成为可能？

1.1.2　为什么是现在

生成式人工智能的成功有以下几个因素。

- **改进的算法**
- **计算力和硬件设计的长足进步**

- **大量标注数据集的可用性**
- **活跃的合作研究社区**

此外，开发更复杂的数学和计算方法对生成式模型的发展起到了至关重要的作用。20 世纪 80 年代提出的反向传播算法就是这样一个示例。它提供了一种有效训练多层神经网络的方法。

21 世纪初，随着研究人员开发出更复杂的架构，神经网络开始重新流行起来。然而，深度学习(一种具有多层的神经网络)的出现标志着这些模型性能和能力的重大转折。有趣的是，虽然深度学习的概念已存在一段时间，但生成式模型的发展和扩展与硬件的显著进步密切相关，特别是**图形处理器(Graphics Processing Units，GPU)**，它在推动更深层模型发展中发挥了重要作用。这是因为深度学习模型的训练和运行需要大量的计算力。这涉及处理能力、内存和磁盘空间等各个方面。

一旦大规模语言模型变大，其能力就会发生巨大变化。一个模型的参数越多，它捕捉单词和短语之间知识关系的能力就越强。举一个简单的例子来说明这些高阶相关性，一个大规模语言模型可以学习到：如果单词“cat”前面有“chase”，那么单词“dog”后面更有可能出现“cat”，即使中间还有其他单词。一般来说，一个模型的困惑度越低，它的表现就越好，例如在回答问题方面。

特别是，在拥有 20 亿～70 亿个参数的模型中，似乎出现了新的能力，例如能够生成诗歌、代码、脚本、音乐作品、电子邮件和信件等格式的创意文本，甚至能够以信息丰富的方式回答开放式和具有挑战性的问题。

1.2 了解大规模语言模型

大规模语言模型是一种擅长理解和生成人类语言的深度神经网络。这些模型可实际应用于内容创建和 NLP 等领域，其最终目标是创建能够理解和生成自然语言文本的算法。

目前的新一代大规模语言模型(如 GPT-4 等)是一种深度神经网络架构，它利用 Transformer 模型，通过对大量文本数据使用无监督学习进行预训练，使模型能够学习语言模式和结构。模型的发展日新月异，能够创建适用于各种下游任务和模式的通用基础人工智能模型，最终推动各种应用和行业的创新。

作为对话界面(聊天机器人)，最新一代大规模语言模型的显著优势在于，即使是在开放式对话中，它们也能生成连贯并与上下文相符的响应。根据前面的单词反复生成下一个单词，模型生成的文本流畅连贯，与人类生成的文本往往无法区分。

语言建模，乃至更广泛的 NLP，其核心都在很大程度上依赖于表征学习的质量。生成式语言模型对其训练过的文本信息进行编码，并根据学习结果生成新的文本，从而完成文本生成的任务。

注意：

***表征学习*(representation Learning)**是指模型通过学习原始数据的内部表征来执行机器学习任务，而不是仅仅依赖于工程特征提取。例如，基于表征学习的图像分类模型可以根据边缘、形状和纹理等视觉特征来表征图像。模型不会被明确告知要寻找哪些特征，而是学习原始像素数据的表征，从而帮助它做出预测。

最近，大规模语言模型已被应用于论文生成、代码开发、翻译和理解基因序列等任务。更广泛地说，语言模型的应用涉及多个领域。

- **问题回答：**人工智能聊天机器人和虚拟助理可以提供个性化和高效的帮助，缩短客户支持的响应时间，从而提升客户体验。这些系统可用于餐厅预订和机票预订等特定场景。
- **自动摘要：**语言模型可以创建文章、研究论文和其他内容的简明摘要，使用户能够快速使用和理解信息。
- **情感分析：**通过分析文本中的观点和情感，语言模型可以帮助企业更有效地理解客户的反馈和意见。
- **主题建模：**大规模语言模型可以发现文档语料库中的抽象标题和主题，可以识别词簇和潜在的语义结构。
- **语义搜索：**大规模语言模型可以理解单个文档中的含义。它使用 NLP 解释单词和概念，以提高搜索相关性。
- **机器翻译：**语言模型可以将文本从一种语言翻译成另一种语言，为企业的全球扩张提供支持。新的生成式模型可以与商业产品(如谷歌翻译)相媲美。

尽管取得了令人瞩目的成就，但语言模型在处理复杂的数学或逻辑推理任务时仍有局限性。目前还不能确定，不断扩大语言模型的规模是否一定能带来新的推理能力。此外，众所周知，大规模语言模型会根据上下文返回最有可能的答案，这有时会产生捏造的信息，称为幻觉。这既是一个功能，也是一个缺陷，因为它凸显了大规模语言模型的创造潜力。

第 5 章将讨论幻觉问题，现在，让我们更详细地讨论大规模语言模型的技术背景。

1.2.1　GPT 模型是如何工作的

Transformer 是一种深度学习架构，于 2017 年由谷歌和多伦多大学的研究人员首次推出(文章名为“Attention Is All You Need”，Vaswani 等)，它由自注意力和前馈神经网络组成，能够有效捕捉句子中的单词关系。

2018 年，研究人员通过创建生成式预训练 Transformer(Generative Pre-Trained Transformer，GPT)(见 *Improving Language Understanding by Generative Pre-Training*；Radford 等)，将 Transformer 提升到了一个新的水平。这些模型通过预测序列中的下一

个单词来进行训练，就像一个大型的猜词游戏，帮助它们掌握语言模式。经过这种预训练过程后，GPT 可以针对翻译或情感分析等特定任务进一步完善。这就将无监督学习(预训练)和有监督学习(微调)结合起来，从而在各种任务中取得更好的性能。也降低了大规模语言模型的训练难度。

1. Transformer

基于 Transformer 的模型优于以往的方法，例如使用循环神经网络，特别是长短期记忆(LSTM)网络。LSTM 等循环神经网络的记忆有限，这对于长句子或复杂的想法来说可能是个问题，因为早期的信息仍然是相关的。

Transformer 的工作方式与此不同，这意味着它们可以利用完整的上下文，并在处理句子中的更多单词时不断学习和完善自己的理解。这种在整个句子中充分利用上下文的能力，可以提高翻译、摘要和问题回答等任务的性能。该模型可以捕捉到长句的细微差别和词语之间的复杂关系。从本质上讲，Transformer 取得成功的一个关键原因是，与循环神经网络等其他模型相比，它能够在长序列中更好地保持性能。

Transformer 模型架构具有编码器-解码器结构，其中编码器将输入序列映射到隐藏状态序列，解码器将隐藏状态映射到输出序列。隐藏状态表示不仅考虑单词的固有含义(语义值)，还考虑单词在序列中的上下文。

编码器由相同的层组成，每个层有两个子层。输入嵌入通过注意力机制传递，第二个子层是一个全连接的前馈网络。每个子层之后都有一个残差连接和层归一化。每个子层的输出是子层输入与输出的总和，然后进行归一化处理。

解码器利用这些编码信息，结合之前生成的项的上下文，逐个生成输出序列。解码器与编码器的模块相同，也有两个子层。

此外，解码器还有第三个子层，该子层对编码器堆栈的输出执行**多头注意力(Multi-Head Attention，MHA)**。解码器还使用残差连接和层归一化。解码器中的自注意子层经过了修改，目的是防止位置编码影响到后续位置。这种掩码加上输出嵌入偏移一个位置，用来确保对位置 i 的预测只能依赖于小于 i 的已知输出位置。

Transformer 的成功得益于以下架构特点，如图 1.4 所示。

- **位置编码**：由于 Transformer 不是按顺序处理单词，而是同时处理所有单词，因此它缺乏单词顺序的概念。为了解决这一问题，使用位置编码将单词在序列中的位置信息添加到模型中。这些编码被添加到代表每个单词的输入嵌入中，从而使模型能够考虑单词在序列中的顺序。
- **层归一化**：为了稳定网络的学习，Transformer 使用了层归一化技术。该技术在特征维度上(而不是在批归一化中的批量维度)对模型的输入进行归一化，从而提高学习的整体速度和稳定性。
- **多头注意力**：Transformer 不是一次应用注意力，而是多次并行应用，这提高了模型关注不同类型信息的能力，从而捕获更丰富的特征组合。

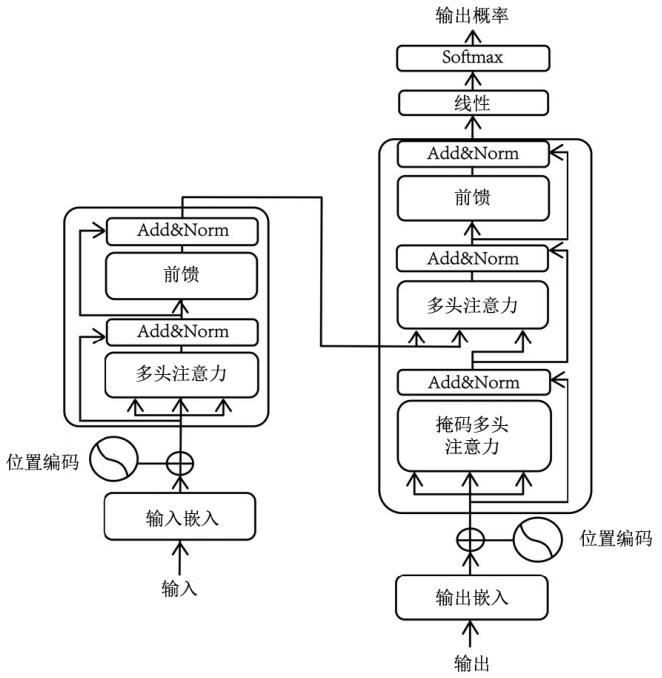

图 1.4　Transformer 架构

注意力机制背后的基本思想是，根据当前位置与所有其他位置之间的相似性，计算与输入序列中每个位置相关的值(通常称为值或内容向量)的加权和。这个加权和被称为上下文向量，然后被用作模型后续层的输入，使模型在解码过程中能够选择性地关注输入的相关部分。

为了增强注意力机制的表现力，通常会将其扩展到多个所谓的“头”，每个“头”都有自己的一组查询、键和值向量，使模型能够捕获输入表征的各个方面。然后，每个头的单个上下文向量会以某种方式进行串联或组合，形成最终输出。

早期的注意力机制与序列长度(上下文大小)成二次方关系，因此不适用于长序列设置。为了缓解这一问题，人们尝试了不同的机制。许多大规模语言模型都使用了某种形式的**多查询注意力(Multi-Query Attention，MQA)**，包括 OpenAI 的 GPT 系列模型、Falcon、SantaCoder 和 StarCoder。

多查询注意力是多头注意力的扩展，在多查询注意力中注意力计算被多次复制。多查询注意力提高了语言模型在各种语言任务中的性能和效率。通过从某些计算中移除头注意力并优化内存使用，与没有多查询注意力的基线模型相比，多查询注意力可使推理任务的吞吐量提高 11 倍，延迟降低 30%。

LLaMa 2 和其他一些模型使用了**分组查询注意力(Grouped-Query Attention，GQA)**，这是自回归解码中应用的一种实践，可以缓存序列中前一个词元的键(K)和值(V)对，从而加快注意力的计算速度。然而，随着上下文窗口或批次大小的增加，多头注意

力模型中与 KV 缓存大小相关的内存成本也会显著增加。为了解决这个问题，可以在多个头之间共享键和值预测，而不会降低性能。

为了提高效率，还提出了许多其他方法，如稀疏、低秩自注意力和潜在瓶颈等。其他研究还试图将序列扩展到固定输入大小之外的方法中；transformer-XL 等架构通过存储已编码句子的隐藏状态来重新引入递归，以便在下一个句子的后续编码中利用。

结合这些架构特点，GPT 模型可以成功地处理理解和生成人类语言和其他领域文本的任务。绝大多数大规模语言模型都是 Transformer，后文还将遇到许多其他更先进的模型，包括图像、声音和 3D 对象模型。

顾名思义，GPT 的特殊性在于预训练。下面看看这些大规模语言模型是如何训练的！

2. 预训练

Transformer 的训练分为两个阶段，采用无监督预训练和区别特定任务微调相结合的方法。预训练的目标是学习一种通用的表征，以适应各种任务。

无监督预训练可以遵循不同的目标。在由 Devlin 等(2019)于“BERT: Deep Bidirectional transformer for Language Understanding”中引入的**掩码语言建模(Masked Language Modeling，MLM)**中，输入被掩码，模型试图根据非掩码部分提供的上下文预测缺失的词元。例如，如果输入句子是“The cat [MASK] over the wall，”理想情况下，模型将学习预测被掩码的“jumped”(跳)。

在这种情况下，训练目标是根据损失函数来最小化预测与掩码词元之间的差异。然后根据比较结果对模型参数进行迭代更新。

负对数似然(Negative Log-Likelihood，NLL)和**困惑度(Perplexity，PPL)**是用于训练和评估语言模型的重要指标。负对数似然是机器学习算法中使用的损失函数，旨在使正确预测的概率最大化。负对数似然越低，表明网络已经成功地从训练集中学习到了模式，因此可以准确预测训练样本的标签。值得一提的是，负对数似然是一个被限制在正区间内的值。

另一方面，困惑度是负对数似然的指数化，可用于更直观地了解模型的性能。困惑度值越小，表明网络从训练集中成功地学习到了模式，因此能准确地预测训练样本的标签。困惑度值越大，表明学习性能越差。直观地说，低困惑度意味着模型对下一个单词预测程度较低。因此，预训练的目标就是尽量降低困惑度，这意味着模型的预测结果与实际结果更加一致。

在比较不同的语言模型时，困惑度通常被用作各种任务的基准指标。它能让人了解语言模型的性能如何，较低的困惑度表示模型对其预测更有把握。因此，与其他复杂度较高的模型相比，困惑度较低的模型会被认为性能更好。

训练大规模语言模型的第一步是**词元化(tokenization)**。这个过程包括建立一个词汇表，将词元映射为唯一的数字表征，以便模型能对其进行处理，因为大规模语言模型是需要数字输入和输出的数学函数。

3. 词元化

词元化是指将文本划分成词元(单词或子单词)，然后将文本中的单词映射到相应整数列表的查找表，从而将词元转换为 ID。

在训练大规模语言模型之前，通常会将词元分析器(更准确地说，是其字典)与整个训练数据集进行匹配，然后冻结。需要注意的是，词元分析器不会产生任意整数。相反，它们会在特定范围内输出整数——从 0 到 *N*，其中 *N* 表示词元分析器的词汇量大小。

定义

词元：词元是字符序列的一个实例，通常由单词、标点符号或数字组成。词元是构建文本序列的基本元素。

词元化：词元化是指将文本划分成词元的过程。词元分析器会根据空白和标点符号将文本划分成单个词元。

举例说明：

考虑以下文本:

The quick brown fox jumps over the lazy dog!

这句话会被划分成以下词元:

[the、quick、brown、fox、jumps、over、the、lazy、dog、!]

每个单词和标点符号都是一个单独的词元。

许多词元分析器根据不同的原理工作，但模型中常用的词元分析器类型有**字节对编码(Byte-Pair Encoding，BPE)**、单词片段(WordPiece)和句子片段(SentencePiece)。例如，LLaMa 2 的 BPE 词元分析器将数字划分成单个数字，并使用字节来分解未知的 UTF-8 字符。总词汇量为 32000(表示 32KB)个词元。

有必要指出的是，大规模语言模型只能根据不超过其上下文窗口的词元序列生成输出。上下文窗口指的是大规模语言模型可以使用的最长词元序列的长度。大规模语言模型的典型上下文窗口大小约为 1000～10 000 个词元。

在预训练之后，一个重要步骤是如何通过微调或提示让模型为特定任务做好准备。下面来看看这种任务调节是怎么回事！

4. 调节

调节大规模语言模型指的是针对特定任务调整模型，包括微调和提示工程。

- **微调(Fine-tuning)**是指通过监督学习对特定任务进行训练，以修改预训练的语言模型。例如，为了更适合与人类聊天，模型需要在表述为自然语言指令形式的任务示例上进行训练(即指令微调)。为了进行微调，通常会使用**基于人类反馈的强化学习(Reinforcement Learning from Human Feedback，RLHF)**再次训练预训练模型，使其既能提供帮助又不产生危害。

- **提示工程(prompting techniques)**以文本形式将问题呈现给生成式模型。有很多不同的提示方法，从简单的问题到详细的说明。提示可能包含类似问题及其解决方案的示例。零样本提示不包含任何示例，而少样本提示则包括少量相关问题和解决方案的示例。

这些调节方法不断发展，在广泛的应用中变得更加有效和实用。提示工程和调节方法将在第 8 章中进一步探讨。

1.2.2 GPT 模型是如何发展的

GPT 模型的开发取得了长足的进步，OpenAI 的 GPT-n 系列在创建基础人工智能模型方面处于领先地位。一个主要驱动因素是模型参数的规模，但其他驱动因素也发挥了作用。

注意：

基础模型(有时也称为**基本模型**)是一种大模型，它在大量数据上进行大规模训练，因此可以适应各种下游任务。在 GPT 模型中，这种预训练是通过自监督学习完成的。

除了简单地扩大模型规模之外，最近的重点已经转向探索其他方法，以提高模型在大规模多任务语言理解等基准上的性能。重点关注的一个关键领域是训练数据的处理和质量。精心选择和筛选训练数据以确保其相关性、多样性和质量，可以显著影响模型的性能，尤其是在测试各种知识和推理能力的基准上。

另一个关键的创新领域是模型架构。例如，Mixtral 和 Leeroo 模型采用了专家混合方法，即针对不同任务专门设置模型参数的不同子集，从而可能提高性能和计算效率。

通过探索这些替代方法，同时继续努力扩大规模，该领域正在努力开发具有更鲁棒的语言理解能力和跨领域推理能力的语言模型。

模型训练的计算要求和成本一直很高，未来可能还会增加。大规模语言模型的计算成本足以让你的钱包哭泣。不过不用担心！在探索减轻负担的方法之前，先来探讨一下是什么让这些模型如此沉重：它们的规模！

1. 模型大小

大规模语言模型的训练语料规模一直在急剧增加。2018 年，OpenAI 推出了 GPT-1，在拥有 9.85 亿单词的 BookCorpus 上进行了训练。同年发布的 BERT 是在 BookCorpus 和英文维基百科的合并语料库上训练的，总计 33 亿单词。现在，大规模语言模型的训练语料库已达到数万亿词元。

OpenAI 一直对**模型**的技术细节守口如瓶，但有消息称，GPT-4 拥有约 1.8 万亿个参数，是 GPT-3 的 10 倍多。此外，OpenAI 能够通过使用**专家混合(Mixture of Expert，MoE)模型**使成本保持在合理水平，该模型由 16 位专家组成，每位专家有约 1110 亿个

参数。

显然，GPT-4 是在大约 13 万亿个词元上训练出来的。不过，这些词元并不是唯一的，因为算上了每个迭代周期中重复呈现的数据。对基于文本的数据训练了两个迭代周期，对基于代码的数据训练了 4 个迭代周期。对于微调后的数据集是由数百万行指令微调而来的。另一个传言是，OpenAI 可能会对 GPT- 4 的推理应用推测解码，即一个较小的模型(Oracle 模型)可能会预测大模型的响应，而这些预测响应可以通过将其输入大模型来帮助加快解码速度，从而跳过词元。这是一种有风险的策略，因为根据 Oracle 响应的置信度阈值，质量可能会下降。

语言模型规模的扩大是其令人印象深刻的性能提升的主要推动力，谷歌的 Gemini 等模型不断突破规模和能力的极限。图 1.5 说明了大规模语言模型的增长情况。

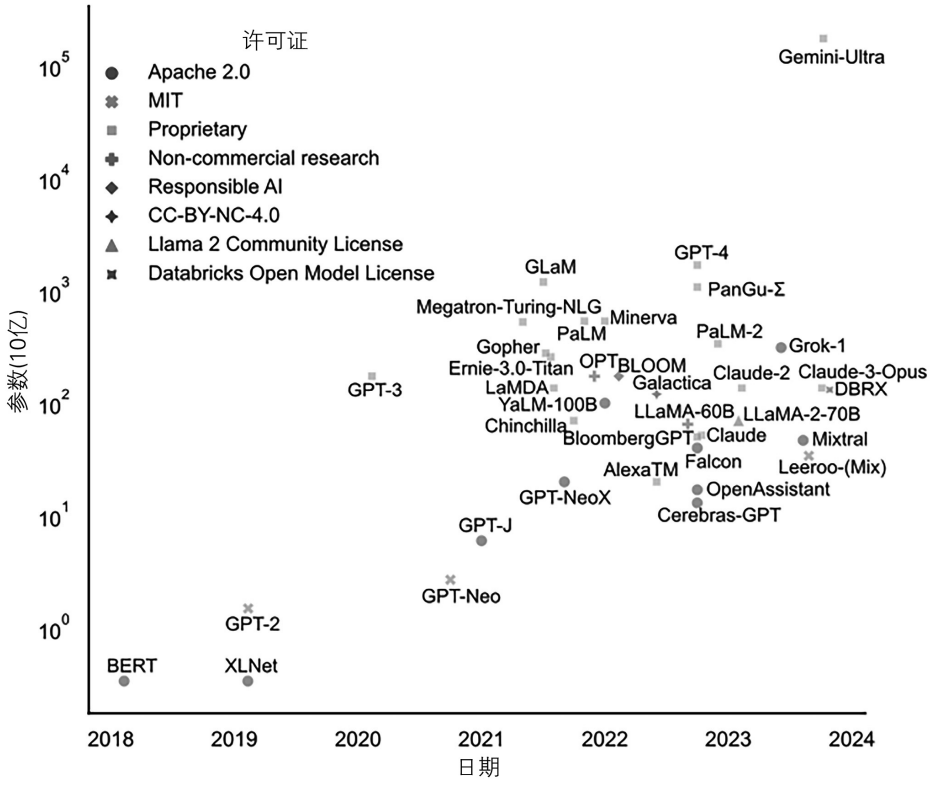

图 1.5　从 BERT 到 GPT-4 的大规模语言模型规模(参数量)、许可证。对于专有模型，参数规模通常是估算值

从图 1.5 中描述的历史进程可以看出，大规模语言模型的规模在不断扩大，参数数量也在不断增加。这一趋势与机器学习中观察到的更广泛的模式相一致，在机器学习中，提高模型性能往往需要扩大模型规模。Kaplan 等在 2020 年发表的 OpenAI 论文“Scaling laws for neural language models”中讨论了缩放法则和参数选择问题。

他们发现了一种幂律关系，表明大规模语言模型性能的提高与数据集规模、模型规模的增加成正比。具体来说，要想将性能提高某个系数，数据集或模型的规模以幂律形式呈指数增长。为获得最佳结果，这两个要素应同时扩展，从而防止模型训练和性能出现潜在瓶颈。

除数据集和模型大小外，考虑训练预算也很重要，因为它对训练过程的效率和结果有重大影响。训练预算包括分配给模型训练的计算力和时间等因素。这一指标可替代以迭代周期为单位的训练衡量标准，在确定停止训练的最佳时间点时更具灵活性和精确性。鉴于大规模语言模型的复杂性和广泛的训练要求，精确定位收敛点可能具有挑战性。因此，训练预算在有效管理资源，同时努力实现最高模型性能方面起着至关重要的作用。

DeepMind 的研究人员 Hoffmann 等在 2022 年发表的“An empirical analysis of compute-optimal large language model training”一文中分析了大规模语言模型的训练计算和数据集大小。他们得出结论：根据缩放法则，大规模语言模型在计算预算与数据集规模之间存在训练不足的情况。

他们预测，如果规模更小、训练时间更长，大模型将表现更出色。他们通过对比拥有 700 亿个参数的 Chinchilla 模型和由 2800 亿个参数组成的 Gopher 模型在基准测试中的结果来验证他们的预测。

不过，最近微软研究院的一个团队对这些结论提出了质疑，并让所有人大吃一惊(“Textbooks Are All You Need”，Gunaseka 等，2023 年 6 月)。他们发现，在高质量数据集上训练的小型网络(3.5 亿个参数)的性能非常具有竞争力。第 6 章将再次讨论这一模型，并在第 10 章中探讨缩放的意义。

可能会看到，将性能与数据质量联系在一起的新的缩放法则，观察大规模语言模型的模型规模是否以与以往相同的速度增长，很有启发意义。这是一个重要问题，因为它决定了大规模语言模型的开发是否会牢牢掌握在大型组织手中。在一定规模下，可能会出现性能饱和，只有改变方法才能解决这一问题。不过，还没有看到这种平缓的趋势。

2. GPT 模型系列

GPT-3 在 3000 亿个词元上进行了训练，有 1750 亿个参数，这是深度学习模型前所未有的规模。GPT-4 是该系列中的最新版，但出于竞争和安全方面的考虑，它的规模和训练细节尚未公布。不过，不同的估计表明，它的参数数量为 2000 亿～5000 亿。OpenAI 首席执行官 Sam Altman 曾表示，GPT-4 的训练成本超过 1 亿美元。

ChatGPT 由 OpenAI 于 2022 年 11 月推出，是在早期 GPT 模型(尤其是 GPT-3)基础上开发的对话模型。它专为对话量身定制，采用人类角色扮演场景和实例相结合的方式，引导模型做出预期行为，并通过基于人类反馈的强化学习(RLHF)得到显著增强。RLHF 不是根据任务表现从预先设定的奖励中学习，而是利用人类的反馈来训练模型，以了解好(高奖励)和坏(低奖励)的响应是什么样的。事实证明，RLHF 能有效地使人工智能模型更符合人类的价值观和偏好，适用于对话智能体和计算机视觉等领域。

2023 年 3 月推出的 GPT-4 标志着在能力方面得到了进一步飞跃。GPT-4 在各种评估任务中表现出色，而且由于在训练过程中进行了 6 个月的迭代微调，对恶意或挑衅性查询的响应避免能力显著提高。

图 1.6 显示了不同模型迭代的时间轴：

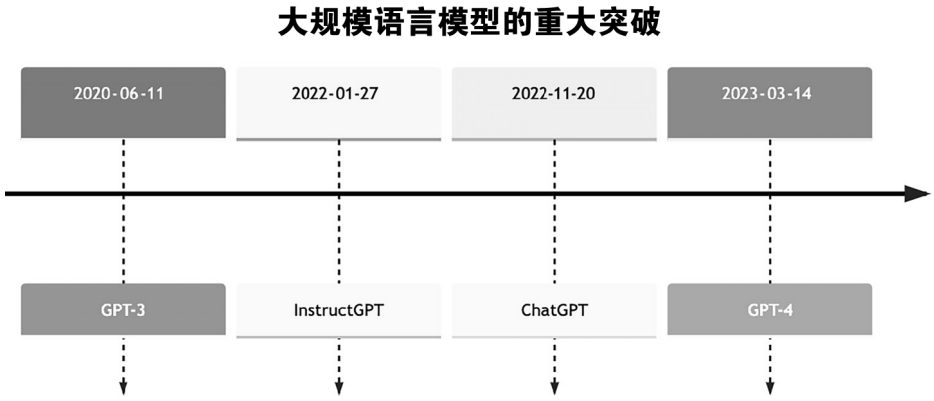

图 1.6　OpenAI GPT 模型系列的发展历程

GPT- 4 还有一个多模态版本，其中包含一个独立的视觉编码器，通过图像和文本数据进行训练，从而使模型能够阅读网页并转录图像和视频内容。

从图 1.5 中可以看出，除了 OpenAI 的模型之外，还有很多开源和闭源模型，其中一些在性能上已经接近 OpenAI 模型。

3. PaLM 和 Gemini

PaLM 2 于 2023 年 5 月发布，其训练重点是提高多语言和推理能力，同时提高计算效率。通过对不同计算缩放的评估，作者估算出了训练数据规模和参数的最佳缩放。PaLM 2 的规模更小，推理速度更快，效率更高，因此可以进行更广泛的部署，响应速度更快，交互节奏更自然。

对不同大小的模型进行的广泛基准测试表明，与前代 PaLM 相比，PaLM 2 在下游任务(包括多语言常识和数学推理、编码和自然语言生成)上的质量有了显著提高。

PaLM 2 还在各种专业语言水平考试中进行了测试，包括汉语(HSK 7-9 写作和 HSK 7-9 综合)、日语(J-Test A-C 综合)、意大利语(PLIDA C2 写作和 PLIDA C2 综合)、法语(TCF 综合)和西班牙语(DELE C2 写作和 DELE C2 综合)。这些考试的目标是测试 C2 级水平，根据 **CEFR(Common European Framework of Reference for Languages，欧洲语言共同参考框架)**，C2 级被视为精通或高级专业水平。

谷歌于 2023 年 12 月发布的 Gemini 是一个功能强大的多模态模型系列，可对图像、音频、视频和文本数据进行联合训练。最大版本的 Gemini Ultra 在语言、编码、推理和 MMMU(大规模多学科多模态)等多模态任务的 30 项基准测试中创造了新的先进结果。

它展示了令人印象深刻的跨模态推理能力、理解能力以及跨文本、图像和音频等不同模态的推理能力。

4. LLaMa 和 LLaMa 2

Meta AI 公司分别于 2023 年 2 月和 7 月发布了 **LLaMa** 和 **LLaMa 2** 系列模型(有多达 700 亿个参数)。这些模型极具影响力，因为社区能够在这些模型的基础上进行构建，开启了开源大规模语言模型的寒武纪爆发。LLaMa 引发了 Vicuna、Koala、RedPajama、MPT、Alpaca 和 Gorilla 等模型的诞生。自最近发布以来，LLaMa 2 已经激发了几个非常有竞争力的编码模型的诞生，如 WizardCoder。

大规模语言模型针对对话用例进行了优化，在发布之初，大规模语言模型在大多数基准测试中都优于其他开源聊天模型，而且根据人工评估，似乎与一些闭源模型不相上下。LLaMa 2 70B 模型在几乎所有基准测试中的表现都与 PaLM (5400 亿个参数)相当或更好，但 LLaMa 2 70B 与 GPT- 4 和 PaLM-2-L 之间仍有很大的性能差距。

LLaMa 2 是 LLaMa 1 的升级版，在新的公开数据组合上进行了训练。预训练的语料规模增加了 40%(2 万亿个词元)，模型的上下文长度增加了 1 倍，并采用了分组查询注意力。不同参数大小(70 亿、130 亿、340 亿和 700 亿)的 LLaMa 2 变体已经发布。虽然 LLaMa 是在非商业许可证下发布的，但 LLaMa 2 对公众开放，供研究和商业使用。

与其他开源和闭源模型相比，LLaMa 2-Chat 已进行了安全性评估。人类评测员在大约 2000 次对抗性提示(包括单轮提示和多轮提示)中对各代模型的安全违规行为进行了评测。

5. Claude 1 – 3

Claude、Claude 2 和 Claude 3 是 Anthropic 创造的人工智能助手。Claude 2 在乐于助人、诚实和减少偏见等方面对以前的版本进行了改进。主要的增强功能包括大量增加了上下文窗口，可容纳多达 20 万个词元，并在编码、摘要和长文档理解任务中表现出色。

最新发布的 Claude 3 是一个全新的大型多模态模型系列，包括旗舰 Claude 3 Opus(能力最强)、Claude 3 Sonnet(兼顾技能和速度)和 Claude 3 Haiku(速度最快、成本最低)。凭借视觉功能，它们在包括大规模多任务语言理解在内的各种基准测试中均表现出强劲的性能。值得注意的是，Claude 3 Opus 在聊天机器人竞技场排行榜上超过了 OpenAI 的 GPT-4，同时表现出更高的多语言流畅性。

6. 专家混合(MoE)

最近，MoE 模型成功地以较低的资源使用率实现了高性能。Mistral AI 公司的 Mixtral 8x7B 是一种稀疏 MoE 模型，在各种基准测试中的表现均优于或不逊色于 Llama 2 70B 和 GPT-3.5，尤其是在数学、代码生成和多语言任务方面。其指令调整版本 Mixtral 8x7B-Instruct 在人类基准测试中超过了其他几个著名模型，如 GPT-3.5 Turbo 和

Claude-2.1。

Grok-1 是 xAI 从头开始训练的一个拥有 3140 亿个参数的 MoE 大规模语言模型，以 Apache 2.0 许可证发行。xAI 使用基于 JAX 和 Rust 的定制训练堆栈训练 Grok-1，展示了他们在开发尖端大规模语言模型方面的专业知识。Leeroo 使用 Leeroo 编排器提出了一种整合多个大规模语言模型的架构，以创建一种新的先进模型。它以更低的计算成本实现了与 Mixtral 不相上下的性能，甚至在大规模多任务语言理解基准测试中超过了 GPT-4 的准确性，并进一步降低了成本。

DBRX 是 Databricks 的开放式大规模语言模型，它建立了跨标准基准的开放式大规模语言模型。它超越了 GPT-3.5，并与 Gemini 1.0 Pro 竞争。作为一种代码模型，其性能优于 CodeLLaMA-70B 等专用模型。DBRX 采用细粒度 MoE 架构，推理速度比 LLaMA2-70B 快 2 倍，参数数量比 Grok-1 少 40%，从而提高了效率。在 Mosaic AI 模型服务上，它生成文本的速度高达 150 个词元/秒/用户。在相同质量的情况下，训练的计算效率是密集模型的 2 倍，在总体计算量减少至原来的 1/4 的情况下，达到了以前 MPT 模型的质量。其他为大规模语言模型进步做出贡献的著名模型包括 DeepMind 的 Chinchilla、Meta 的 OPT、Google 的 Gopher、Hugging Face 的 BLOOM，以及 EleutherAI 的 GPT-NeoX 等研究小组的各种模型。

1.2.3　如何使用大规模语言模型

可以通过 OpenAI、Google 和 Anthropic 的网站或 API 访问它们的大规模语言模型。如果想在笔记本电脑上尝试其他大规模语言模型，开源大规模语言模型是一个不错的开始。这里有整个模型所需要的一切！你可以通过 Hugging Face 或其他提供商访问这些模型(参见第 3 章)。你甚至可以下载这些开源模型，对其进行微调或从头开始训练。第 8 章将介绍如何对模型进行微调。

大规模语言模型的许可证不同，对如何使用、修改和进一步开发大规模语言模型以用于商业或研究目的有很大影响。一些用于训练、训练数据集和权重本身的代码已向社区开放，供本地运行、调查、进一步开发、微调和改进。其他模型则被隐藏在 API 之后，其性能背后的秘密只能靠谣传和猜测。以下是一些关键许可证类型及其影响的细分。

开放源码许可证(如 Apache 2.0、MIT)：

- 允许为商业和非商业目的自由使用、修改和再发布。
- 允许创作衍生作品并将模型集成到产品/服务中。
- 研究机构和商业实体可以在这些模型的基础上进行构建和扩展。
- 例如 BERT、Mistral。

非商业许可证(如 CC-BY-NC-4.0，非商业研究)：

- 仅允许为非商业研究目使用和修改。
- 商业实体不能直接使用或将这些模型集成到产品/服务中。
- 研究人员可以在学术环境中研究、评估和构建这些模型。

- 例如 Galactica、OPT、Llama 60B。

专有许可证：

- 模型是闭源的，不能自由使用、修改或再发布。
- 商业实体保留完全控制权，可将模型作为产品/服务盈利。
- 研究机构可为评估/基准测试目的有限度地使用模型。
- 例如 GPT - 4、Claude、Gemini。

新许可证，如 Databricks 开放模型许可证和 Llama 2 社区许可证：

- 允许为商业和非商业目的使用、修改和创作衍生作品。
- 可能对再发布、赔偿或使用跟踪设置某些条件。
- 在开源访问和商业利益之间取得平衡。

一般来说，开放源码许可证促进了模型的广泛采用、合作和创新，有利于研究和商业开发。专有许可证给了公司独家控制权，但可能会限制学术研究的进展。非商业许可证在促进研究的同时，也限制了商业使用。新许可证旨在减少这些权衡。

1.3 节将回顾最先进的文本条件图像生成方法，将重点介绍该领域迄今为止取得的进展，也会讨论现有的挑战和未来的方向。

1.3 什么是文本到图像模型

文本到图像模型是一种功能强大的生成式人工智能，可通过文本描述生成逼真的图像。这类模型在创意产业和设计领域有多种应用案例，可用于生成广告、产品原型、时尚图像和视觉效果。主要应用有以下几种。

- **文本条件图像生成：**根据文本提示创建原始图像，如“一幅花丛中的猫的画”。可用于艺术、设计、原型设计和视觉效果。
- **图像补全：**根据周围环境填补图像缺失或损坏的部分，可以恢复损坏的图像(去噪、去雾和去模糊)或编辑掉不需要的元素。
- **图像到图像的转换：**将输入图像转换为通过文本指定的不同风格或领域，如“让这张照片看起来像莫奈的画”。
- **图像识别：**大规模基础模型可用于图像识别，包括场景分类和目标检测，如检测人脸。

Midjourney、DALL-E 2 和 Stable Diffusion 等模型可以提供派生自文本输入或其他图像的创造性的逼真图像。这些模型的工作原理是在大型图像-文本对数据集上训练深度神经网络。所使用的关键技术是扩散模型，该模型从随机噪声开始，通过反复的去噪步骤，逐渐将噪声还原成图像。

Stable Diffusion 和 DALL-E 2 等流行模型使用文本编码器将输入文本映射到嵌入空间。文本嵌入被输入到一系列条件扩散模型中，这些模型在连续的阶段中对潜在图像进

行去噪和细化。模型的最终输出是与文本描述一致的高分辨率图像。

主要使用两类模型：**生成对抗网络(Generative Adversarial Networks，GAN)**和扩散模型。生成对抗网络模型(如 StyleGAN 或 GANPaint Studio)可以生成高度逼真的图像，但训练不稳定且计算成本高昂。这类模型由生成器和判别器两个网络组成，这两个网络在类似游戏的环境中相互竞争，生成器负责从文本嵌入和噪声中生成新图像，判别器负责估计新数据是真实数据的概率。随着这两个网络的竞争，生成对抗网络在生成真实图像和其他类型数据的任务中表现得越来越好。

训练生成对抗网络的设置如图 1.7 所示(摘自“A Survey on Text Generation Using Generative Adversarial Networks”，G de Rosa 和 J P. Papa，2022；https://arxiv.org/pdf/2212.11119.pdf)。

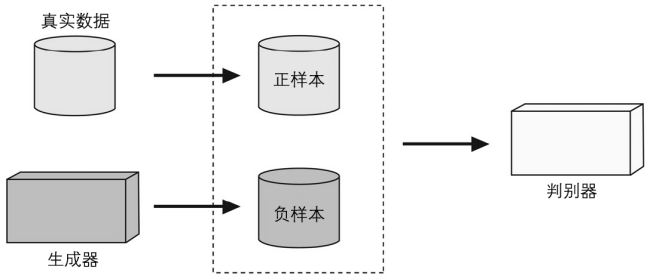

图 1.7　生成对抗网络训练

扩散模型在各种生成任务(包括文本到图像的合成)中很受欢迎，前景广阔。与以往的方法(如生成对抗网络)相比，扩散模型的优势是降低计算成本和减少序列误差积累。扩散模型的运行过程与物理学中的扩散过程类似。它们遵循**前向扩散过程(forward diffusion process)**，向图像中添加噪声，直到图像变得不再有特征，不再嘈杂。这一过程类似墨水滴入水杯中逐渐扩散的过程。

生成式图像模型的独特之处在于**后向扩散过程(reverse diffusion process)**，即模型试图从噪声、无意义的图像中恢复原始图像。通过迭代应用去噪变换，模型生成的图像分辨率越来越高，与给定的文本输入达成一致。最终输出的图像是根据文本输入修改过的图像。这方面的一个示例是 Imagen 文本到图像模型(谷歌研究院 2022 年 5 月发表的“Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”)，该模型结合了来自大规模语言模型的固定文本嵌入，在纯文本语料库中进行了预训练。文本编码器首先将输入文本映射到嵌入序列。一系列条件扩散模型将文本嵌入作为输入并生成图像。

图 1.8 展示了去噪过程(来源：用户 Benlisquare 通过维基共享资源提供)。

图 1.8 仅展示了 40 步生成过程中的部分步骤。你可以看到图像生成的各个步骤，包括使用去噪扩散隐式模型**(Denoising Diffusion Implicit Model，DDIM)**采样方法的 U-Net 去噪过程，该方法可反复去除高斯噪声，然后将去噪后的输出解码到像素空间。

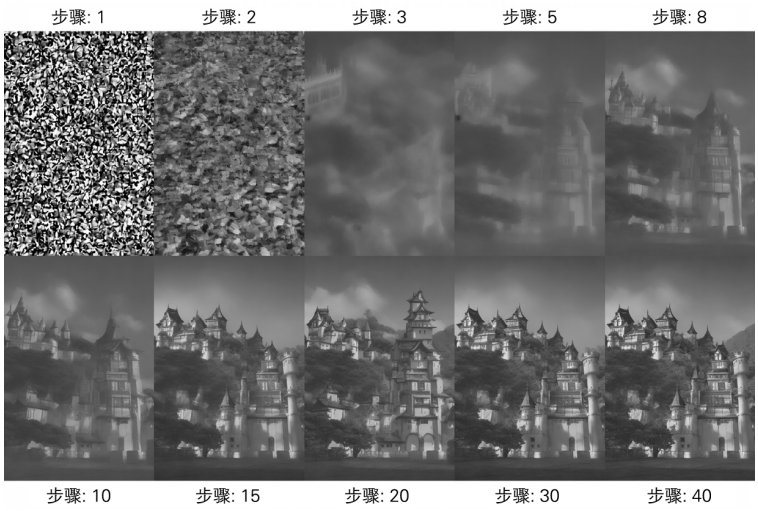

图 1.8　使用 Stable Diffusion V1-5 AI 扩散模型创建的日本欧式城堡

使用扩散模型，只需要对模型的初始设置或(在本例中的)数字求解器和采样器进行最小限度的修改，就能看到各种结果。虽然这些模型有时会产生惊人的结果，但不稳定性和不一致性是广泛应用这些模型的一个重大挑战。

Stable Diffusion 模型是由慕尼黑大学的 CompVis 小组开发的。与之前的(基于像素的)扩散模型相比，Stable Diffusion 模型大大减少了训练成本和采样时间。该模型可在配备普通 GPU(如 GeForce 40 系列)的消费级硬件上运行。通过在消费级 GPU 上根据文本创建高保真图像，Stable Diffusion 模型实现了访问的平民化。此外，该模型的源代码甚至权重都是在 CreativeML OpenRAIL-M 许可证下发布的，该许可证对重用、分发、商业化和改编不加限制。

值得注意的是，为了提高计算效率，Stable Diffusion 模型在潜在(低维)空间表征中引入了操作，以捕捉图像的基本属性。VAE 提供潜在空间压缩(在本文中称为感知压缩)，而 U-Net 执行迭代去噪。

Stable Diffusion 模型通过以下几个清晰的步骤根据文本提示生成图像。

(1) 在潜在空间中生成一个随机张量(随机图像)，作为初始图像的噪声。

(2) 噪声预测器(U-Net)同时接收具有潜在噪声的图像和提供的文本提示，并预测噪声。

(3) 模型从具有潜在噪声的图像中减去潜在噪声。

(4) 如图 1.8 所示，重复第(2)步和第(3)步进行一定次数的采样，如 40 次。

(5) VAE 的解码器组件将具有潜在噪声的图像转换回像素空间，提供最终的输出图像。

VAE 将数据编码到已学习到和较小的表征模型中。然后，这些表征可用于生成与训练所用数据类似的新数据(解码)。这种 VAE 要先经过训练。

注意：

U-Net 是一种流行的**卷积神经网络(Convolutional Neural Network，CNN)**，具有对称的编码器-解码器结构。它通常用于图像分割任务，但在 Stable Diffusion 中，它可以帮助引入和去除图像中的噪声。U-Net 将噪声图像(种子)作为输入，通过一系列卷积层对其进行处理，以提取特征并学习语义表征。

这些卷积层通常以收缩路径组织，在增加通道数量的同时降低空间维度。一旦收缩路径到达 U-Net 的瓶颈，U-Net 就会通过对称扩展路径进行扩展。在扩展路径中，转置卷积(也称为上采样或展开卷积)被应用于逐步上采样空间维度，同时减少通道数量。

为了在潜在空间本身(**潜在扩散模型**)中训练图像生成式模型，需要使用损失函数评估生成图像的质量。一种常用的损失函数是**均方差(Mean Squared Error，MSE)**损失，它量化了生成图像与目标图像之间的差异。模型进行了优化，以最小化这一损失，从而促使其生成的图像与需要输出的图像非常相似。

这种训练是在 LAION-5B 数据集上进行的，该数据集由数十亿图像-文本对组成，源自 Common Crawl 数据，其中包括来自 Pinterest、WordPress、Blogspot、Flickr 和 DeviantArt 等的数十亿图像-文本对。

图 1.9 说明了通过扩散从文本提示生成图像的过程。

(来源：Ramesh 等，“Hierarchical Text-Conditional Image Generation with CLIP Latents”, 2022; https://arxiv.org/abs/2204.06125)

充满活力的萨尔瓦多•达利的肖像画，半张机器人脸　　戴贝雷帽、穿黑色高领毛衣的柴犬

图 1.9　根据文本提示生成图像

总体而言，Stable Diffusion 和 Midjourney 等图像生成式模型利用前向扩散和后向扩散过程，将文本提示转化为生成式图像，并在较低维度的潜在空间中运行，以提高效率。但是，在文本到图像的用例中，如何对模型进行调节？

调节过程允许这些模型受到特定输入文本提示或输入类型(如深度图或轮廓)的影响，从而更精确地创建相关图像。然后，文本 Transformer 会对这些嵌入进行处理，并将其输入噪声预测器，引导其生成与文本提示一致的图像。

所有模式的生成式人工智能模型的全面介绍超出了本书的讨论范围。不过，我们可以简单介绍模型在其他领域的作用。

1.4 人工智能在其他领域的作用

生成式人工智能模型在声音、音乐、视频和三维形状等各种模式中都展现了令人印象深刻的能力。在音频领域，模型可以合成自然语音，生成原创音乐作品，甚至可以模仿说话者的声音、节奏和韵律的模式。

语音到文本系统可以将口头语言转换成文本[**自动语音识别(Automatic Speech Recognition，ASR)**]。在视频方面，人工智能系统可以根据文字提示创建逼真的视频片段，并进行复杂的编辑，如删除物体。三维模型可以根据图像重建场景，并根据文本描述生成复杂的物体。

生成式模型有很多种类型，可以处理不同领域的不同数据模式，如表 1.1 所示。

表 1.1 音频、视频和其他领域的模型

类型	输入	输出	示例
文本到文本	文本	文本	Mixtral，GPT-4，Claude 3，Gemini
文本到图像	文本	图像	DALL-E 2，Stable Diffusion，Imagen
文本到音频	文本	音频	Jukebox，AudioLM，MusicGen
文本到视频	文本	视频	Sora
图像到文本	图像	文本	CLIP，DALL-E 3
图像到图像	图像	图像	超分辨率，风格迁移，绘画填充
文本到编码	文本	编码	Stable Diffusion，DALL-E 3，AlphaCode，Codex
视频到音频	视频	音频	Soundify
文本到数学	文本	数学表达式	ChatGPT，Claude
文本到科学	文本	科学输出	Minerva，Galactica
算法发现	文本/数据	算法	AlphaTensor
多模态输入	文本/图像	文本，图像	GPT-4V

要考虑的模式组合还有很多，以上只是我遇到的一些。此外，还可以考虑文本的子类别，如文本到数学，即根据文本生成数学表达式，ChatGPT 和 Claude 等模型在这方面大放异彩；或文本到代码，即根据文本生成编程代码的模型，如 AlphaCode 或 Codex。一些模型专门用于科学文本，如 Minerva 或 Galactica；或专门用于算法发现，

如 AlphaTensor。

还有一些模型使用多种输入或输出模式。例如，OpenAI 的 GPT- 4V 模型(GPT- 4 版本)就展示了多模态输入的生成能力，该模型于 2023 年 9 月发布，可以同时接收文本和图像，并且具有比老版本更好的**光学字符识别(Optical Character Recognition，OCR)**，可以从图像中读取文本。图像可以翻译成描述性的单词，然后应用现有的文本过滤器。这降低了生成无限制图像标题的风险。

如表 1.1 所示，文本是一种常见的输入模式，可以转换为图像、音频和视频等各种输出。这些输出也可以转换回文本或在同一模式下转换。大规模语言模型推动了以文本为中心的领域快速发展。这些模型通过不同的模式和领域实现了各种功能。大规模语言模型是本书的重点，不过，我们也会偶尔介绍其他模型，尤其是文本到图像的模型。这些模型通常基于 Transformer 架构，通过自监督学习在海量数据集上进行训练。

快速的发展显示了生成式人工智能在不同领域的潜力。在业内，人们对人工智能的能力及其对业务运营的潜在影响越来越兴奋。但是，还有一些关键挑战需要解决，如数据可用性、计算要求、数据偏差、评估困难、潜在滥用及其他社会影响，第 10 章将讨论这些问题。

这些创新成果的基础是生成对抗网络、扩散模型和 Transformer 等深度生成式架构的进步。谷歌、OpenAI、Meta 和 DeepMind 等领先的人工智能实验室正在引领创新。

1.5　小结

随着计算力的提升，深度神经网络、Transformer、生成对抗网络和 VAE 能够比前几代模型更有效地模拟现实世界数据的复杂性，从而推动了人工智能算法的发展。本章探讨了深度学习、人工智能以及大规模语言模型和 GPT 等生成式模型的近代史，以及它们的理论基础，尤其是 Transformer 架构。本章还解释了图像生成式模型的基本概念，如 Stable Diffusion 模型，最后讨论了文本和图像以外的应用，如声音和视频。

第 2 章将利用 LangChain 框架探讨生成式模型(尤其是大规模语言模型)的工具化，重点关注基本原理、实现以及如何利用这一特殊工具开发和扩展大规模语言模型的能力。

1.6　问题

我认为，在阅读技术类书籍时，检查自己是否消化了书中的内容是一个好习惯。为此，我设计了几个与本章内容相关的问题。看看你能否回答出来。

1. 什么是生成式模型？
2. 生成式模型有哪些应用？
3. 什么是大规模语言模型？它有什么用处？

4. 如何提高大规模语言模型的性能？
5. 要使这些模型成为可能需要哪些条件？
6. 哪些公司和组织是开发大规模语言模型的主要参与者？
7. 如何获取大规模语言模型许可证？举例说明。
8. 什么是 Transformer？它由哪些部分组成？
9. GPT 代表什么？
10. Stable Diffusion 是如何工作的？