# 自序

自 2022 年 11 月 30 日 OpenAI 发布 ChatGPT 以来，大模型技术掀起了新一轮人工智能浪潮。ChatGPT 在各个领域（如人机对话、文本摘要、内容生成、问题解答、识图、数学计算、代码编写等）取得了比之前算法好得多的成绩，在很多方面超越了人类专家的水平，特别是人机对话具备了一定的共情能力，这让 AI 领域的从业者和普通大众相信通用人工智能（Artificial General Intelligence，AGI）时代马上就要来临了。

大模型除了对话能力达到了真正可以互动的水准，更厉害的是当模型参数达到一定规模（100B[①]以上）时，会涌现新的能力，即大模型具备举一反三、任务分解、逻辑推理、解决未知任务的能力，这在之前的机器学习范式中是从未出现过的。之前的机器学习模型都是为某个具体任务构建的，只能完成特定的任务，对于新任务，必须训练新的机器学习模型。

最近 7~8 年，没有哪一项科技进步如 ChatGPT 这般吸引全球的目光[②]。除了媒体的大肆报道，国内外各类科技公司、科研机构、高等院校也在跟进大模型技术，基于大模型的创业公司如雨后春笋般层出不穷。

不到一年时间，国外就涌现了上百家大模型应用的创业公司，做得优秀的如 Midjourney、Runway、Inflection AI、Anthropic 等，都获得了上亿甚至几十亿美元的融资，估值达数十亿、上百亿美元。此外，一系列优秀的大模型相继被发布，如 Anthropic 的 Claude、Google 的 Bard、Meta 的 LLaMA 等。

国内也不甘落后，各个大厂、创业公司、科研院校相继发布了大模型产品，如智谱 AI 的 ChatGLM、百度的文心一言、华为云的盘古大模型、阿里云的通义千问、科大讯飞的星火大模型等，也有不少“大佬”亲自下场做大模型。

---

① 这里的 B 代表 Billion（十亿）。

② 上一次引发全球关注的 AI 大事件是 2016 年的 AlphaGo 战胜人类顶尖围棋手。

以 ChatGPT 为核心的大模型相关技术，可以应用于搜索、对话、内容创作等众多领域。在推荐系统领域的应用也不例外，在这方面已经有广泛的学术研究，学术界发表了大量的相关论文。我相信，在不久的将来（2024 年—2025 年），大模型相关技术会在工业界被大量用于推荐系统，并成为推荐系统的核心技术，就像 2016 年开始的深度学习技术对推荐系统的革新一样。

ChatGPT 和大模型相关技术不能被任何人、任何行业忽视，它在各行各业一定会呈现“井喷式”的发展。从 2023 年年初开始，我一直关注大模型相关技术的进展及其在行业上的应用，特别是在推荐系统领域的应用。在几乎每天都有大模型相关重磅突破发布的当下，我们必须跟上技术发展的步伐。为此，我花了近一年的时间写了这本《大模型推荐系统：算法原理、代码实战与案例分析》，希望能抛砖引玉，为国内普及、推广大模型在推荐系统上的应用发挥作用。

在开启我们的学习之旅前，我先用非常直白、浅显的方式来简单说明为什么大模型能应用于推荐系统，有了这个基础认识，我相信你可以更好地学习本书中的知识。

本书需要读者具备一定的背景知识。例如，熟悉推荐系统、有一定的编程能力，如果了解 NLP 就更好了。读者如果在学习过程中发现对某个知识点不熟悉，可以自行补充学习。

大模型在底层构建 Transformer 架构，通过海量的互联网文本信息预测下一个 token[①]出现的概率来预训练模型[②]。由于有海量的互联网文本数据，模型的预训练过程不需要人工标注（但需要对数据进行预处理），一旦模型完成预训练，就可以用于语言理解和语言生成任务。简单来说，大模型基于海量文本中的 token 序列中下一个 token 出现的概率进行统计建模，通过学习在给定语言片段后出现下一个 token 的概率来完成下游任务（例如文本摘要、翻译、生成文本等）。

对于推荐系统，用户过往的操作就是一个有序的序列，每个用户的操作序列就像一篇文本，所有用户的操作序列就像大模型的预训练语料库，预测用户下一个操作就像预测词序列的下一个 token[③]。通过这个简单的类比，我们知道推荐系统可以被嵌入大模型的理论框架中。因此，直观地看，大模型一定可以用于解决推荐系统问题。

上面的思路比较简单，只用到了用户与物品的交互信息。实际上，推荐系统的数据来源更复杂，除了用户历史交互序列，还有用户画像、物品画像信息等。部分用户画像、物品画像信息，例如用户的年龄、性别、偏好等，物品的标题、标签、描述文本等，可以利用自然语言来呈现，用户历史交互序列、用户画像、物品画像等信息都可以被输入大模型，为大模型提供更多的背景知识，最终的推荐结果会更加精准。

---

① token 既可能是一个单词，也可能是一个单词的一部分。

② BERT 等模型基于左右两边的 token 预测中间的 token，这属于模型架构上的不同。

③ 这里推荐系统的物品类似语言模型中的一个 token。

推荐系统涉及很多多模态数据，例如物品有描述文本、图片，甚至视频介绍等，这类异构的信息对于推荐系统的效果相当重要。多模态数据可以被转化为文本信息供大模型使用，目前的多模态大模型可以直接处理多模态数据，因此也可以直接用于推荐系统。

目前，即使不使用图片、视频等多模态数据，利用好文本数据也能让大模型很强大。大模型的强大之处是具备 zero-shot、few-shot 的能力[①]，很多推荐机制都利用了大模型的这个能力，只不过需要在使用大模型的过程中设计一些提示词（prompt）和模板（template）来激活大模型的推荐能力。

说一下我个人对激活的理解。大模型有上百亿、上千亿，甚至上万亿个参数，是一个非常庞大的神经网络。当用一些提示词告诉大模型作为推荐系统进行个性化推荐时，就激活了深度神经网络中的某些连接，这些连接是神经网络的某个子网络，而这个子网络具备进行个性化推荐的能力，这个过程非常类似人类大脑神经元的工作机制。例如当你看到美食时，就会激活大脑中负责进食的区域，这个区域是大脑整个复杂神经元网络的子网络，导致可能产生流口水、吞咽等行为，这里“看到美食”就类似大模型的提示词。另外，我们在进行头脑风暴时，突然被别人启发想到某个绝妙的创意也是一种激活过程。few-shot 更复杂一些，需要在提示词中告诉大模型一些过往的推荐案例，例如用户看了 A、B、C 三个视频后，会看另一个视频 D，让它临时学习如何推荐。

提示词学习没有改变大模型的参数，即没有进行梯度下降的反向传播训练，但为什么具备 zero-shot、few-shot 的能力呢？提示词作为一个整体，激活了大模型神经网络的某个功能区域。大模型具备多轮对话能力的道理也是类似的，我们可以将多轮对话作为一个整体，这个整体激活了大模型在某个对话主题下的功能区域，导致大模型能“记住”多轮对话之前的信息[②]。由于目前的大模型不具备增量学习[③]的能力，对话完成后，对话中的新信息并没有被大模型学习到。

除了直接利用大模型的 zero-shot、few-shot 能力进行推荐，我们还可以按照大模型的输入、输出范式准备推荐系统的相关数据，然后通过监督学习微调大模型，让大模型更好地适配具体的推荐场景，这也是将大模型应用于推荐系统的一个非常有价值的方向。

另外，大模型压缩的世界知识、大模型的涌现能力、大模型的自然对话能力都可以很好地

---

① zero-shot 指预训练后可以直接完成未知的下游任务。few-shot 指给出几个示例，大模型可以解决类似的问题，即所谓的上下文学习能力，也就是举一反三的能力。

② 这个对话是作为整体输入大模型的，或者可以理解为整个对话过程就是一次连贯的语言生成过程，只不过部分话语是人类给出的，模型接着人类的话语继续生成。

③ 增量学习指遇到一个新信息马上学习到模型的参数中，人是具备增量学习能力的，增量学习肯定是大模型未来最重要的一个研究方向。

被用于推荐系统，解决深度学习推荐系统很难解决的问题，下面举两个例子说明。

首先，大模型有助于缓解数据稀疏问题，特别是冷启动问题[①]，这是当前深度学习推荐系统的主要瓶颈。通过从在不同模型架构中学习的预训练模型中提取和迁移知识，可以提高推荐系统的通用性、稀疏性、效率和有效性等性能。

其次，大模型一个很大的优势是可以利用对话的方式跟用户互动，就像 ChatGPT 所呈现的那样。如果能将推荐系统设计成一个跟用户互动的对话式推荐引擎，那么大模型可以利用自然语言响应用户的个性化需求，从而提升用户的整体体验和参与度。

通过前面的介绍，相信你已经大致知道为什么大模型可以被应用于推荐系统了，也知道了将大模型应用于推荐系统的独特优势，那么如何将大模型应用于推荐系统呢？这就是本书的核心内容，你将从这本书中找到答案。

① 大模型学习的是海量的互联网知识，对于新物品、新用户都可以很好地进行知识迁移。

# 前言

## 为什么写作本书

我从 2010 年开始研究、实践推荐系统，属于国内最早从事推荐系统工作的一批人。过去 15 年，我有过至少 4 次从零开始成功构建推荐系统的经历，曾经负责构建的推荐系统最高有超过千万日活跃用户数量（Daily Active User，DAU）。我出版过两本推荐系统相关图书，分别是《构建企业级推荐系统：算法、工程实现与案例分析》（2021.09）和《推荐系统：算法、案例与大模型》（2024.04）。过去十几年的经历让我见证了推荐系统技术的完整发展过程，我也一直跟随推荐系统的发展大势进行学习、实践。

2022 年 11 月底 ChatGPT 发布后，全世界掀起了新一轮的人工智能浪潮，以 ChatGPT 为代表的大模型技术逐步渗透到各个应用场景和领域，当然也包括推荐系统。将大模型应用在推荐系统上的相关学术研究非常多，截至 2024 年 5 月，已有上百篇相关的研究论文发表。大模型在产业上的应用也逐步开始：阿里巴巴在淘宝上内测了淘宝问问—— 一个对话式推荐引擎；Meta 在尝试利用大模型技术实现万亿级参数的新一代推荐系统；百度正在利用大模型重构底层的核心搜索、推荐模块……

ChatGPT 和大模型带来的影响是空前的，过去没有哪一项 AI 技术对全球的冲击像 ChatGPT 和大模型这么大。我预见到 ChatGPT 和大模型会革新推荐行业，因此在过去的一年多时间里，我阅读了上百篇大模型、大模型推荐系统相关的论文，并且跟行业专家进行了密切的交流。

另外，我从 2023 年 4 月开始创业，公司业务为 B 端的数智化转型，主要方向是精细化运营、大模型搜索推荐、大模型智能知识顾问等。过去一年，我针对将大模型应用于推荐、搜索等话题与相关的企业进行交流并付诸实践。

结合过往的学习、交流、实践，我准备将自己掌握的大模型推荐系统的知识框架进行系统梳理，整理成一本系统介绍大模型推荐系统的图书，希望为推荐行业提供一套完整的、基于大模型的方法论和实践指南。

希望我的经验和经历可以帮助想学习大模型在推荐系统上应用的读者体系化了解并快速掌握大模型推荐系统。

## 读者对象

本书主要讲解大模型在推荐系统中的应用，既有算法原理，又有代码实现，聚焦于如何利用最新的大模型技术赋能、革新、重构现有的推荐系统。本书需要读者有一定的推荐系统基础知识，了解大模型的一些基本原理，本书适合以下读者。

1. 推荐系统开发人员及推荐算法研究人员。
2. 期望从事推荐系统相关工作的学生。
3. 在高校从事推荐算法研究，希望对大模型在推荐系统中的应用有全面了解的科研人员。
4. 对大模型在推荐、搜索中的应用感兴趣的产品和运营人员。
5. 期望将大模型引入推荐、搜索产品的公司管理层人员。

## 如何阅读本书

本书分为 9 章，包含大模型基本原理介绍、将大模型应用于推荐系统的思路和方法、在电商推荐场景中使用大模型等。下面分别对各个章节进行简单介绍。

第 1~3 章是准备部分。第 1 章介绍大模型的基础知识，包括大模型的发展历史、数据资源、数据预处理、大模型预训练、大模型微调、大模型推理、大模型部署、相关软件和框架等，这一章是后续章节的理论基础，是为没有大模型基础的读者准备的。如果你已经非常熟悉大模型，那么可以略过这一章。第 2 章对后续章节用到的数据和开发环境进行介绍，本书的代码实现基于微软的 MIND 数据集和 Amazon 电商数据集，开发环境包括 Python 沙盒、CUDA 和 MacBook。第 3 章将大模型在推荐系统中的应用抽象为 4 种范式——生成范式、预训练范式、微调范式、直接推荐范式。

第 4~7 章展开说明第 3 章介绍的 4 种范式，每章既包含算法原理，又包含相关的案例说明，同时会基于 MIND 数据集给出对应的代码实现。

第 8 章是一个完整的实战案例，基于 Amazon 商品评论数据，利用大模型解决电商推荐问题。本章讲解如何利用大模型来解决生成用户兴趣画像、生成个性化商品描述信息、猜你喜欢推荐、关联推荐、冷启动、推荐解释、对话式推荐 7 类问题，针对每类问题都提供完整的步骤及对应的代码实现。其中，前 6 类问题是大模型对传统推荐算法解决方案的有效补充和突破，有了大模型的支持，传统推荐算法有更新颖、更好的解决方案；而对话式推荐是基于传统推荐

系统的一种新的推荐形态，借助大模型的自然语言对话能力，我们可以采用对话的方式为用户进行个性化推荐，在提升用户体验的同时，带来更高的商业价值。

第 9 章为将大模型推荐系统更好地应用于工业界提供了一些思路和方法，包含大模型的高效预训练、高效推理，以及真实业务场景中的一些问题和建议。

本书是专门为方便读者快速上手大模型推荐系统而准备的。本书主要有三个特点：一是将大模型应用于推荐系统抽象为 4 种范式，在这个统一的框架下，你可以体系化地学习大模型推荐系统；二是包含电商场景的最佳实践，针对每种问题，提供完整的利用大模型解决推荐问题的思路、步骤和方法；三是包含代码实战内容，第 2~8 章都有完整的代码实现。**本书所有的代码都可以从 GitHub 的代码仓库 liuq4360/llm4rec_abc 中获取，读者也可以根据封底处提示，加入本书读者群，获取本书代码、链接及参考文献。**

希望本书能够成为有志于利用大模型技术更好地赋能传统推荐、搜索系统的爱好者和从业者的方法论和落地实战指南！

## 勘误和支持

由于作者水平和写作时间有限，书中难免有所纰漏，恳请读者批评指正。你可以将书中描述不准确的地方或者错误告诉我，以便本书重印或者再版时更正。你可以通过微信 liuq4360 与我取得联系，或者发送邮件至邮箱 891391257@qq.com，我很期待看到你真挚的反馈。

## 致谢

首先，感谢 AI 行业的技术信仰者，正是他们的不懈努力为大模型技术带来快速的发展和广阔的前景。同时，感谢过去一年来给予我信任的个人和企业，通过与他们合作，我更好地理解和实践了大模型相关技术在产业上的应用。

其次，感谢电子工业出版社张爽老师，在她的耐心指导和建议下，我一步步优化了本书的结构和内容，让这本书的质量得到了保证。

最后，感谢我的父母和家人，是他们的无私付出让我有足够的时间学习、实践、创业！

谨以此书，献给所有懂我、关心我、支持我的家人和朋友！

刘强

2024 年 5 月 9 日于上海图书馆东馆

# 目录