



第 1 章

ChatGPT 简介 与基本原理

2022 年 11 月 30 日，OpenAI 宣布正式发布 ChatGPT 3.5，其惊人的自然语言理解和生成能力犹如一块巨石投入平静的湖面，迅速激起了全社会的热情和关注，成为备受瞩目的全球话题，并开始在各行业被广泛应用。仅仅两个月，ChatGPT 的月活跃用户数便突破 1 亿大关，创下了消费者应用史上用户数增长最快的辉煌战绩。三个多月后，ChatGPT 4.0 横空出世，它不仅在自然语言理解、逻辑推理等方面相较 3.5 版本有了巨大提升，还拥有更加深入的“知识储备”，让世人不由赞叹人工智能技术的日新月异。而在我们的身边，也常常能在各种自媒体上看到诸如“最容易被 ChatGPT 取代的工作岗位”“ChatGPT：人工智能觉醒自我意识？”等抓人眼球的内容。

那么，ChatGPT 到底是什么？它是如何一步一步发展到今天的？它有着怎样的底层原理？又有着哪些应用场景？它如何赋能我们的数据分析工作？让我们带着这些问题一探究竟。

1.1 ChatGPT 简介

本节首先讲述 ChatGPT 的概念和发展历程，再讨论 ChatGPT 的特点，为后续介绍 ChatGPT 原理和应用做铺垫。

1.1.1 ChatGPT 前世今生

时至今日，ChatGPT 的爆火和“出圈”已经无须赘述，可以说大街小巷、老少妇孺间都有关于它的传说。那么，ChatGPT 到底是什么？我们不妨带着“我是谁？我从哪来？要到哪去？”的终极问题向 ChatGPT 提问，先听听它的自我介绍：

Question:

你好，ChatGPT，请问你是谁？你从哪来？要到哪去？

ChatGPT 的回答如下：

Answer:

你好！我是ChatGPT，一个由OpenAI开发的人工智能语言模型。我是基于GPT（Generative Pretrained Transformer）系列模型的一个实例，这是一种深度学习模型，专门用于理解和生成人类语言。

我从大量的文本数据中学习而来。OpenAI的研究人员使用了各种来源的文本数据来训练我，包括书籍、网站和其他类型的文本。通过这种方式，我学习了语言的语法、词汇和一些基本的世界知识。

我的主要目标是帮助人们解决问题和完成任务。我可以帮助回答问题、写作、编程、学习新的概念，甚至可以进行一些基本的情感支持。我还在不断学习和进化，希望能够更好地理解 and 满足用户的需求。

至于我要到哪里去，这取决于OpenAI的研究人员和用户的需求。我可能会继续学习新的知识和技能，也可能被用于新的应用场景。但无论如何，我都会致力于提供有用、准确和友好的帮助。

借由它的自我介绍，我们可以总结：ChatGPT 是由 OpenAI 开发的基于 GPT（生成式预训练转换器）模型的一个变体，它可以用语言交流形式实现与人的交互，是一款强大的 AI 聊天机器人。

目前，ChatGPT 已经发展到 ChatGPT 4.0 了，而 ChatGPT 4.0 和它的爆火当然不是像空中楼阁那样凭空出现的，而是建立在 OpenAI 对于语言模型，尤其 NLP 模型领域的持续研究与创造性探索之上的。我们可以大致将 ChatGPT 的发展历程分为四个阶段。

1. 初试宝刀

2018 年，OpenAI 推出第一代生成式预训练模型 GPT-1。这个模型采用了生成式预训练 Transformer 模型，这是一种采用自注意力机制的深度学习模型。GPT-1 采用了半监督学习模式，通过无监督学习的预训练，运用大量未标注数据增强模型的语言能力，再进行有监督的微调，与大规模数据集集成以提高模型在 NLP 任务中的表现。

2. 渐露锋芒

2019 年，OpenAI 推出了 GPT-2，这是一个更大的模型，拥有超过 15 亿个参数，并且可以生成达到人类编写水平的文本。GPT-2 的出现与性能提升，进一步验证了无监督学习的力量，即通过海量数据与大规模参数训练而成的 NLP 模型能够无须额外训练具备迁移到其他类别任务的能力。

3. 爆火“出圈”

2020 年，OpenAI 推出了 GPT-3，参数量达到了 1750 亿，GPT-3 删去微调步骤，直接输入自然文本作为指令，提升了 GPT 在阅读文本后可接续问题的能力以及任务主题的广泛性。而 2022 年 11 月推出的 GPT-3.5 版本接近人类语言反应能力，GPT-3.5 的主要杰作就是近期大火的 ChatGPT。它使用了微软 Azure AI 超级计算基础设施上的文本和代码数据进行训练，在训练参数上增加到 GPT-3 的 10 倍以上，延续了 OpenAI 对大规模数据的追求。此外，它颠覆性地使用大量人工标注数据与有人类反馈的强化学习，使得 ChatGPT 表现出出色的上下文对话能力甚至编程能力。

4. 步入多模态

2023 年 3 月 15 日，OpenAI 推出了 ChatGPT 4.0，这是一个使用前所未有的计算和数据规模进行训练的模型，参数量呈指数级增长，是目前为止功能最强大的模型。ChatGPT 4.0 在 GPT-3.5 的基础上支持了图像的输入，将语言到多模态的连通从可能变成了现实。引入多模态数据不

仅支持了各种“图像 - 语言”任务的完成，同时也利用图像数据提升了语言处理能力，使模型的性能提升不再仅依赖于大量的语言数据。

纵观 ChatGPT 系列的发展，我们可以看到 OpenAI 始终贯彻大规模数据与生成式预训练 Transformer 模型的技术路线，并且在追求更大规模的模型和更强大的能力的同时，也在不断探索新的训练方法和模型架构。由此，模型的规模和能力都在持续增长，对人类语言的理解和生成能力也在不断提升。

1.1.2 ChatGPT的特点

现在我们已经知道 ChatGPT 的发展脉络了，那么 ChatGPT 到底具备了哪些特点，才让大家如此热烈地追捧呢？

这里我们可以总结出 ChatGPT 的七大特点。

(1) 对话交互。这是 ChatGPT 最显而易见的特征，它以对话方式进行交互，你问我答、有问必答。

(2) 连续多轮互动。ChatGPT 可以根据上下文了解用户意图，并且能够记忆先前用户的对话信息，以回答某些假设性的问题，从而极大地提升了对话交互模式下的用户体验，完成多轮对话，真正像人类一样进行聊天交流。比如你可以先说“我喜欢吃巧克力”，然后再问“你能推荐一款吗？” ChatGPT 会记住你之前的对话，并给出推荐的巧克力，如下所示。

```
Question:
我喜欢吃巧克力。
Answer:
那真是太好了！巧克力是一种非常美味的甜点，它不仅口感丰富，而且还含有能够提振人的心情的化合物。
你有特别喜欢的巧克力品牌或者类型吗？比如黑巧克力、牛奶巧克力或者是带有坚果的巧克力？
Question:
你可以推荐一款给我吗？
Answer:
当然可以。如果你喜欢口感丰富且品质上乘的巧克力，我会推荐你试试Lindt的瑞士巧克力。Lindt是一家瑞士的巧克力品牌，它的巧克力口感细腻、味道丰富，非常受欢迎。特别是它的Lindt Excellence系列，有多种可口的口味，比如70%的黑巧克力、橙味黑巧克力，以及带有海盐、辣椒或者杏仁的巧克力，都是非常值得一试的。
当然，巧克力的口味因人而异，我建议您尝试多种口味，找到最适合你的那一款。
```

(3) 多场景应用。ChatGPT 能够完成多种日常任务，是一个功能强大的贴心助手，无论你是需要写一封邮件、创作一个视频脚本还是设计一段广告文案，甚至进行语言翻译或编写代码，ChatGPT 都能帮你完成。

(4) 问答系统。ChatGPT 是聊天机器人和问答系统的结合，还可以帮助我们完成一些问答的功能。你可以向它提各种问题，比如“第一次鸦片战争发生在哪一年”或者“ $E=mc^2$ 是什么意思？”它都能给出相应的答案。

(5) 承认无知和错误。ChatGPT 能够主动承认自身的无知和错误，并在用户指出错误后进行优化或者给出解决途径。比如：

User:
ChatGPT4用到了多少参数?

Answer:
截止到我最后一次更新(2021年9月),OpenAI发布的最新版本是GPT-3,它使用了1750亿个参数。GPT-4尚未发布,因此我无法提供关于GPT-4的参数数量的信息。如果有新的版本发布,我建议您查阅OpenAI的官方文档或者相关的科学论文来获取最新和最准确的信息。

(6) 主动质疑不正确的问题。当遇到不正确的问题时,ChatGPT能够指出问题的错误并调整输出结果,比如:

Question:
汉高祖刘邦出生于公元1996年。

Answer:
对不起,这个信息不准确。汉高祖刘邦实际上出生于公元前256年,他是中国历史上的一位重要人物,是西汉的创始人。他在公元前202年建立了汉朝,这一朝代持续了四个世纪,对中国的历史产生了深远影响。公元1996年是近代,距离刘邦的生活时期已经过去了两千多年。

(7) 支持多模态。ChatGPT 4.0 可以支持“图像-语言”任务的完成,比如我们可以问ChatGPT“请用图表展示和简单说明 ChatGPT 的版本迭代”,于是 ChatGPT 可以输出如图 1.1 所示的示例。

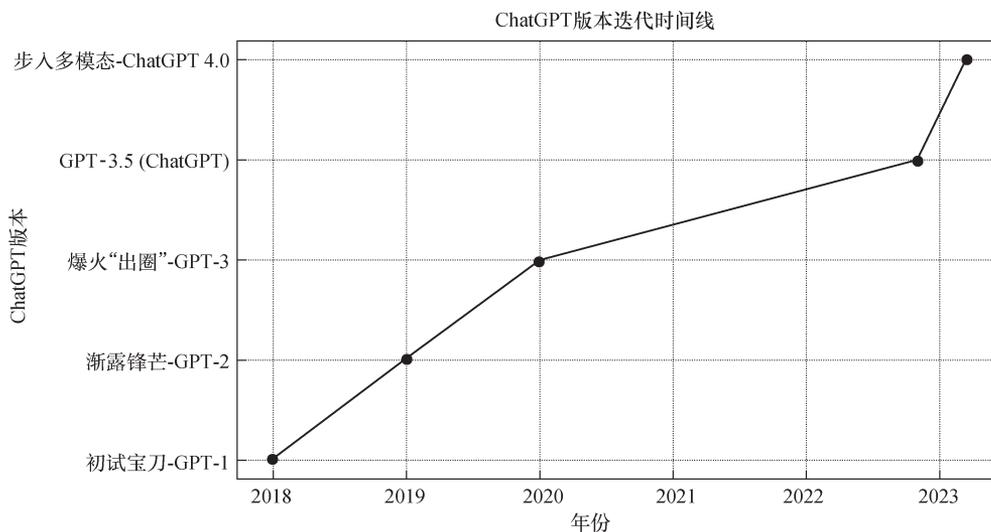


图1.1 ChatGPT “图像-语言”任务示例

当然,世界上不存在绝对完美的事情,ChatGPT也是如此,这里我们也需要点出一些它的缺陷,主要体现在如下几个方面。

(1) 数据泄露。ChatGPT是在大量的互联网文本上训练的,尽管它不会记住特定的数据源或文档,但它可能会生成在训练数据中看到过的信息,从而使它生成一些不准确或误导性的

信息。

(2) 生成能力的限制。虽然 ChatGPT 可以生成流畅且看似有意义的文本，但它并不真正理解它正在说什么，从而导致它可能会“一本正经地胡说八道”。

(3) 可能会生成不适当的内容。尽管 OpenAI 已经采取了一些措施来防止 ChatGPT 生成不适当或冒犯性的内容，但它仍然有概率生成一些不适当的回答。

1.2 ChatGPT和数据分析

在 1.1 节，我们深入探讨了 ChatGPT 的概念和原理，了解了它如何学习、理解和生成人类语言。但是，这些只是理论上的知识，我们可能会好奇，ChatGPT 在实际数据分析应用中能做什么？它的能力如何被转化为实际的价值？所以在接下来的一节中，我们将探讨 ChatGPT 在数据分析中的应用。

1.2.1 数据分析的基本概念

在讲解 ChatGPT 在数据分析的应用场景之前，我们有必要先梳理一遍数据分析的各种相关概念。

数据分析是用适当的统计分析方法对收集来的大量数据进行分析，将它们加以汇总和理解并消化，以求最大化地开发数据的功能和发挥数据的作用。数据分析的目的是把隐没在一大批看来杂乱无章的数据中的信息集中、萃取和提炼出来，以找出所研究对象的内在规律。

数据分析可以伴随企业经营或者产品运营的全过程，在企业或产品的“过去”“现在”和“将来”三个时间维度上，数据分析都能发挥极大的作用。

对于“过去”，数据分析可以帮助企业回顾和理解历史数据，进行原因分析。比如，企业可以通过分析过去的销售数据，了解哪些产品或服务最受欢迎，哪些销售策略最有效，从而为未来的决策提供依据。对于“现在”，数据分析可以帮助企业实时监控业务运行状态，洞察业务整体运作情况，及时发现和解决问题。比如，企业可以通过实时分析生产数据，发现生产过程中的瓶颈或故障，及时进行调整，确保生产的顺利进行。对于“将来”，数据分析可以帮助企业预测未来的趋势，为公司制订业务目标，并提供有效的战略参考和决策依据提前做好准备。比如，企业可以通过分析历史数据和市场趋势，预测未来的销售情况，从而提前调整生产计划，避免库存积压或缺货的情况。无论是回顾过去、把握现在，还是预测未来，数据分析都是企业不可或缺的工具。

那么，数据分析的基本流程是怎样的呢？我们可以把数据分析的流程总结为如下六个步骤，如图 1.2 所示。

(1) 定义问题和目标。在开始数据分析之前，我们需要问自己：为什么要进行数据分析？这次分析能解决什么问题？一旦目标明确，我们就可以开始规划分析策略、构建分析框架。我们需要将大目标分解为几个小目标，明确每个小目标需要从哪些角度进行分析，需要使用哪些指标。

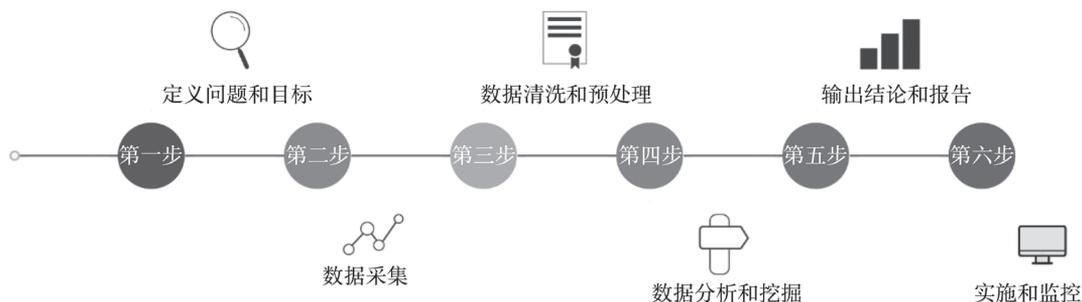


图1.2 数据分析全流程

(2) 数据采集。在明确了问题和目标之后，我们需要采集相关的数据。数据分为两种：第一手数据为可直接获取的数据，第二手数据为加工整理后得到的数据。常用的数据来源包括自有数据库、公开数据集、互联网爬虫、调查问卷等。

(3) 数据清洗和预处理。初步采集到的数据往往是“脏的”，也就是说，它们可能包含错误、缺失值、异常值等各种问题。在这个阶段，可能需要进行数据清洗、数据转换、数据提取和数据计算等一系列操作，从而使数据变得规整且可用。

(4) 数据分析和挖掘。在数据处理完成后，我们可以使用适当的方法和工具对数据进行分析，提取有价值的信息，得出有效的结论。在这一步，我们既可以对数据进行探索性分析，了解数据的基本特性，也可以选择合适的模型对数据进行建模。需要特别注意的是，如果数据本身存在错误，那么即使我们使用最先进的分析方法，得到的结果也可能是错误的。

(5) 输出结论和报告。数据分析报告是对整个数据分析过程的总结和呈现。通过报告，我们可以完整地展示数据分析的起因、过程、结果和建议，供相关人员参考。

(6) 实施和监控。在报告完成之后，我们可以根据报告的建议或决策实施，并对实施结果进行监控和评估。

在数字化时代，数据分析将不仅仅是专业的数据从业者才被要求掌握的能力，它将越来越多地融入各个岗位，成为不可或缺的技能之一。

1.2.2 ChatGPT在数据分析中的应用场景

前文我们在讲解 ChatGPT 特点时提到过它可以在多个场景应用，包括但不限于写邮件、写文章甚至编写代码等，那么 ChatGPT 会如何赋能数据分析呢？

实际上，当 ChatGPT 在搭配上各种实用插件后，其功能之强大让人赞叹，几乎可以在数据分析的各个流程、各个环节中都起到极大的辅助乃至主力作用。对照前文的数据分析流程，ChatGPT 数据分析全流程应用如图 1.3 所示。

在明确分析目的和确定思路阶段，ChatGPT 可以帮助数据分析师定义问题和设计分析方案。例如，你可以向 ChatGPT 描述你的业务问题，然后它可以帮助你将这个问题转化为一个或多个具体的数据分析任务，并给出你一些思路去尝试。

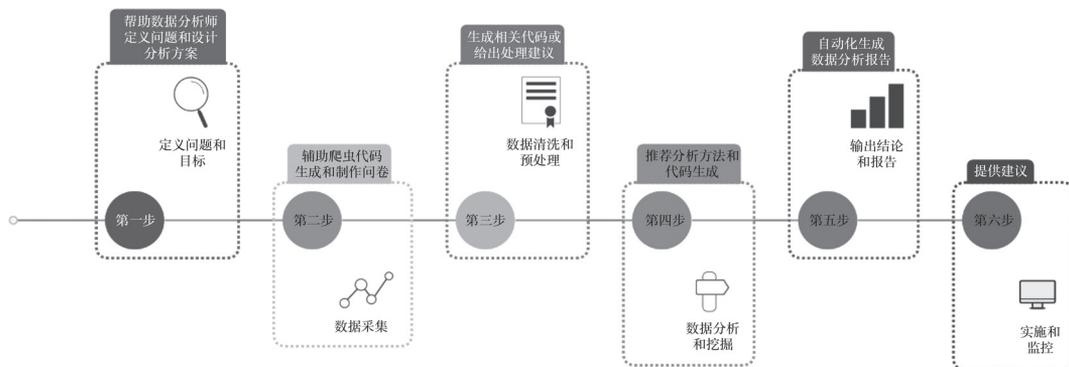


图1.3 ChatGPT数据分析全流程应用

在数据采集阶段，ChatGPT一方面可以辅助编写代码制作爬虫采集数据，另一方面可以通过一些插件去制作问卷，从而辅助数据的收集。此外，如果你向 ChatGPT 描述你的数据需求，那么它可以给你提供一些可能的数据源或者数据收集方法。

在数据清洗和预处理阶段，ChatGPT 可以生成数据清洗和预处理的代码，帮助你更高效地处理数据，当前一些智能化的插件甚至可以主动地帮助你指出需要处理的问题并生成相应代码。

在数据分析和挖掘阶段，ChatGPT 可以帮助你选择合适的数据分析方法和算法模型，并生成相应的代码。例如，你可以向 ChatGPT 描述你的分析目标，然后它可以推荐一些可能的分析方法，生成相应的代码，并且可以生成可视化图表。

在输出结论和报告阶段，ChatGPT 可以自动化生成数据分析报告。例如，你可以将你的数据分析结果或者数据集输入 ChatGPT 中，然后让它自动生成一份详细的数据分析报告。

总而言之，ChatGPT 可以作为数据从业者或者有数据需求的人员的强大助手，帮助他们更高效、更快捷地完成数据分析的各个阶段的工作。