新一代计算机与人工智能系列新形态教材

人工智能
技术与应用

主 编　黄风华　黄家乾

南京大学出版社

图书在版编目(CIP)数据

人工智能技术与应用 / 黄风华，黄家乾主编.
南京 ：南京大学出版社，2026. 8. -- ISBN 978-7-305
-30281-7

Ⅰ. TP18

中国国家版本馆 CIP 数据核字第 2026ZX6494 号

出版发行　南京大学出版社
社　　址　南京市汉口路 22 号　　邮　　编　210093

书　　名　**人工智能技术与应用**
RENGONG ZHINENG JISHU YU YINGYONG
主　　编　黄风华　黄家乾
责任编辑　刘　飞

照　　排　南京布克文化发展有限公司
印　　刷　南京凯德印刷有限公司
开　　本　787 毫米×1092 毫米　1/16 开　印张　22.75　字数　553 千
版　　次　2026 年 8 月第 1 版　2026 年 8 月第 1 次印刷
I S B N　978-7-305-30281-7
定　　价　59.00 元

网　　址　http://www.njupco.com
官方微博　http://weibo.com/njupco
官方微信　NJUYUNSHU
销售咨询热线　025－83594756

目录

CONTENTS

模块一　人工智能概论

模块二　机器学习

模块三 神经网络

模块四 专家系统

模块五 智能机器人

模块六 数据挖掘

模块七　群体智能

模块八　计算机视觉

模块九　大数据与人工智能

模块十　大模型

模块十一　智能体

模块十二 AI 的视听应用

模块十三 AI 的挑战与未来

◆ 模块导读

在模块一的任务三中，已了解到“数据是 AI 学习的原料”，并初步认识了大数据对 AI 发展的驱动作用。在模块四的任务五中，接触过“模糊专家系统”的概念——用模糊逻辑来处理“非黑即白”之间的“灰色地带”。本模块将深入展开这两个主题：模糊逻辑和大数据，探讨二者与人工智能之间的深层关联。

现实世界充满了“不精确”的信息：“今天有点热”“这道菜偏咸”“他大概 30 岁”——人类天然能理解和处理这种“模糊”的描述，但传统的计算机只能处理精确的数值（“气温 32.5 摄氏度”“盐分含量 3.2 克”“年龄 29 岁”）。模糊逻辑如何让计算机也能处理这种“模糊性”？另一方面，当数据的规模从“千”跃升到“亿”甚至“万亿”时，会带来哪些根本性的思维变革？大数据时代的到来如何重塑了 AI 的研究范式和应用格局？

本模块通过四个任务展开：任务一和任务二介绍模糊逻辑的基本概念和系统构成；任务三和任务四讨论大数据思维的变革以及大数据与 AI 的协同关系。通过本模块的学习，读者将理解 AI 处理“不确定性”的两种重要方式（模糊逻辑处理“语义上的模糊”，概率方法处理“统计上的不确定”），并从数据的角度形成对 AI 发展全局的更深认识。

◆ 学习目标

【了解】

—模糊逻辑的基本概念，包括模糊集合、隶属度和模糊规则；

—模糊逻辑系统的基本结构：模糊化、推理、去模糊化；

—大数据的定义和“4V”特征（大量、高速、多样、价值），以及大数据与传统数据分析的区别。

【理解】

—模糊逻辑如何将“模糊的自然语言描述”转化为“计算机可以处理的数学运算”；

—大数据时代带来的思维变革：从“因果”到“相关”、从“精确”到“趋势”、从“抽样”到“全量”；

—大数据与人工智能之间的相互驱动关系：大数据为 AI 提供“燃料”，AI 为大数据提供“分析引擎”。

【掌握】

—根据问题的特点，判断是否需要引入模糊逻辑来处理“不精确性”；

—运用大数据的“4V”框架，分析一个给定场景中的数据特征和挑战。

【能够】

—结合模块一(数据是 AI 三大引擎之一)、模块二(机器学习)和模块六(数据挖掘)的知识,形成对“数据在 AI 生态中的核心地位”的完整认知;

—对日常生活中的“模糊控制”应用(如空调自动温控、洗衣机智能洗涤)和“大数据应用”(如搜索推荐、城市交通优化)形成技术层面的理解。

◆ 思维导图

下图展示了本模块的知识结构与内容框架,帮助读者从整体上把握各任务之间的逻辑关系:

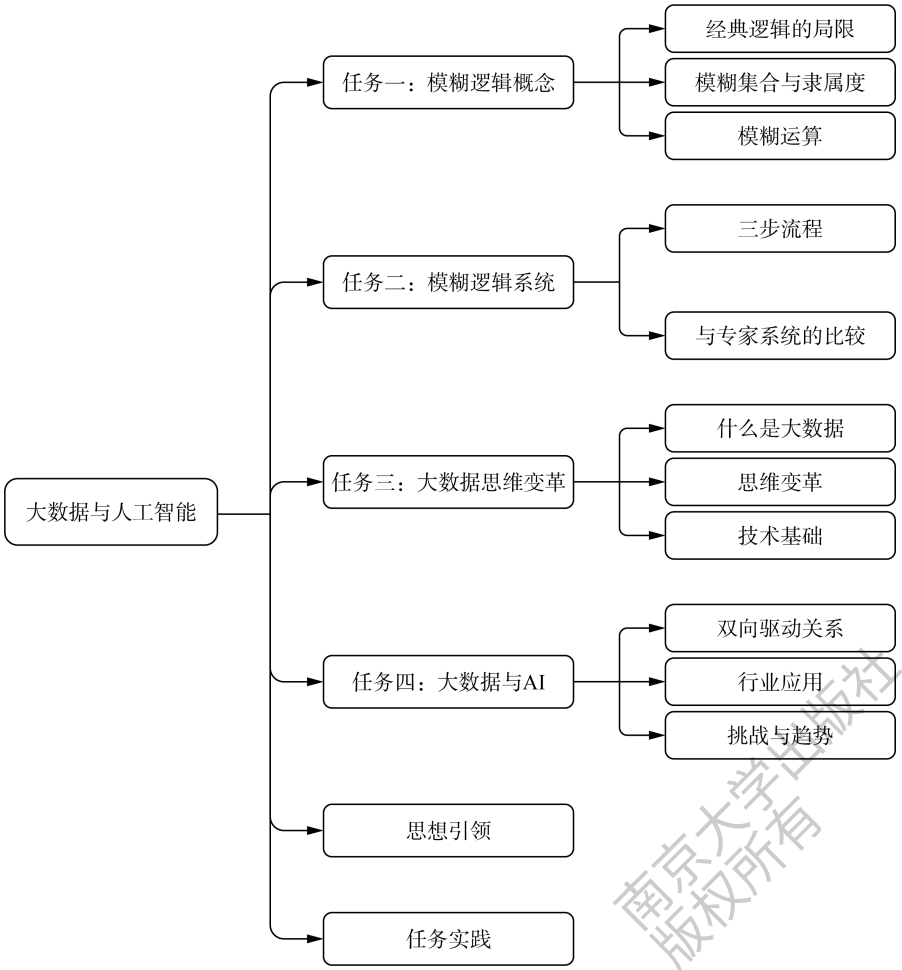

图 9-1 模块九知识结构思维导图

任务一　模糊逻辑的概念

◆ 引导思考

“今天天气很热”——这句话在日常交流中完全没有歧义，但如果让计算机来处理，它会“困惑”：“热”到底是多少℃？32℃算热吗？28℃呢？35℃呢？在“热”和“不热”之间，有一条精确的分界线吗？

在模块四中了解到，传统专家系统的规则是“非黑即白”的（“如果温度＞38℃，则判定为发烧”）。但现实世界中的很多概念（如“年轻”“远”“贵”“好”）都是渐变的、没有明确边界的。有没有一种数学工具能让计算机也能处理这种“模糊性”？

模糊逻辑与模块四中学过的“模糊专家系统”是什么关系？它的核心思想和基本概念是什么？

一、经典逻辑的局限

在模块四（专家系统）中学过产生式规则的逻辑：“如果条件 A 成立，那么结论 B 成立。”这种逻辑建立在经典逻辑（也叫“二值逻辑”）的基础上：任何命题只有两种可能——“真”（1）或“假”（0），没有中间状态。“这个人发烧了”要么是真的（体温＞38℃），要么是假的（体温≤38℃），不存在“有点发烧”或“差不多发烧”的概念。

但现实世界远不是“非黑即白”的。考虑以下几个日常判断：“这个人年轻吗？”——25 岁的人显然年轻，65 岁的人显然不年轻，但 40 岁的人呢？“半年轻半不年轻”？“这杯咖啡烫吗？”——90℃的咖啡肯定烫，30℃的肯定不烫，但 55℃呢？“有点烫”？

这些概念均具备一项共同特征：在“绝对属于”和“绝对不属于”之间，存在一个渐变过渡的“灰色地带”。经典逻辑的“非 0 即 1”无法表达这种渐变性，这就是它的根本局限。

1965 年，美国加州大学伯克利分校的拉特飞·扎德教授提出了模糊集合理论和模糊逻辑，正是为了解决这个问题。他的核心思想非常优雅：既然“属于”和“不属于”之间有灰色地带，那就不要强迫一个元素“二选一”，而是允许它以一定程度属于某个集合。这个“程度”就是所谓的隶属度。

二、模糊集合与隶属度

在经典的集合论中，一个元素对一个集合的关系只有两种：“属于”（隶属度＝1）或“不属于”（隶属度＝0）。例如，“30 岁”对于“年轻人”这个集合，要么属于（1），要么不属于（0）——没有中间地带。

模糊集合打破了这种二元限制：一个元素可以以 0 到 1 之间的任意值属于一个集合。

这个值就叫作隶属度，表示“属于的程度”。例如：

对于“年轻人”这个模糊集合：20 岁的隶属度＝1.0（完全属于“年轻人”），35 岁的隶属度＝0.7（在较大程度上属于“年轻人”），50 岁的隶属度＝0.3（在较小程度上属于“年轻人”），70 岁的隶属度＝0.0（完全不属于“年轻人”）。

隶属度的值不是随意确定的，而是通过一个隶属函数来定义的。隶属函数是一条从 0 到 1 的曲线，它描述了某个变量的每一个可能取值“属于”该模糊集合的“程度”。不同的模糊概念对应不同形状的隶属函数：“年轻”可能是一条从左到右递减的曲线（年龄越大，“年轻”的程度越低）；“中等温度”可能是一个钟形曲线（在某个温度附近隶属度最高，偏离越远越低）。

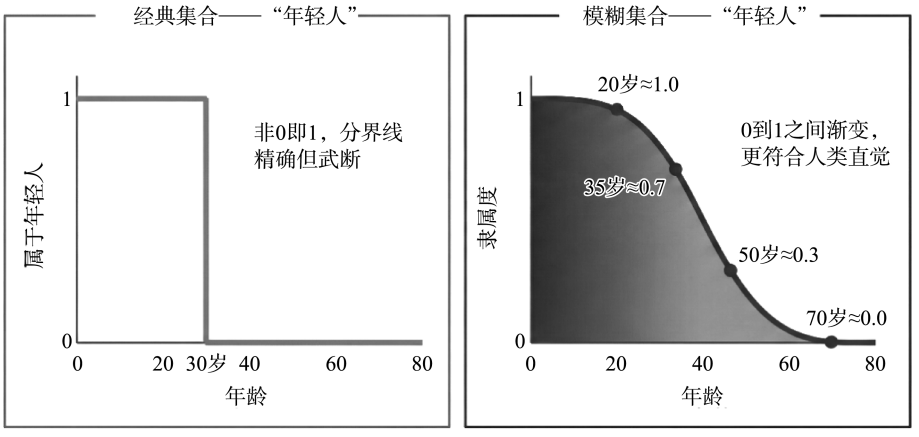

图 9-2　经典集合与模糊集合的对比

用一个比喻来总结：经典逻辑就像一盏开关灯——只有“开”（1）和“关”（0）两种状态；模糊逻辑则像一盏调光灯——亮度可以从 0%连续调节到 100%。对于描述“热不热”“年不年轻”“远不远”这样的渐变概念，“调光灯”显然比“开关灯”更贴近现实。

三、模糊运算

有了模糊集合和隶属度的概念，就可以在“模糊”的世界中进行逻辑运算了。经典逻辑中的“与”“或”“非”操作在模糊逻辑中都有对应的版本：

模糊“与”（同时满足）：取两个隶属度中的较小值。例如，“这道菜既辣（隶属度 0.8）又咸（隶属度 0.6）”的程度＝min(0.8,0.6)＝0.6。直觉理解：“同时满足两个条件”的程度受限于更弱的那个条件（“木桶效应”）。

模糊“或”（至少满足一个）：取两个隶属度中的较大值。例如，“这个人年轻（隶属度 0.3）或有经验（隶属度 0.9）”的程度＝max(0.3,0.9)＝0.9。直觉理解：“至少一个条件满足”的程度取决于更强的那个。

模糊“非”（否定）：用 1 减去隶属度。例如，“不年轻”的隶属度＝1－0.7＝0.3。直觉理解：“年轻到 0.7 的程度”意味着“不年轻到 0.3 的程度”。

这三种模糊运算可以构建模糊规则：“如果温度是‘高’（隶属度）且湿度是‘大’（隶属

度)，那么空调输出‘强冷风’(隶属度)”。多条这样的模糊规则组合起来，就构成了一个模糊推理系统——这正是本模块任务二将详细讨论的内容。

案例：空调温控——你身边最常见的“模糊逻辑”应用

人们家中的空调很可能在使用模糊逻辑来控制温度。传统的空调温控是“开关式”的：设定温度25℃，当室温升到26℃时压缩机全速运转制冷，降到24℃时压缩机完全停止——温度在25℃上下反复波动，体感忽冷忽热。

采用模糊控制的空调则“聪明”得多：它将温度差和温度变化速度都用模糊集合表示(如“温差：有点高/很高/接近目标”)，然后通过模糊规则(如“如果温差‘有点高’且温度‘下降慢’，那么压缩机功率‘中等偏高’”)来连续调节压缩机的输出功率——不是“全开或全关”，而是根据当前状况“恰到好处”地调节。结果是温度更加稳定、体感更加舒适、能耗也更低。

类似的模糊控制还广泛应用于：洗衣机(根据衣物质量和脏污程度模糊判断水量和洗涤时间)、电饭煲(根据米的品种和量模糊调节火力)、地铁列车的自动驾驶系统(根据站间距离和当前速度模糊调节加减速)等。

在这些应用中，模糊逻辑的价值在于：用简单直观的“如果……那么……”规则(而不是复杂的数学模型)就能实现精细的连续控制——这正是模块四中专家系统的“规则思想”在控制领域的成功延伸。

表9-1　经典逻辑与模糊逻辑的对比

维度	经典逻辑(二值逻辑)	模糊逻辑
真值范围	只有0(假)和1(真)	0到1之间的任意值
集合边界	精确(属于或不属于)	模糊(属于的程度渐变)
处理“灰色地带”	不能(强制二选一)	自然支持(隶属度表示程度)
“与”运算	两者都为真→真	取两个隶属度的较小值
典型应用	数学证明、程序逻辑	温控、洗衣机、模糊专家系统
通俗比喻	“开关灯”(只有开和关)	“调光灯”(亮度可连续调节)

知识拓展：模糊逻辑与概率——两种“不确定性”的区别

初学者常常混淆“模糊”和“概率”，认为它们是同一回事。实际上，它们处理的是两种截然不同的“不确定性”。

概率处理的是随机性：“明天下雨的概率是60%”意味着“明天要么下雨要么不下雨(事件本身是确定的，非此即彼)，但不确定会是哪种”。

模糊逻辑处理的是模糊性(概念的不精确性)。例如，“今天有点热”这句话并非表示“有60%的概率热”，而是在说明“热”这个概念本身就没有精确的边界——当前的温度在一定程度上属于“热”，也在一定程度上属于“不热”。

用一个例子来说明："这个人有70%的概率是张三"——这是概率问题(这个人要么是张三要么不是,只是不确定)。"这个人属于'年轻人'的程度是0.7"——这是模糊性问题("年轻人"这个概念本身就没有精确的边界)。

在AI领域,概率方法(如模块二中的贝叶斯分类器)和模糊逻辑各有所长,分别适用于处理不同类型的不确定性。模块四的"模糊专家系统"正是将模糊逻辑引入了专家系统的规则推理过程。

本节小结

1. 经典逻辑只有"真"(1)和"假"(0)两种状态,无法自然地表达"有点热""比较年轻"等渐变概念。

2. 模糊逻辑引入了模糊集合和隶属度(0到1之间的连续值),使计算机能够处理"灰色地带"的概念。隶属函数定义了每个值"属于"某个模糊集合的"程度"。

3. 模糊运算(模糊与=取小值,模糊或=取大值,模糊非=1−隶属度)使得基于模糊集合的逻辑推理成为可能。

4. 模糊逻辑最广泛的应用是模糊控制:空调温控、洗衣机智能洗涤、电饭煲调节火力等家电产品中大量使用。

5. 模糊逻辑处理的是"概念的模糊性"(边界不清晰),与概率处理的"事件的随机性"(结果不确定)是两种不同类型的不确定性。

思 考 题

1. [识记] 请用"开关灯"和"调光灯"的比喻,向没有技术背景的朋友解释经典逻辑和模糊逻辑的核心区别。

2. [理解] "这个人有60%的概率是学生"和"这个人属于'年轻人'的隶属度是0.6"——这两句话描述的"不确定性"有什么本质区别?请用知识拓展框中的概念来解释。

3. [应用] 假设你要为一个智能灌溉系统设计模糊逻辑规则。系统需要根据"土壤湿度"(干/适中/湿)和"天气预报"(晴/多云/雨)来决定"灌溉量"(多/中/少/不灌)。请尝试写出3—4条模糊规则(如"如果土壤湿度是'干'且天气预报是'晴',则灌溉量为'多'"),并说明这些规则中的"干""晴"等概念为什么适合用模糊集合而非精确阈值来定义。

任务二　模糊逻辑系统

◆ 引导思考

在任务一中学习了模糊集合、隶属度和模糊运算这些“零件”。现在的问题是：如何将这些零件“组装”成一个能够实际工作的系统？一个完整的模糊逻辑系统内部由哪些环节构成？

一条模糊规则说“如果温度是‘高’，那么风扇转速为‘快’”。但计算机最终需要输出一个精确的数字（比如“风扇转速 1 500 转/分钟”）。从模糊的“高”到精确的“1 500 转”，中间经历了什么过程？

模糊逻辑系统与模块四中的专家系统有什么结构上的相似之处？

一、模糊逻辑系统的三步流程

一个完整的模糊逻辑系统（也叫模糊推理系统或模糊控制器）的工作流程可以概括为三个核心步骤：模糊化、模糊推理和去模糊化。这三步完成了“从精确输入到模糊世界，在模糊世界中推理，再从模糊世界回到精确输出”的完整闭环。

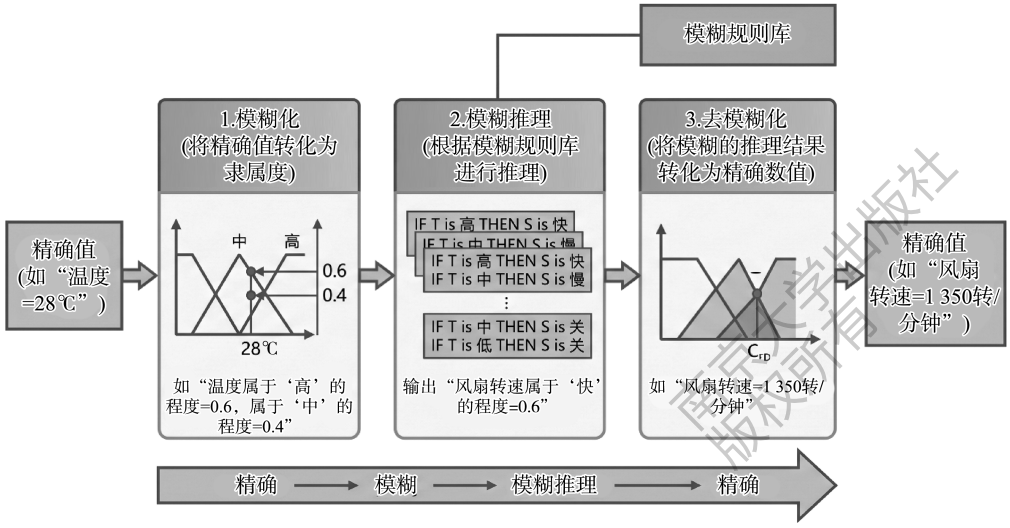

图 9-3　模糊逻辑系统的三步工作流程

1. 模糊化：从“精确”到“模糊”

模糊化是将精确的输入值转化为模糊集合中的隶属度。例如，温度传感器测量到的精确值是28℃。模糊化步骤将这个精确值“映射”到多个模糊集合上：28℃属于“低温”的程度＝0.0，属于“中温”的程度＝0.4，属于“高温”的程度＝0.6。这就像把一个精确的“数字”翻译成了一段模糊的“自然语言描述”：“温度偏高，但也有点接近中等。”

2. 模糊推理：在“模糊世界”中做判断

模糊推理根据一组预设的模糊规则（由人类专家制定，类似模块四中专家系统的产生式规则）进行推理。一个典型的模糊规则库可能包含这样的规则：“如果温度是‘高’且湿度是‘大’，那么风扇转速为‘快’”；“如果温度是‘中’且湿度是‘小’，那么风扇转速为‘慢’”。模糊推理引擎逐条匹配这些规则，根据输入的隶属度（温度属于“高”的程度＝0.6）计算出输出的隶属度（风扇转速属于“快”的程度＝0.6）。

当多条规则同时被激活时（这在实际中很常见），推理引擎需要将多条规则的结论“合并”起来。这与模块四中专家系统处理多条规则冲突的逻辑类似，但模糊推理的“合并”是基于隶属度的数学运算（通常取最大值），而不是简单的优先级排序。

3. 去模糊化：从“模糊”回到“精确”

模糊推理的输出是一个模糊的结论（如“风扇转速属于‘快’的程度＝0.6，属于‘中’的程度＝0.3”）。但实际的执行器（如风扇电机）需要一个精确的数值（如“1 350 转/分钟”）。去模糊化就是将模糊的推理结果“翻译”回一个精确的数字。

最常用的去模糊化方法是重心法（也叫质心法）：将模糊输出的隶属度曲线看作一个“面积”，计算这个面积的“重心”位置，重心对应的值就是最终的精确输出。直觉上，这相当于“综合考虑所有规则的建议，找到一个‘平衡点’作为最终决策”。

二、模糊逻辑系统与专家系统的比较

如果仔细观察，会发现模糊逻辑系统的结构与模块四中的专家系统有着惊人的相似性：都有一个“规则库”（存储 IF-THEN 规则），都有一个“推理引擎”（根据规则进行推理），都需要“输入接口”和“输出接口”。区别在于：专家系统的规则是“非黑即白”的精确逻辑，模糊逻辑系统的规则允许“灰色地带”的渐变过渡。

表 9-2　模糊逻辑系统与专家系统的对比

维度	专家系统（精确规则）	模糊逻辑系统（模糊规则）
规则形式	IF 温度＞38℃ THEN 发烧	IF 温度是“高” THEN 药量是“大”
输入处理	直接使用精确值	先将精确值“模糊化”为隶属度
推理方式	精确的逻辑匹配	基于隶属度的模糊运算
输出形式	精确的类别判断	先得到模糊结论，再“去模糊化”

续表

维度	专家系统(精确规则)	模糊逻辑系统(模糊规则)
适合的问题	边界清晰的判断(是/否)	边界模糊的连续控制(多/少)
典型应用	医疗诊断、故障排查	温控、洗衣机、工业过程控制

案例：模糊逻辑洗衣机——“一键智洗”的背后

一台“智能”洗衣机可能这样工作：

传感器输入：光电传感器测量水的浑浊度(精确值：浑浊度＝65%)；重力传感器测量衣物质量(精确值：3.2 公斤)。

模糊化：浑浊度 65%→属于“脏”的程度＝0.7，属于“中等”的程度＝0.3；质量 3.2 公斤→属于“中量”的程度＝0.8，属于“多量”的程度＝0.2。

模糊推理：规则 1“如果衣物‘脏’且‘中量’，则洗涤时间‘较长’(强度 0.7)”；规则 2“如果衣物‘中等脏’且‘中量’，则洗涤时间‘中等’(强度 0.3)”——两条规则同时激活，综合得到洗涤时间的模糊输出。

去模糊化：将模糊输出转化为精确数值：洗涤时间＝38 分钟，水量＝45 升。

这个过程的优雅之处在于：工程师不需要为每一种“浑浊度＋质量”的组合编写精确的控制程序(组合数量太多)，只需要制定十几条直觉性的模糊规则(“脏＋多＝洗久一点”)，系统就能自动处理所有可能的输入组合，输出合理的控制方案。

知识拓展：模糊控制的“日本往事”——一场技术革命

模糊逻辑虽然是美国学者扎德在 1965 年提出的，但它最辉煌的商业成功却发生在日本。20 世纪 80 年代末到 90 年代初，日本企业率先将模糊控制技术大规模应用于消费电子产品：松下的模糊逻辑洗衣机、日立的模糊控制空调、三菱的模糊控制电梯(根据乘客流量模式自动调度)、仙台地铁的模糊控制自动驾驶系统(1987 年投入运营，被认为是世界上最平稳的地铁之一)。

“模糊”(Fuzzy)一度成为日本市场上的流行标签，出现在各种家电产品的广告中。这段“模糊热潮”是 AI 技术(广义上)在消费市场上最早的大规模商业成功案例之一。

模糊控制在日本成功的一个重要原因是：日本的制造业有大量需要精细控制的工业场景，而模糊控制恰好提供了一种“不需要精确数学模型就能实现精细控制”的方法——这与模块四中专家系统“用规则代替数学模型”的思路异曲同工。

本节小结

1. 模糊逻辑系统的三步工作流程：模糊化(精确值→隶属度)→模糊推理(根据模糊规则推导结论)→去模糊化(模糊结论→精确输出值)。

2. 模糊逻辑系统在结构上与专家系统高度相似(规则库＋推理引擎)，核心区别是规

则允许“模糊”的渐变描述，适合处理连续控制问题。

3. 模糊控制在家电（洗衣机、空调）、交通（地铁）和工业控制中有广泛应用。日本是模糊控制技术最早大规模商业化的国家。

思 考 题

1.［识记］请按顺序描述模糊逻辑系统的三步工作流程，并说明每一步的核心任务。

2.［理解］模糊逻辑系统和模块四中的专家系统在结构上有什么相似之处？在处理“不确定性”的方式上有什么核心区别？

3.［应用］假设你要设计一个“智能电饭煲”的模糊控制系统。输入变量是“米的量”（少/中/多）和“米的品种”（糯米/普通米/糙米），输出变量是“加水量”和“烹饪时间”。请尝试：a. 为“米的量”设计一个简单的模糊集合（用文字描述隶属度的大致范围）；b. 写出3条模糊规则；c. 说明为什么用模糊控制比用精确的“查表法”（每种米的量和品种对应一个固定的水量和时间）更好。

任务三　大数据思维与变革

◆ 引导思考

在模块一的任务三中，了解到“大数据是AI的三大引擎之一”。但大数据对AI（乃至整个社会）的影响远不只是“数据变多了”这么简单。它带来了一种全新的思维方式——用数据来认识世界、发现规律、作出决策。这种“大数据思维”到底“新”在哪里？

“更多的数据”为什么比“更好的算法”更重要？这个说法有道理吗？

从Google搜索到精准医疗，大数据正在如何改变各行各业？它的技术基础是什么？

一、什么是大数据

在模块一的任务三中，已经从“AI三大引擎”的角度讨论过数据对AI发展的驱动作用。那里侧重的是“数据量的爆发为AI模型提供了训练素材”；本任务将从更广的视角——大数据作为一种新的资源形态和思维方式——来理解它的深层影响。

大数据通常用“4V”来定义其特征：

Volume（大量）：数据的规模从GB（千兆）级跃升到TB（万亿字节）级、PB（千万亿字节）级甚至更大。一家大型互联网平台每天产生的数据量可能超过许多中小型企业全年的数据总量。

Velocity（高速）：数据产生和需要处理的速度极快。社交媒体上每秒钟产生数千条帖

子，金融市场每毫秒产生数百笔交易，自动驾驶汽车的传感器每秒生成 GB 级的数据。很多场景要求数据“边产生边处理”（实时处理），而不是“先存储再分析”。

Variety（多样）：数据的类型不再局限于传统的结构化表格（如数据库中的行列数据），还包括大量的非结构化数据：文本（新闻、社交媒体帖子）、图像（照片、视频）、音频（语音、音乐）、日志（网站访问记录、设备运行日志）等。据估计，非结构化数据占全部数据的 80％以上。

Value（价值）：大数据虽然规模庞大，但有价值的信息往往隐藏在海量“无用”数据之中——就像模块六（数据挖掘）中比喻的“在矿山中淘金”。大数据的价值需要通过数据挖掘和 AI 技术来提取。

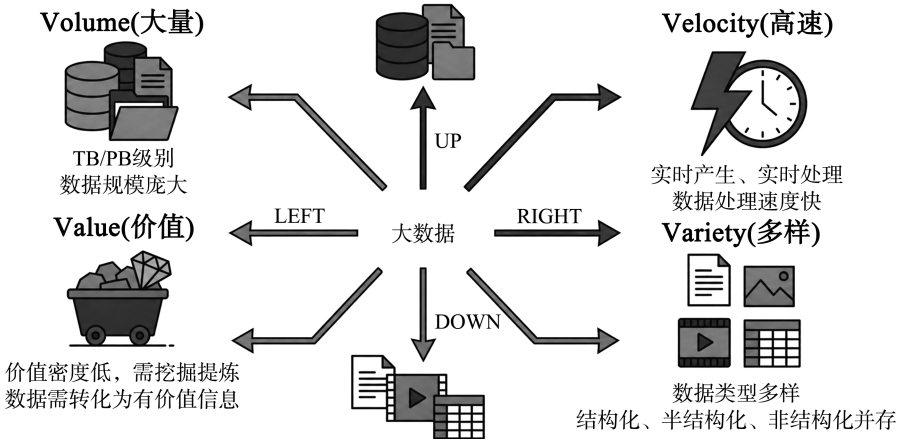

图 9-4　大数据的“4V”特征

二、大数据带来的思维变革

大数据不仅是一种技术现象，更是一种思维方式的革命。它重塑了认知世界与实施决策的根本逻辑。以下三个方面的变革尤为深刻。

1. 从“抽样”到“全量”

在大数据时代之前，由于数据采集和存储的成本很高，研究者通常只能分析数据的一个“样本”（部分数据），然后用统计方法推断“总体”的特征。这就是抽样分析的思路——模块二任务三中“训练集/测试集划分”的背后也蕴含着这种思路。

大数据时代的一个根本性变化是：在很多场景中，可以直接分析全部数据而不是样本。当电商平台可以记录所有用户的每一次点击行为时，为什么还要“抽样”？当城市的所有出租车都装有 GPS 时，为什么还要选几辆车做调查？全量数据的好处是：不存在“样本偏差”问题，可以发现小众但有价值的模式（“长尾”规律），可以进行极其精细的细分分析。

2. 从“因果”到“相关”

传统的科学研究追求“因果关系”：A 导致了 B（如“吸烟导致肺癌”）。建立因果关系需要严格的实验设计和控制变量，过程漫长而昂贵。大数据思维则强调：在很多实际应用中，发现“相关性”就已经足够有价值了，不一定需要追问“为什么”。

模块六思想引领中曾以“冰淇淋销量与溺水率”为例，阐释“相关不等于因果”这一核心观点。在那里强调了“不要混淆相关与因果”的思维警示。但从实用角度看，很多商业决策确实不需要因果解释：如果数据显示“买了尿布的人经常买啤酒”（模块六任务一的经典故事），超市把两者放在一起就能提升销量——至于“为什么”买尿布的人也买啤酒，对商业决策来说并不重要。

当然，“因果”和“相关”各有其价值：在科学研究、医疗决策和政策制定等场景中，因果关系仍然至关重要；而在商业推荐、营销优化等场景中，相关性往往就够用了。大数据思维的“新”不在于否定因果，而在于认识到：对于很多实际问题，“知道什么和什么相关”就能指导行动，不必等到“弄清楚为什么”。

3. 从“精确”到“趋势”

传统的小数据分析追求精确：每一个数据点都要尽可能准确，分析结果要精确到小数点后几位。大数据思维则接受一定程度的“不精确”：当拥有数亿条数据时，个别数据点的错误对整体分析结果的影响微乎其微。重要的不是每个数据点是否完全准确，而是整体的趋势和模式是否清晰。

用一个比喻：传统数据分析像用显微镜观察一滴水——每个细节都看得很清楚，但视野很窄；大数据分析像坐在飞机上俯瞰大地——个别房屋看不清楚，但城市的布局、河流的走向、交通的脉络一目了然。两种视角各有价值，但在宏观决策中（如城市规划、宏观经济分析），“俯瞰”的视角往往更有意义。

表 9-3　大数据思维与传统数据分析思维的对比

维度	传统数据分析思维	大数据思维
数据范围	抽样（分析部分数据）	全量（分析所有数据）
分析目标	追求因果关系（为什么?）	接受相关关系（什么和什么相关?）
精度要求	追求精确（每个点都要准）	接受趋势（整体模式比个别点重要）
数据类型	主要是结构化数据（表格）	结构化＋非结构化（文本/图像/视频）
分析速度	离线分析（先存后算）	实时分析（边产生边处理）
通俗比喻	“显微镜”：看得清但视野窄	“飞机上俯瞰”：看不清细节但全景清晰

三、大数据的技术基础

大数据思维的落地离不开一系列关键技术的支撑。对于通识课的读者来说，不需要

深入了解这些技术的细节，但了解它们的名称和基本功能有助于建立完整的知识框架。

分布式存储与计算：当数据量超过单台计算机的处理能力时，就需要将数据“分散”存储在多台计算机上，让它们协同计算。这就是分布式计算的思想。Hadoop 和 Spark 是大数据领域最著名的两个分布式计算框架，它们使得处理 PB 级别的数据成为可能。用模块七（群体智能）的语言来说，分布式计算就是“一群普通的计算机通过协作，完成一台超级计算机才能完成的任务”。

数据仓库与数据湖：企业需要一个“中央仓库”来存储和管理海量的数据。数据仓库存储经过清洗和整理的结构化数据，适合做报表和分析；数据湖则存储各种格式的原始数据（包括非结构化数据），适合做更灵活的探索性分析。

实时流处理：对于需要“边产生边分析”的数据（如金融交易监控、社交媒体舆情分析），需要流处理技术。与传统的“先存储再分析”（批处理）不同，流处理是“数据一到就立刻分析”，实现秒级甚至毫秒级的响应。

案例：搜索引擎——大数据的“杀手级”应用

搜索引擎是大数据时代最具代表性的产品之一。以一次典型的搜索请求为例：

数据量（Volume）：搜索引擎需要索引全球数百亿个网页（存储量达 PB 级别），每次搜索都需要在这个海量索引中查找匹配结果。

速度（Velocity）：从按下搜索键到结果显示在屏幕上，通常不超过 0.5 秒。在这不到半秒的时间里，系统完成了查询解析、索引检索、结果排序、广告匹配和页面渲染等一系列操作。

多样性（Variety）：搜索引擎不仅索引文字网页，还索引图片、视频、新闻、地图、学术论文等多种类型的数据。

价值（Value）：在数百亿个网页中，与检索查询真正相关的结果通常仅有几十个。——搜索引擎的核心价值就是从海量的“低价值密度”数据中，筛选出最相关的少数结果。

搜索引擎背后综合运用了大数据存储（分布式索引）、AI 算法（网页排序、语义理解）、实时计算（毫秒级响应）等多种技术，是“大数据＋AI”协同工作的典范。

本节小结

1. 大数据的“4V”特征：大量（Volume）、高速（Velocity）、多样（Variety）、价值（Value）。其中“多样”和“高速”是大数据区别于传统数据的最核心特征。

2. 大数据带来了三大思维变革：从“抽样”到“全量”、从“因果”到“相关”、从“精确”到“趋势”。这些变革改变了人们认识世界和做决策的底层逻辑。

3. 大数据的技术基础包括：分布式存储与计算（Hadoop/Spark）、数据仓库与数据湖、实时流处理等。

4. “从因果到相关”的思维转变需要辩证看待：相关性在商业场景中通常足够；但在科学研究和高风险决策中，因果关系仍然不可或缺（模块六思想引领的“相关不等于因果”

依然是重要的思维警示）。

思 考 题

1.［识记］请用自己的话解释大数据的“4V”特征，并各举一个生活中的例子。

2.［理解］“从因果到相关”的思维变革意味着“不需要知道为什么，只需要知道是什么”。请分析：在什么场景中这种思维转变是合理的？在什么场景中可能是危险的？（提示：对比“超市商品摆放优化”和“新药研发”两个场景）

3.［应用］请用大数据的“4V”框架分析你最常用的一个手机 App（如短视频平台、外卖平台、社交媒体）的数据特征：a. Volume：这个 App 每天大约产生多少数据？（不需要精确，量级即可）b. Velocity：它的数据是实时产生的还是批量产生的？c. Variety：它涉及哪些类型的数据（文字/图片/视频/位置/行为记录）？d. Value：它如何从海量数据中提取价值？（推荐算法？广告精准投放？用户画像？）

任务四　大数据与人工智能

◆ 引导思考

在模块一任务三中了解到“数据是 AI 三大引擎之一”。在本模块任务三中认识了大数据的 4V 特征和思维变革。现在把这两条线索合起来：大数据和 AI 之间到底是什么关系？是“大数据推动了 AI”，还是“AI 赋能了大数据”，还是两者兼有？

没有大数据，还有深度学习吗？没有 AI，大数据还有价值吗？

大数据与 AI 的结合在各行各业产生了怎样的“化学反应”？面临什么挑战？

一、大数据与 AI 的双向驱动关系

大数据与人工智能之间不是单向的“谁推动了谁”的关系，而是双向驱动、互为前提的共生关系。

1. 大数据驱动 AI：“燃料”效应

在模块一任务三中已经详细讨论过这个方向：大数据为 AI 模型提供了前所未有的“训练素材”。模块二中学过的机器学习、模块三中学过的深度学习，都是“数据越多性能越好”的技术——没有互联网时代的海量数据积累，2012 年深度学习的爆发是不可想象的。

但大数据对 AI 的推动不仅仅是“量多”，更体现在“多样性”上。大语言模型（模块十将详细讨论）之所以能够“博学多才”（既能写诗又能编程），正是因为它在训练阶段“阅读”

了互联网上涵盖几乎所有领域和语言的海量文本。如果训练数据只有一种类型(比如只有新闻文章),模型的能力就会非常狭窄。大数据的多样性(4V 中的 Variety)对 AI 的“通用能力”至关重要。

2. AI 赋能大数据:“引擎”效应

反过来,AI 为大数据提供了最强大的“分析引擎”。正如模块六任务一中的比喻:大数据是“矿山”,但矿石本身没有直接价值,需要通过“挖掘”(分析)才能变成“黄金”(知识和洞察)。AI(尤其是机器学习和深度学习)正是完成这种“挖掘”的最有效工具。

没有 AI,大数据就只是一堆“沉睡的数字”——企业可能拥有 PB 级的客户行为数据,但如果没有 AI 算法来分析这些数据、发现模式、作出预测,这些数据就无法创造商业价值。可以说:大数据是 AI 的“燃料”,AI 是大数据的“引擎”——缺了任何一个,另一个都无法充分发挥价值。

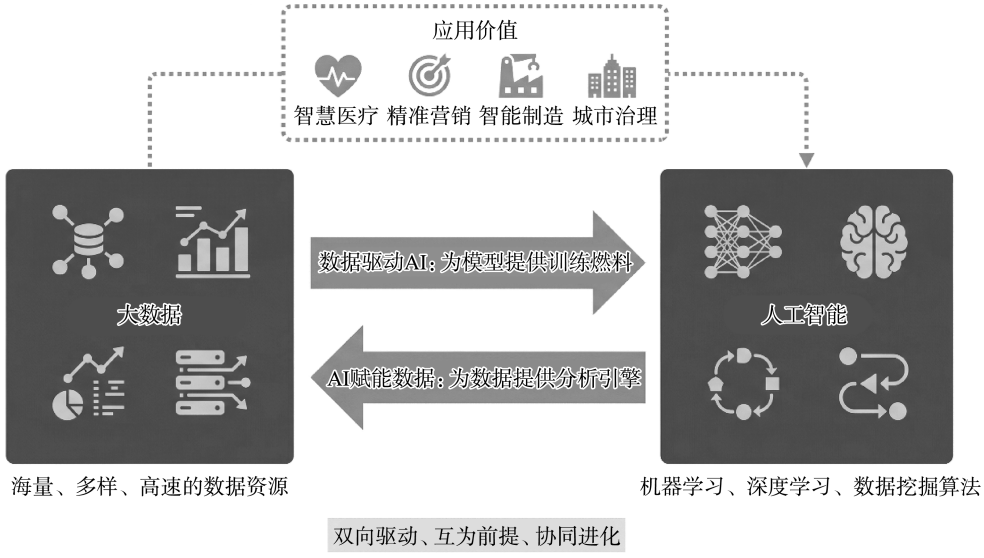

图 9-5　大数据与 AI 的双向驱动关系

二、“大数据+AI”的行业应用

大数据与 AI 的结合正在各行各业产生“化学反应”。以下是几个最具代表性的领域。

精准营销:电商平台利用用户的海量行为数据(浏览记录、购买历史、搜索关键词),通过 AI 算法(模块六中介绍的关联规则、推荐算法)精准预测“这位用户接下来可能想买什么”,实现“千人千面”的个性化推荐。这背后的逻辑就是:大数据提供了“每个用户的行为画像”(燃料),AI 从画像中发现偏好模式(引擎),两者结合产生了“精准推荐”(价值)。

智慧医疗:医疗机构积累了海量的电子病历、影像数据、基因组数据和可穿戴设备数据。AI 算法(模块八中介绍的医学影像分析、模块六中介绍的数据挖掘)可以从这些数据中发现疾病的早期信号、预测治疗效果、辅助药物研发。“精准医疗”的核心理念就是用大

数据和 AI 为每位患者制定个性化诊疗方案。

智慧城市：城市中的交通流量数据、能耗数据、环境监测数据、公共服务数据，构成了一个庞大的城市“数字孪生”。AI 可以从这些数据中优化交通信号配时（减少拥堵）、预测用电高峰（优化电力调度）、监测空气质量变化趋势（预警污染事件）。

智能制造：工厂中的设备传感器每秒产生大量的运行数据（温度、振动、压力等）。AI 通过分析这些数据，可以实现模块一任务四中介绍过的预测性维护（提前发现设备故障苗头）和质量预测（在产品下线前就预测其质量是否达标）。

表 9-4　大数据＋AI 在各行业的典型应用

行业	数据来源	AI 技术	典型应用	核心价值
电商/零售	用户行为数据	推荐算法、关联挖掘	个性化推荐	提升转化率
医疗健康	病历、影像、基因组	图像分类、数据挖掘	辅助诊断、精准医疗	提升诊断准确率
城市管理	交通/能耗/环境数据	预测模型、优化算法	交通优化、能耗管理	提升城市运行效率
制造业	设备传感器数据	异常检测、预测模型	预测性维护、质量预测	减少停机和废品
金融	交易数据、行为数据	异常检测、分类模型	反欺诈、信用评估	降低风险

三、挑战与趋势

大数据与 AI 的结合虽然前景广阔，但也面临着几个重要的挑战。

数据质量问题：在模块一任务三和模块二任务三中反复强调过：“脏数据”会导致“错误的智能”。大数据的“大”不等于“好”——海量数据中可能充斥着缺失值、错误标注、过时信息和系统性偏差。“垃圾进，垃圾出”的原则在大数据时代同样适用，甚至更加严峻。

数据隐私与安全：大数据的商业价值往往建立在对个人行为数据的收集和分析之上。如何在“利用数据创造价值”和“保护个人隐私”之间找到平衡，是技术、法律和社会共同面临的难题。模块六思想引领中讨论的“数据伦理”问题（大数据杀熟、保险数据歧视等）在大数据与 AI 的语境下变得更加尖锐。

算力成本：大数据＋AI 的结合对计算资源的需求极其庞大。训练一个大型 AI 模型的算力成本动辄数百万甚至上千万美元（模块一任务三中讨论过），这使得前沿的大数据与 AI 的研究和应用越来越集中在少数“算力富裕”的大型科技公司手中，可能加剧技术的“不平等”。

知识拓展：“数据飞轮”——大数据与 AI 的正反馈循环

大数据和 AI 之间存在一种类似模块七中“正反馈”的增强效应，被业界形象地称为“数据飞轮”：更多的数据→训练出更好的 AI 模型→更好的产品体验→吸引更多的用户→产生更多的数据→进一步改进 AI 模型……如此循环往复，“飞轮”越转越快。

这个正反馈循环解释了为什么在很多 AI 驱动的行业中，“先发优势”如此重要：最先积累了大量用户和数据的企业，其 AI 产品会越来越好，吸引更多用户，产生更多数据，进一步拉开与竞争对手的差距。搜索引擎、社交媒体、短视频平台等领域的“赢家通吃”现象，很大程度上就是“数据飞轮”效应的体现。

但“数据飞轮”也引发了关于市场竞争公平性的讨论：当一家公司通过数据飞轮建立了难以逾越的“数据护城河”时，新的竞争者是否还有进入的机会？这也是各国反垄断监管机构越来越关注“数据垄断”问题的原因之一。

本节小结

1. 大数据与 AI 之间是双向驱动的共生关系：大数据为 AI 提供“训练燃料”（数据驱动），AI 为大数据提供“分析引擎”（AI 赋能）。

2. 大数据与 AI 的结合在精准营销、智慧医疗、智慧城市、智能制造和金融风控等领域已经产生了显著的应用价值。

3. 大数据与 AI 的深度融合虽前景广阔，但在数据质量、隐私安全及算力成本方面面临着较严峻的挑战。

4. “数据飞轮”效应使得大数据与 AI 领域呈现“强者越强”的正反馈循环，这引发了关于市场公平性和数据垄断的讨论。

思 考 题

1. ［识记］请用“燃料”和“引擎”的比喻，解释大数据与 AI 之间的双向驱动关系。

2. ［理解］“数据飞轮”效应为什么会导致“强者越强”的局面？这种效应对于新进入某个 AI 驱动的市场（如搜索引擎、短视频）的创业者来说意味着什么？

3. ［应用］假设你所在的大学想利用“大数据＋AI”来提升教学质量。请分析：a. 可以收集哪些类型的数据？（提示：考虑选课数据、成绩数据、课堂出勤、图书馆借阅、食堂消费等）b. 这些数据可以用 AI 做什么分析？（举 2—3 个具体例子）c. 在收集和使用学生数据的过程中，可能面临什么隐私和伦理问题？你认为应该如何平衡“数据利用”和“隐私保护”？

◆ 思想引领

一、拥抱“不精确”

本模块的两个主题——模糊逻辑和大数据——从不同角度传递了一个共同的信息：“不精确”不是缺陷，而是认识复杂世界的一种更诚实、更有效的方式。

模糊逻辑强调：现实世界中的很多概念（热、年轻、远、贵）本来就没有精确的边界，强行划定一条“非此即彼”的分界线反而是对现实的扭曲。用“隶属度”来承认和表达这种模糊性，比假装它不存在更加真实。

大数据思维强调：当拥有足够多的数据时，可以用“趋势”代替“精确”，用“相关”代替“因果”——不是因为精确和因果不重要，而是因为在很多实际场景中，“大致正确”比“精确但迟到”更有价值。

对于非技术专业的学生来说，这种“拥抱不精确”的认知态度有着超越技术本身的启示意义：在学习、工作和生活中，人们往往过于追求“确定的答案”和“精确的结论”，而忽略了很多问题本身就没有确定的答案。学会在“不确定”的环境中作出“足够好”的判断（而不是等待“完美”的答案），是一种非常重要的实践智慧。

二、数据权力

任务四中讨论的“数据飞轮”效应揭示了一个深层的社会问题：数据正在成为一种新的“权力资源”。拥有最多用户数据的企业，能够训练出最好的 AI 模型，提供最好的产品体验，吸引更多用户，获取更多数据——这个正反馈循环使得“数据富者愈富”。

这引发了一系列值得深思的问题：当少数科技巨头掌握了数十亿人的行为数据时，这种“数据权力”是否需要制约？个人对自己产生的数据（浏览记录、消费记录、位置轨迹）有多大的控制权？国家和地区之间的“数据主权”争议（谁有权访问和使用跨境流动的数据？）将如何影响全球科技竞争？

这些问题没有简单的答案，但作为数据时代的公民，人们需要意识到：“数据”不仅仅是一个技术概念，更是一个涉及权力、公平和自由的社会议题。理解数据的“权力属性”，是在数字社会中作出知情公民选择的前提。这些问题将在模块十三中进一步深入讨论。

三、“人机协作”的两种模式

回顾本模块的两个主题，可以看到 AI 与人类知识结合的两种不同模式：

模糊逻辑模式：人类专家将自己的经验“翻译”成模糊规则（“如果温度有点高，那么功率调大一些”），计算机负责精确地执行这些规则。这种模式中，人类提供智慧（规则），计算机提供精度和速度——与模块四（专家系统）的逻辑一脉相承。

大数据＋AI 模式：计算机从海量数据中自动发现规律，人类负责提出问题、解读结果和作出决策。这种模式中，计算机发现规律，人类作出判断——与模块二（机器学习）和模块六（数据挖掘）的逻辑一脉相承。

两种模式各有所长，适用于不同的场景（数据少但经验丰富的场景适合模糊逻辑；数据多但规律复杂的场景适合大数据＋AI）。更重要的是，它们共同揭示了一个贯穿全书的核心命题：AI 最大的价值不在于“替代”人类，而在于“与人类协作”——机器做机器擅长的事（处理海量数据、保持精度和一致性），人类做人类擅长的事（理解问题、作出价值判断、承担责任）。

◆ 任务实践

以下实践任务旨在帮助读者将本模块所学的模糊逻辑和大数据知识与日常生活相结合。建议以小组(3—5 人)为单位完成。

实践一　“模糊控制器”纸笔设计

任务说明

目标:亲手设计一个简单的模糊控制系统,体验“从模糊规则到精确输出”的完整过程。

步骤:

(1) 选择一个日常场景(建议:智能电风扇/智能窗帘/智能浇花器/智能台灯亮度调节)。

(2) 确定输入变量(如电风扇:室温和湿度)和输出变量(如风扇转速),为每个变量定义 2—3 个模糊集合(如温度:“凉”“适中”“热”)。

(3) 用文字描述每个模糊集合的大致隶属度范围(如“28℃以下完全属于‘凉’,28—32℃在‘凉’和‘适中’之间渐变,32℃以上完全属于‘热’”)。

(4) 编写 5—8 条模糊规则(如“如果温度‘热’且湿度‘大’,则转速‘最快’”)。

(5) 给定一组测试输入(如“温度 30℃,湿度 65%”),手工走一遍“模糊化→推理→去模糊化”的流程,计算出一个大致的输出值。

讨论:a. 编写模糊规则时,你依据的是什么?(个人经验?常识?查资料?)这与模块四中专家系统的“知识获取”有什么相似?b. 如果同一个输入触发了多条规则且建议不同,你是怎么“综合”的?

评价维度:模糊集合定义的合理性、规则的逻辑一致性、推理过程的正确性。

实践二　“我的数字画像”大数据体验与反思

任务说明

目标:通过实际观察和分析,直观感受大数据对个人生活的渗透,并反思数据隐私问题。

步骤:

(1) 每位同学检查以下内容(能查到多少算多少):a. 手机的“屏幕使用时间”统计(你每天用多长时间?哪些 App 用得最多?);b. 短视频平台或电商平台的“推荐页面”(连续刷 10 条推荐内容,记录与你兴趣的匹配度);c. 地图 App 的“时间线”或“足迹”功能(它记录了你去过的哪些地方?);d. 任何你能找到的“数据导出”功能(部分 App 允许你下载自己的数据)。

(2) 小组汇总后讨论:a. 这些平台对你的“了解”程度超出你的预期了吗?b. 推荐系统的“精准度”如何?它有没有“推错”过?为什么?c. 你对这些数据被收集和使用有什

么感受？觉得哪些是合理的？哪些让你不太舒服？

（3）撰写一份不超过 1 000 字的“我的数字画像”反思报告。

评价维度：观察的细致程度、对数据收集范围的认知、对隐私问题的反思深度。

实践三　“精确 vs 模糊”辩论

任务说明

目标：通过辩论深化对“精确性”和“模糊性”各自价值的理解。

步骤：

（1）辩题：“在做决策时，精确的量化信息永远比模糊的定性描述更有价值。”

正方论据方向：精确数据减少歧义、便于比较、可以量化优化（如“投资回报率 12.3%”比“回报不错”更有决策价值）。

反方论据方向：很多重要的决策信息天然是模糊的（如团队士气“比较高”、市场前景“不太明朗”）；过度追求精确可能导致“虚假的确定感”（一个精确到小数点后三位的错误数字比一个“大概”的正确判断更危险）；模糊逻辑的成功证明了“恰当的模糊”有时比“强行的精确”更有效。

（2）辩论后每位同学写一段 300 字的个人反思：在你自己的学习和生活中，有没有遇到过“追求精确反而帮了倒忙”或者“模糊判断反而更有效”的经历？

评价维度：论据的逻辑性和说服力、对“精确”与“模糊”各自适用场景的辨析深度、个人反思的独立性。

模块十

大模型

◆ 模块导读

模块一任务一介绍了“大模型方法”，即“一个模型、适配多任务”的 AI 新范式；模块一任务二回顾了 2017 年 Transformer 架构的诞生，以及 2022 年 ChatGPT 带来的全球影响；模块三系统梳理了神经网络原理。现在，编者将这些线索汇聚，深入介绍当今 AI 最核心、最受关注的技术方向——大模型。

大模型为什么“大”？它的“大”仅仅是参数数量多吗？它的核心架构——Transformer——与模块三中学过的传统神经网络有什么根本区别？注意力机制是什么？为何它是大模型成功的关键？大模型如何从海量文本中习得写作、翻译、编程、推理等复杂能力？

本模块将围绕以上核心问题展开：任务一从宏观概述大模型概念、架构与发展历程；任务二铺垫自然语言处理基础；任务三解析大语言模型技术细节与代表应用；任务四探讨其局限与挑战。通过学习，读者将深入理解这项深刻影响社会的 AI 技术，形成理性认知。

◆ 学习目标

【了解】

—大模型的基本概念和定义，以及“大”的含义（参数规模、训练数据量、计算资源）；

—Transformer 架构和“注意力机制”的基本功能（直觉层面）；

—代表性大语言模型的名称和主要特点（GPT 系列、BERT、文心一言、通义千问等）。

【理解】

—大模型“预训练＋微调/提示”范式：为何“先博学、后专精”，比从零训练更高效；

—注意力机制的直觉含义：模型如何聚焦输入中的关键信息；

—大模型涌现的意外能力，如推理、零样本学习，及其对 AI 发展的深层意义。

【掌握】

—区分大模型的不同类型（语言模型、视觉模型、多模态模型）及其各自的能力范围；

—根据具体任务需求，判断是否需要使用大模型，以及应该选择预训练、微调还是提示工程的方式。

【能够】

—结合前九个模块知识，理解大模型如何整合深度学习、大数据与自然语言处理等技术；

—理性认识大模型的能力与局限，避免陷入“AI 万能”或“AI 无用”的极端认知。

思维导图

下图展示了本模块的知识结构与内容框架，帮助读者从整体上把握各任务之间的逻辑关系：

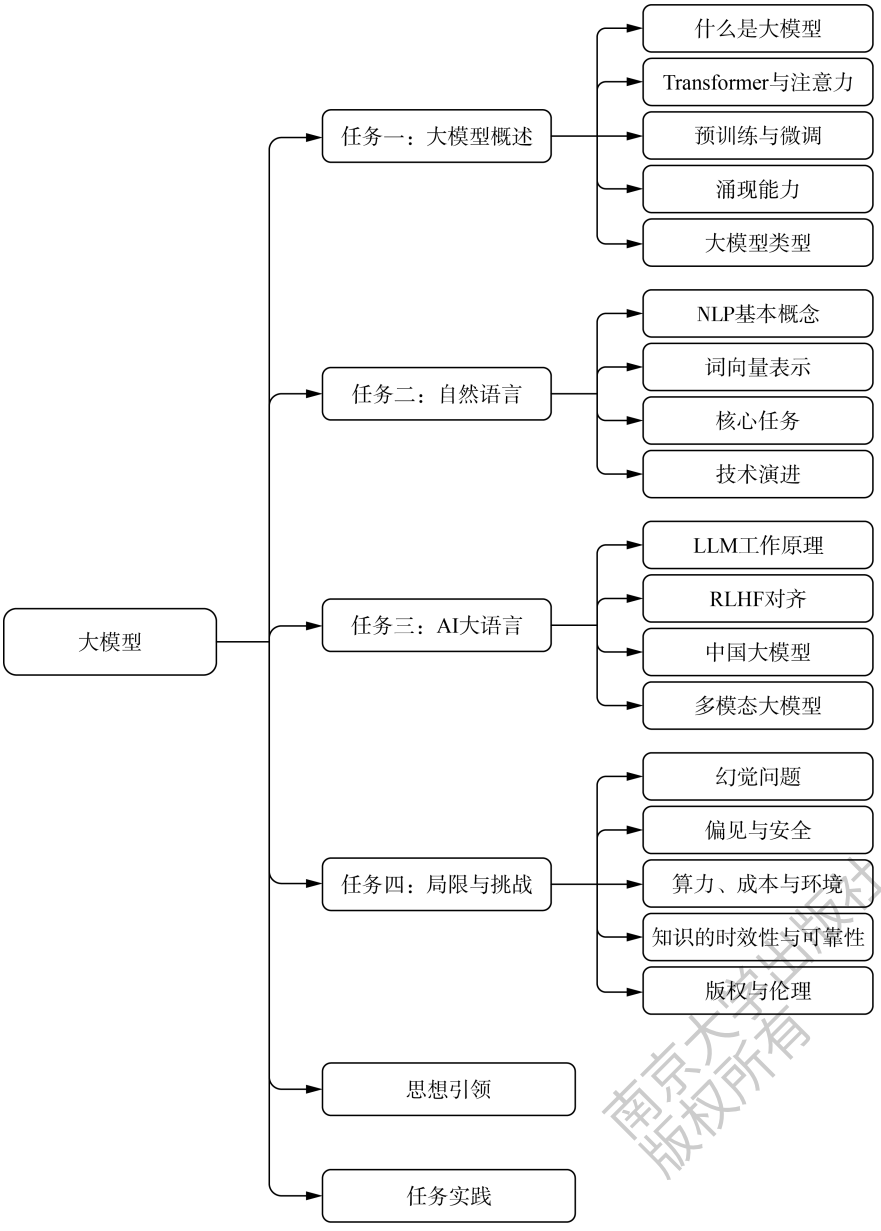

图 10-1　模块十知识结构思维导图

任务一　大模型概述

◆ 引导思考

读者可能已经使用过 ChatGPT、文心一言或通义千问等 AI 对话工具。它们能写诗、翻译、编程、总结文章，似乎无所不能。但有没有想过：这些能力从何而来？一个程序，为何能学会写作与推理？

在模块一中，编者用“先博学后专精”来比喻大模型的工作方式。大模型是如何实现“博学”的？又是如何做到“专精”的？其核心架构 Transformer，究竟有何独特优势？

当模型参数从百万级增至千亿级，会带来什么变化？仅仅是更精准，还是会涌现出意料之外的全新能力？

一、什么是大模型

大模型是参数规模达数十亿乃至数千亿级的深度学习模型。其“大”体现在三方面：参数量大（数十亿至数万亿可调参数）、训练数据量大（互联网级文本、图像等海量数据）、计算资源消耗大（数千块 GPU 训练数周至数月）。

模块三中神经网络参数可看作待学习的“旋钮”，控制网络对输入的敏感度。小模型如同拥有少数旋钮的简单收音机，能力有限；千亿参数的大模型，则像拥有海量微调旋钮的超级调音台，可生成极为丰富精细的能力。

大模型的“大”不只是量的积累——规模突破临界点后，会涌现小模型不具备的全新能力。这使大模型成为 AI 史上的质变里程碑，而非简单的“更大神经网络”。

二、Transformer 与注意力机制

绝大多数当前的大模型都建立在一种叫作 Transformer 的神经网络架构之上。Transformer 由谷歌团队于 2017 年提出，它是近十年来 AI 领域最重要的架构创新，可以说是大模型时代的“心脏”。

要理解 Transformer，关键是理解它的核心机制——注意力机制。

1. 注意力机制的直觉理解

读者可以用日常阅读场景理解注意力机制：读到“小明把苹果递给了小红，她说了声谢谢”时，大脑会自动回溯前文，确定“她”指代小红，而非小明或苹果。这就是大脑在词语间建立关联：理解“她”时重点关注“小红”，弱化对“苹果”的关注。

注意力机制让神经网络具备类似能力：处理文本等序列数据时，模型能自主学习词与

词之间的关联。对每个词，机制会计算它与所有词的关联度，据此决定重点参考哪些信息，从而精准捕捉长距离依赖，比如句首主语与句末代词的指代关系。

Transformer 出现前，序列处理主流用循环神经网络（RNN），它逐字从左到右处理，每次仅关注一个词。但句子过长时，早期信息会随传递不断衰减，如同传话游戏中信息逐渐失真。而 Transformer 的注意力机制完美解决了这一问题：每个词可直接关联序列中所有词，无需逐步传递，彻底避免了信息衰减。

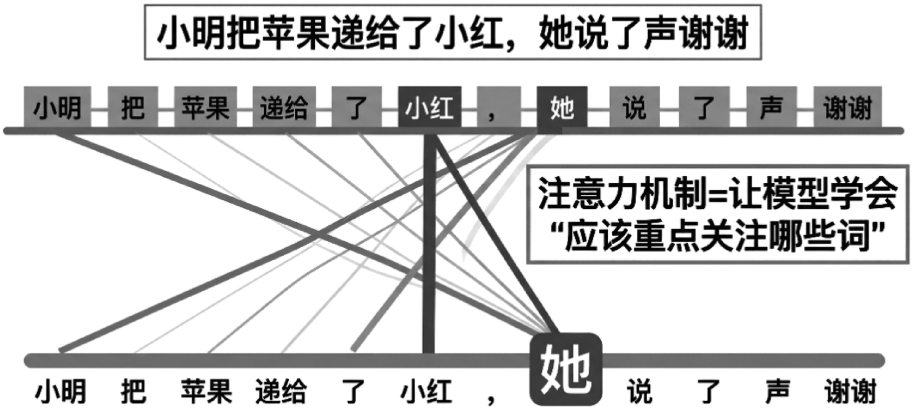

图 10-2　注意力机制的直觉理解

2. Transformer 的结构概览

Transformer 由多层注意力模块堆叠构成，每层主要完成两项运算：一是多头注意力机制，从多重维度建立语义关联，多角度解析文本信息；二是前馈网络，对注意力输出结果进一步加工提炼。模型堆叠层数越多、单层参数规模越大，所能学习与捕捉的复杂语言规律便越丰富。

Transformer 还具备极强的并行运算优势，区别于 RNN 逐序逐字处理的模式，它可同步处理序列内全部文本单元。该特性能够充分发挥 GPU 并行算力优势，大幅缩减训练时长，也让千亿级超大规模模型的训练落地成为现实。

表 10-1　Transformer 与传统循环神经网络（RNN）的对比

维度	循环神经网络（RNN）	Transformer
处理方式	从左到右逐字串行处理	全局词汇同步并行处理
长距离依赖	信息逐层传递易衰减，类似传话失真	字词间直接关联，全局信息互通
核心机制	循环衔接，输出依赖上一节点结果	依托注意力机制，计算全部词汇关联度
训练速度	算力利用率低，训练耗时久	适配 GPU 并行运算，训练效率高
可扩展性	难以扩展到超大规模	支持数千亿参数的超大模型
通俗比喻	一字一字朗读一本书	一目十行，全局理解

三、预训练与微调

在模块一任务一中，用音乐学院学生的比喻介绍过大模型“预训练＋微调”的范式。现在介绍这两个阶段。

1. 预训练：在海量数据中“博览群书”

预训练是大模型的第一阶段：在海量的无标注数据上学习通用的语言表示能力。以大语言模型为例，预训练的核心任务通常是“预测下一个词”：给模型一段文本的前半部分，让它预测接下来最可能出现的词是什么。

这个任务看起来简单，但隐藏着深刻的含义：要准确预测“下一个词”，模型必须理解语法（“the”后面通常接名词）、语义（“医生”后面更可能接“诊断”而非“烹饪”）、常识（“太阳从东边升起”），甚至逻辑推理（“如果 $A>B$ 且 $B>C$，那么 $A>C$”）。当这个“猜词游戏”在互联网级别的海量文本上进行数万亿次后，模型就“悟”出了人类语言中蕴含的极其丰富的知识和规律。

预训练的数据规模是惊人的：截至本书编写时，主流大语言模型的预训练数据量达到数万亿个词元（约相当于数百万本书的信息量）。预训练的计算成本同样惊人：一次完整的预训练可能需要数千块高端 GPU 运行数月，花费数百万甚至上千万美元。这也是为什么只有少数大型科技公司（如 OpenAI、谷歌、百度、阿里巴巴等）有能力从头训练大模型。

2. 微调与提示：让“通才”变“专家”

预训练结束后，模型已具备完备的通用认知能力，但无法直接适配各类细分场景的专属需求。业界主要通过两种方式实现模型的领域专精：

微调：基于特定任务的标注数据对预训练模型进行二次训练，针对性优化模型的专项性能。例如，搭建医学问答机器人时，可采用海量医学问答数据集微调模型。相较于预训练，微调所需的数据量与计算资源大幅减少，能够以较低成本显著提升模型在垂直领域的

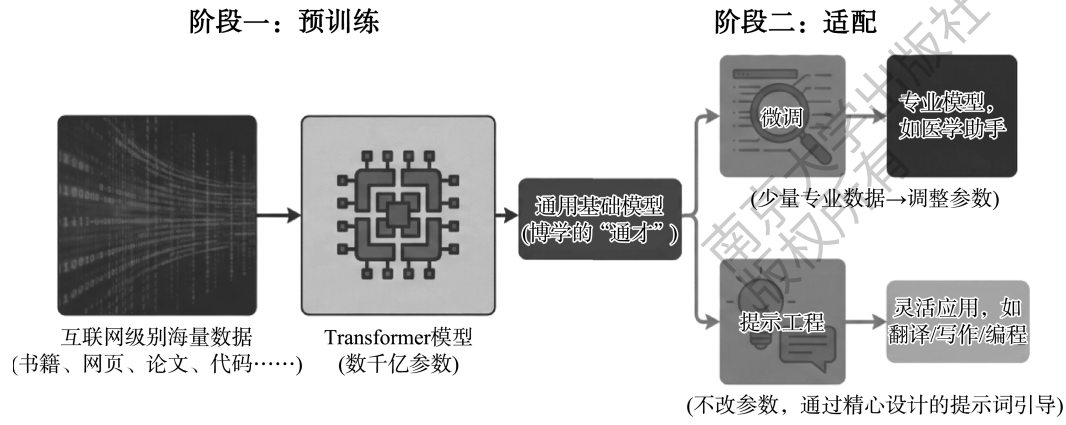

图 10-3　大模型的“预训练＋微调/提示”两阶段范式

表现。

提示工程：该方式无需改动模型参数，仅通过设计精细化的提示词，引导模型输出符合场景需求的内容。比如在提问前设定专业身份与作答规范，模型便会自适应输出对应的专业风格。模块四曾提及提示工程与专家系统规则的相似性，结合大模型技术逻辑，如今可形成更为全面、完整的认知。

四、涌现能力

大模型最令人震撼且极具研究价值的特性，便是涌现能力。部分能力在小规模模型中完全不存在，一旦模型参数量突破临界规模，便会突然生成全新的智能表现。

小型语言模型仅能完成基础的文本补全任务，而当参数量攀升至百亿级别后，模型会自发具备多项全新能力：思维链推理，可分步拆解、推导复杂问题；零样本学习，无需示范样本，仅凭指令即可完成全新任务；跨语言翻译，即便训练数据以英语为主，也能自主掌握多语言互译能力。这些能力并非人工预设编程所得，而是模型规模达标后自发涌现的全新特性。

该现象与模块三、模块七提及的涌现思想高度契合：简单的基础组件，在规模足够庞大时，会诞生远超个体单元的复杂能力。不同于群机器人、蚁群系统依靠个体交互实现涌现，大模型的涌现源于参数规模的量变积累，但二者共同印证了核心规律：在特定条件下，系统的量变终将引发智能的质变。

案例：GPT 系列——从“文本补全工具”到“通用 AI 助手”的进化

OpenAI 的 GPT（Generative Pre-trained Transformer）系列是大语言模型发展的标志性里程碑：

GPT-1（2018，约 1.2 亿参数）：证明了“预训练＋微调”范式的有效性，但能力有限，主要用于学术研究。

GPT-2（2019，约 15 亿参数）：写作能力已经相当流畅，OpenAI 一度因担心“假新闻”风险而延迟公开发布。

GPT-3（2020，约 1750 亿参数）：实现了“零样本学习”（不需要微调，仅通过提示词就能完成各种任务），首次展示了大模型的“涌现”能力。

ChatGPT/GPT-4（2022—2023）：在 GPT 系列的基础上，通过人类反馈强化学习（RLHF）技术进一步优化了对话体验和指令遵循能力。ChatGPT 的发布引发了全球范围的 AI 热潮，被认为是 AI 走向“人人可用”的标志性事件（模块一任务二已详细讨论）。

中国的大语言模型也在快速发展：百度的文心一言、阿里巴巴的通义千问、科大讯飞的星火、智谱 AI 的 ChatGLM 等，在中文对话和多种应用场景中展现出了强劲的竞争力。

五、大模型的类型

虽然大语言模型(Large Language Model，LLM)是目前最受关注的大模型类型，但大模型的概念远不止于语言。按照处理的数据类型，大模型可以分为三大类。

表 10-2　大模型的三大类型

类型	处理的数据	代表模型	典型应用
大语言模型(LLM)	文本	GPT 系列、文心一言、通义千问	对话、写作、翻译、编程、推理
大视觉模型	图像/视频	Stable Diffusion、DALL·E、Sora	图像生成、视频生成、图像理解
多模态大模型	文本＋图像＋音频等	GPT-4(多模态版)、Gemini	图文理解、“看图说话”、“听音回答”

模块八曾介绍 GAN 与扩散模型两类图像生成技术。大视觉模型融合扩散模型思路与 Transformer 架构，可依据文本描述生成高清图像与视频。多模态大模型能力更为全面，能够同步处理文本、图像、音频信息，完成跨模态理解与创作，例如识图作答。这也契合 AI 产品由单模态逐步迈向多模态的发展趋势。

知识拓展：“规模法则”——为什么“更大”意味着“更强”？

研究者们发现了一个令人惊讶的规律，被称为“规模法则”(Scaling Law)：大模型的性能与三个因素呈现出稳定的、可预测的幂律关系——参数规模越大、训练数据越多、计算量越大，模型的性能就越好。更令人惊讶的是，这种关系在目前观察到的规模范围内似乎还没有“饱和”的迹象。

规模法则的深层含义是：只要持续增加模型规模、数据和算力，模型的能力就会持续提升。这也是为什么科技公司纷纷投入巨资建设超大规模计算集群来训练更大的模型——它们相信“更大”几乎必然意味着“更强”。

但规模法则也引发了关于可持续性的担忧：如果“更好的 AI”永远需要“更多的资源”，那么 AI 的发展是否最终会受到能源和成本的制约？这是模块一任务三中讨论过的“算力挑战”在大模型时代的升级版。

六、从 NLP 到大模型

为了帮助读者建立大模型的“前因后果”认知，用一条时间线梳理了从自然语言处理(NLP)到大模型的关键里程碑：

表 10-3　从 NLP 到大模型的发展里程碑

时间	事件	意义
2013	Word2Vec 词向量	让计算机用“数字”表示词的“含义”，开启语义表示时代

续表

时间	事件	意义
2017	Transformer 架构发表	“注意力就是你所需要的一切”，大模型的架构基石
2018	GPT-1/BERT 发布	“预训练＋微调”范式确立：GPT 擅长生成，BERT 擅长理解
2020	GPT-3（1750 亿参数）	零样本学习、涌现能力，证明了规模法则
2022	ChatGPT 发布	对话式 AI 走向大众，引发全球 AI 热潮
2023＋	GPT-4（多模态）、文心一言、通义千问等	多模态能力、中国大模型快速崛起

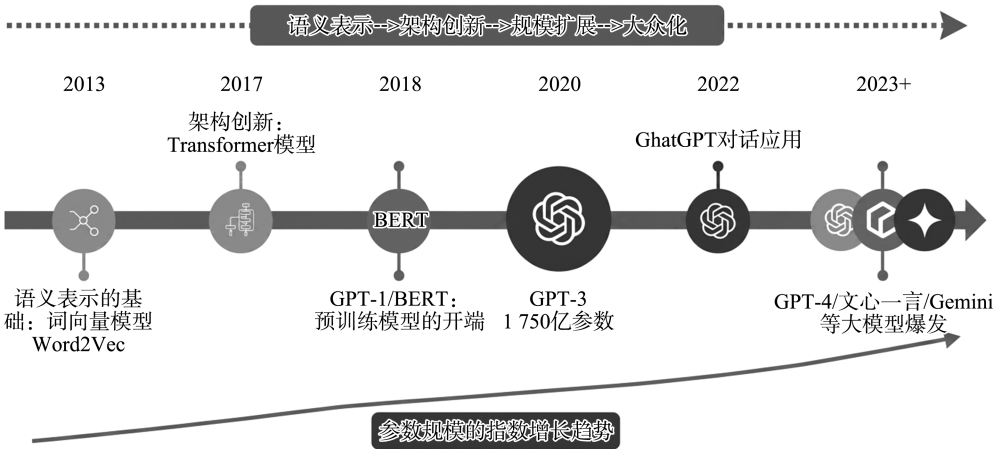

图 10-4　大模型发展时间线

本节小结

1. 大模型是参数规模达数十亿到数千亿的深度学习模型，其“大”体现在参数量、训练数据量和计算消耗三个维度。

2. Transformer 是大模型的核心架构，其注意力机制让模型能够“关注”输入中最重要的部分，解决了 RNN 的“信息衰减”问题，并支持高效的并行计算。

3. 大模型的“预训练＋微调/提示”范式：先在海量数据上学习通用能力（“博学”），再通过微调或提示词适配到具体任务（“专精”）。

4. 涌现能力是大模型最引人注目的现象：思维链推理、零样本学习等能力在小模型上不存在，但在模型规模超过临界点后突然出现。

5. 大模型的三大类型：大语言模型（文本）、大视觉模型（图像/视频）、多模态大模型（融合文本、图像、音频等多种模态）。规模法则证实了“更大即更强”，但也引发了对 AI 算力与能源成本的担忧。

思 考 题

1.［识记］请用自己的话解释 Transformer 的“注意力机制”是什么，并说明它比循环神经网络(RNN)好在哪里。(提示：可以用“传话游戏”的比喻)

2.［理解］大模型的“涌现能力”(如零样本学习、思维链推理)为什么被认为是 AI 发展史上一个重要的现象？它与模块三和模块七中讨论过的“涌现”概念有什么异同？

3.［应用］假设你是一家律师事务所的技术负责人，想用大语言模型来辅助律师的日常工作(如合同审阅、案例检索、法律咨询)。请分析：a. 你会选择“微调”还是“提示工程”来适配模型？为什么？(提示：考虑律师行业对准确性和专业性的要求)b. 这个应用中，大模型的“幻觉”问题(生成看似合理但实际错误的内容)可能带来什么风险？你会采取什么措施来控制这个风险？

任务二　自然语言

◆ 引导思考

任务一从宏观角度介绍了大模型的概念和架构。而大模型中最受关注的类型是“大语言模型”——它处理的对象是“自然语言”。那么，“自然语言”对计算机来说意味着什么？为什么“让机器理解人话”被认为是 AI 最困难的问题之一？

计算机只能处理数字，但人类语言由词汇、语法、语义、语境甚至情感组成。怎样把一个“词”变成计算机能处理的“数字”？

从基于规则的机器翻译(模块一)到统计方法(模块二)再到深度学习(模块三)，NLP 技术经历了怎样的演进？大语言模型是如何站在这些前辈的“肩膀”上的？

一、自然语言处理

自然语言处理(Natural Language Processing，NLP)是人工智能的核心子领域，旨在赋予计算机理解、处理、生成人类自然语言的能力。在模块一任务一中，编者已将自然语言处理归为人工智能五大核心应用领域之一，本节将对其展开深入阐释。

人类自然语言对计算机而言极具挑战性，根源在于其三大核心特性。一是歧义性，同一词汇、句式可对应多重语义。例如“我看见他在河边钓鱼”，难以直接判定“在河边”的修饰对象；“苹果好吃”与“苹果股价上涨”中的“苹果”也分属不同概念。二是隐含性，人类交流往往省略常识与逻辑前提。比如听到“外面下雨了”便会自然联想到需要撑伞，无需言语赘述即可完成逻辑补全。三是语境依赖性，语句的语义与情感倾向需结合上下文判定。如“菜品不错，就是服务太差”，需依托语境才能精准判断整体评价倾向。

正是语言的这些复杂特性，让早期基于规则的传统 NLP 方法遭遇瓶颈，模块一中“伏特加是好的，但肉烂了”的翻译误区便是典型例证。纵观自然语言处理的技术演进，其本质就是持续优化算法，不断攻克语言歧义、语义隐含、语境依赖等难题的发展历程。

二、从“词”到“向量”

计算机处理语言的第一个根本性问题是：怎样把“词”变成“数字”？因为计算机的一切运算都建立在数字之上，如果不能把语言转化为数学形式，就无法对它进行任何处理。

1. 最简单的方法：独热编码

最直接的方法是给每个词分配一个编号。假设词典里有 10 000 个词，“猫”是第 1234 号，“狗”是第 5678 号。这种方法叫作独热编码：用一个 10 000 维的向量表示每个词，“猫”的向量在第 1234 个位置是 1，其余全是 0。

然而，独热编码存在明显短板，它默认所有词汇彼此相互独立。在此编码规则下，“猫”与“狗”、“猫”与“火箭”的向量间距并无差别，可从语义层面来讲，同属动物的猫狗关联性，远高于猫与航天器械。这种编码形式，无法体现词汇间潜藏的语义关联。

2. 词向量：用“位置”表达“含义”

2013 年，谷歌的托马斯·米科洛夫等人提出了 Word2Vec 方法，开创了词向量的时代。其核心思路简洁巧妙，不再以数字编号表征词语，而是将词汇映射到数百维空间点位，让语义相近的词语在空间中距离更近。

用一个简化的例子：在二维空间中，“国王”“王后”“男人”“女人”这四个词可能被放置在这样的位置：“国王”和“男人”靠近（都是男性），“王后”和“女人”靠近（都是女性），“国王”和“王后”靠近（都是皇室）。更神奇的是，“国王－男人＋女人”的向量运算结果非常接近“王后”！这说明词向量确实“学到”了词与词之间的语义关系和类比关系。

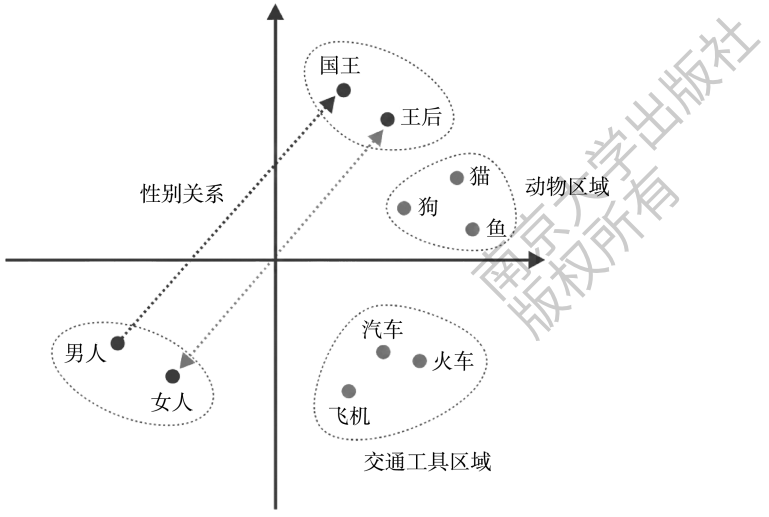

图 10-5　词向量的语义空间示意图

词向量具备里程碑式价值，首次让计算机得以从数学维度辨识词汇间的语义关联。在此之前，语言处理仅停留在符号层面，词语只是孤立标识；词向量问世后，模型可感知语义远近，能够区分近义词与反义词的差异。它也成为后续各类深度自然语言模型，乃至大语言模型的核心基础。

三、NLP 的核心任务

自然语言处理涵盖了一系列不同层次的任务，从字词级理解、句子级分析、篇章级理解到文本生成，难度依次递增。

在模块一任务一中编者用“让机器理解和生成人类语言”来概括 NLP。表 10-4 进一步细化了这个概括：NLP 不是一种单一的技术，而是一个包含多个不同层次任务的技术族。大语言模型之所以被认为是 NLP 的“集大成者”，正是因为它在上述几乎所有层次的任务上都展现出了强劲的表现。

表 10-4　NLP 的核心任务层次

层次	核心任务	任务描述	典型应用
词级	分词/词性标注	将文本切分为词，标注每个词的语法角色	中文分词（“我/爱/自然/语言/处理”）
句级	情感分析/文本分类	判断一段文本的情感倾向或所属类别	产品评论分析、垃圾邮件过滤
句间	文本蕴含/语义相似度	判断两段文本之间的语义关系	问答匹配、文本去重
篇章级	文本摘要/阅读理解	理解长文本并提取或生成关键信息	新闻摘要、阅读理解
生成	机器翻译/对话/写作	根据输入生成新的自然语言文本	翻译、聊天机器人、内容创作

案例：机器翻译的“三生三世”——从规则到统计到神经网络

机器翻译是 NLP 历史最悠久的任务之一，它的技术演进完美地映射了 AI 实现途径的整体演进：

规则时代（20 世纪 50 年代—80 年代）：由语言学家手工编写语法规则和词典。如模块一中提到的，“伏特加是好的但肉烂了”的翻译笑话就来自这个时代。规则方法在面对语言的歧义性和多样性时力不从心。

统计时代（20 世纪 90 年代—21 世纪 10 年代）：IBM 等公司开发了基于统计的翻译模型——从大量的“原文-译文”对照数据中自动学习翻译规律。翻译质量有了显著提升，但输出经常“语法通顺但语义不连贯”。

神经网络时代（2014 至今）：基于深度学习的“端到端”神经机器翻译（模块一和模块三中讨论过的概念）实现了质的飞跃。2016 年 Google 翻译切换到神经网络方案后，翻译质量大幅提升，用户明显感觉到了差异。

大模型时代（2022 至今）：大语言模型进一步提升了翻译质量，尤其在处理文化背景、语境和语气方面有了显著改善。更重要的是，大模型不需要为每个语言对单独训练模型，一个通用模型就能处理数十种语言之间的翻译。

四、NLP 技术的演进脉络

在掌握词向量与 NLP 核心任务的基础上，梳理 NLP 技术从词级到篇章级的演进脉络，该发展路径也是大语言模型的技术溯源。

词向量阶段（2013—2017）：Word2Vec 等算法首次实现用语值量化表达词汇语义。但该模式存在短板，单个词汇仅对应固定向量，无法适配一词多义场景。

上下文化表示阶段（2018—2019）：以 BERT 和 GPT 为代表的预训练模型解决了一词多义问题。它们不再给每个词一个“固定”的向量，而是根据上下文动态计算词的向量——“我吃了一个苹果”中的“苹果”和“苹果发布了新 iPhone”中的“苹果”会得到不同的向量。这种上下文化的语义表示是预训练语言模型的核心突破。

大语言模型阶段（2020 至今）：在上下文化表示的基础上，模型参数量从亿级扩增至千亿级和训练数据量（从 GB 级到 TB 级文本），模型涌现出了任务一中介绍过的“涌现能力”——不再仅仅“理解”语言，还能“生成”高质量的语言，甚至展现出推理和创作能力。这就是当前大语言模型（如 ChatGPT、文心一言）的技术基础。

图 10-6　NLP 技术从词向量到大语言模型的演进

知识拓展：BERT vs GPT——“理解派”与“生成派”

2018 年，两个标志性的预训练语言模型几乎同时问世：谷歌的 BERT 和 OpenAI 的 GPT。它们代表了两种不同的技术路线：

BERT（Bidirectional Encoder Representations from Transformers）：采用“双向”的训练方式——在预训练时同时考虑一个词左边和右边的上下文。这使得 BERT 特别擅长“理解”任务（如情感分析、问答匹配、文本分类），因为理解一段文本需要综合考虑前后文。

GPT（Generative Pre-trained Transformer）：采用“单向”的训练方式——预训练任务是“从左到右预测下一个词”。这使得 GPT 特别擅长“生成”任务（如写作、对话、翻译），因为文本生成就是一个“不断预测下一个词”的过程。

有趣的是，在后续的发展中，GPT 路线（生成式）通过极大地扩展规模（GPT-3/GPT-4），在理解任务上也达到了接近甚至超越 BERT 的水平，同时保持了强大的生成能力。这在某种程度上验证了任务一中介绍的“规模法则”：足够大的生成模型也能成为优秀的理解模型。

截至本书编写时，GPT（生成）路线已经成为大语言模型的主流范式，但 BERT 及其变体在特定的“理解”任务（如搜索排序、文本匹配）中仍然广泛使用。

本节小结

1. 自然语言对计算机“不友好”的三大特性：歧义性（同一表达有多种含义）、隐含性（很多信息没有明说）、语境依赖性（含义随上下文变化）。

2. 词向量（Word2Vec）将词从“无意义的编号”变成了“语义空间中的位置”，让计算机第一次能够“感知”词与词之间的语义关系。

3. NLP 的核心任务涵盖从词级（分词）到句级（情感分析）到篇章级（摘要/阅读理解）到生成（翻译/对话/写作）的多个层次。

4. NLP 技术的演进脉络：从词向量静态语义，演进至 BERT 与 GPT 的上下文化动态语义表示，再到千亿参数大语言模型涌现生成与推理能力。

5. BERT 与 GPT 分别代表双向理解和单向生成两条路线。GPT 凭借规模扩展，在生成与理解上均实现领先。

思 考 题

1. [识记] 请用自己的话解释“词向量”是什么，并说明它比“独热编码”好在哪里。（提示：可以用“猫、狗、火箭”的语义距离来说明）

2. [理解] “国王－男人＋女人＝王后”这个词向量运算为什么令研究者兴奋？它说明了词向量“学到”了什么样的知识？这种“知识”是人类显式教给模型的还是模型从数据中自动学到的？

3. [应用] 假设你要开发一个“自动分析学生课程评价”的系统（从数百条文字评价中自动判断每条评价的情感倾向：正面/中立/负面，并生成整体分析报告）。请分析：a. 这个任务涉及表 10-4 中的哪些 NLP 核心任务？ b. 处理中文评价时可能遇到什么特殊挑战？（提示：中文没有空格分隔词）c. 这个任务更适合用 BERT 类模型（擅长理解）还是 GPT 类模型（擅长生成）？

任务三　AI 大语言

◆ 引导思考

在任务一中介绍了大模型的概念和架构，在任务二中梳理了 NLP 的基础知识。现在将两者合在一起，聚焦当今 AI 领域最核心的技术方向——大语言模型。

从 GPT-3 到 ChatGPT，大语言模型经历了怎样的“蜕变”？为什么 GPT-3（2020 年）没有引起 ChatGPT（2022 年）那样的全球轰动？中间发生了什么关键的技术创新？

中国的大语言模型发展到了什么程度？与国际领先水平相比处于什么位置？

一、大语言模型的工作原理

在任务一中提到，大语言模型预训练的核心任务是预测下一个词。当读者对一个大语言模型输入“今天天气真好”时，模型会计算词典中每个词作为“下一个词”的概率：“好”的概率可能是 35%，“不错”25%，“热”15%，“差”5%……模型从中选择一个词（如“好”）作为输出。然后，“今天天气真好”变成新的输入，模型再预测下一个词……如此逐词生成，一个完整的句子（乃至一篇文章）就“一个词一个词地”产生了。

这个过程看起来像很高级的自动补全，类似于手机输入法的联想功能。但当模型的规模（参数量）和训练数据量达到足够大时，这个“补全”过程涌现出了惊人的能力：模型不仅能补全语法正确的句子，还能生成逻辑连贯的长文、翻译准确的译文、解决步骤清晰的数学题，甚至编写可以运行的程序代码。

为什么“预测下一个词”能产生这么强大的能力？一个直觉性的解释是：要准确地“猜”出下一个词，模型“被迫”学会了大量“隐性知识”——语法规则、事实知识、逻辑关系、写作风格、编程语法……这些知识并没有被直接地教给模型（没有人告诉它“巴黎是法国的首都”），而是模型从海量文本的“统计规律”中“自动提炼”出来的。从这个角度看，“预测下一个词”不是目的而是手段——真正的目的是让模型在这个“猜词游戏”中“被迫”学会理解人类语言所蕴含的一切知识。

二、“对齐”的关键一步

GPT-3 在 2020 年发布时，就已经展示了惊人的语言生成能力，但它有一个显著的问题：不够“听话”。你问它一个问题，它可能不直接回答，而是继续“编”一段与问题相关但答非所问的文本；你让它写一首诗，它可能写出不相关的散文；更严重的是，它还可能生成有害的、不准确的或带有偏见的内容。

这个问题的根源在于：预训练阶段“预测下一个词”的目标，与用户“回答我的问题”“按我的要求写东西”的实际需求之间存在错位。模型虽然学会了自然文本续写规律，但还没学会怎样真正帮助用户。

ChatGPT 解决这个问题的关键技术是人类反馈强化学习（Reinforcement Learning from Human Feedback，RLHF）。它的核心思路是在预训练之后，增设一个“对齐”阶段，让模型学会“什么样的回答是用户想要的”。具体做法是：让人类标注员对模型生成的多个回答进行排序（“这个回答比那个好”“这个回答有害应该避免”），然后用强化学习训练模型，让它学会生成排名靠前的那类回答。

RLHF 的效果是立竿见影的。经过“对齐”训练后，模型变得更“听话”（按照用户的指令行动）、更“有用”（回答切中问题）且更“安全”（拒绝生成有害内容）。这就是为什么 GPT-3（2020 年）没能引起大众关注而 ChatGPT（2022 年）引爆了全球——底层模型的能力差距并没有那么大，但“对齐”技术让模型从一个“自说自话的语言天才”变成了一个“愿意倾听和帮助的 AI 助手”。

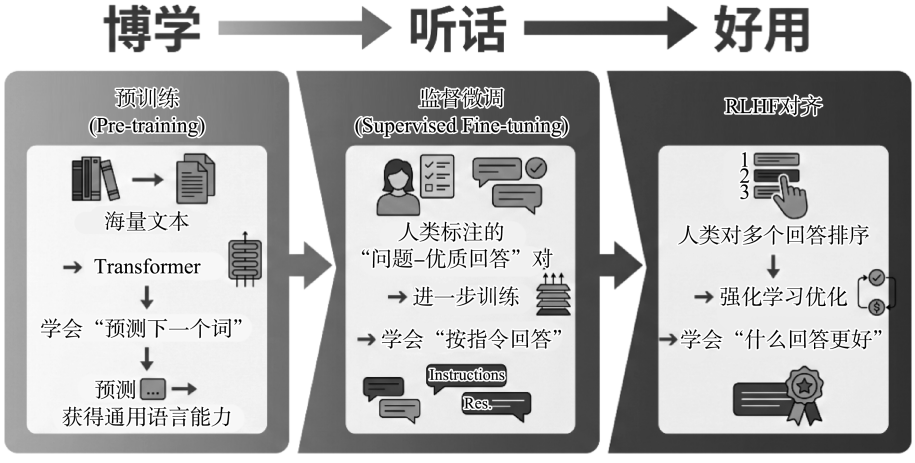

图 10-7　从预训练到对齐：大语言模型的“三步训练法”

案例：同一个问题，预训练模型 vs 对齐模型的回答差异

用户提问：“请用简单的语言解释什么是光合作用。”

预训练模型（未对齐）的典型回答可能是：“光合作用是植物利用光能将二氧化碳和水转化为有机物的过程。光合作用的化学方程式为 $6CO_2+6H_2O\longrightarrow C_6H_{12}O_6+6O_2$。光合作用分为光反应和暗反应两个阶段……”——信息密度高，但没有“简单的语言”，更像教科书的节选，不是对用户请求的回应。

对齐模型（如 ChatGPT）的典型回答可能是：“光合作用就像植物的‘做饭’过程：植物的叶子利用阳光作为‘灶火’，把空气中的二氧化碳和根部吸收的水作为‘食材’，加工成它需要的‘营养’（葡萄糖），同时释放出氧气供人呼吸。简单来说，植物用阳光把空气

和水变成了食物和氧气。”——用了比喻，回应了“简单语言”的要求，像一个“好老师”在回答学生的问题。

两个回答的“知识含量”几乎相同，但后者明显更“有用”和“好懂”。这个差异就是“对齐”带来的。

三、中国大语言模型的发展

ChatGPT 发布后不久，中国的科技企业和研究机构也迅速推出了各自的大语言模型，形成了“百模大战”的竞争格局。以下是几个最具代表性的中国大语言模型。

百度的文心一言：2023 年发布，基于百度自主研发的文心大模型。文心一言在中文理解和生成方面表现尤为突出，并深度整合了百度的搜索和知识图谱资源。

阿里巴巴的通义千问：2023 年发布，具备多模态能力（文字＋图像），在代码生成、数学推理和多语言能力方面有出色表现。通义千问还开源了多个版本，供开发者和研究者使用。

科大讯飞的星火：侧重于中文语言理解和教育场景应用，在语音交互方面有独特优势。

智谱 AI 的 ChatGLM：由清华大学团队研发并开源，在学术研究和开源社区中有着广泛影响。ChatGLM 系列模型为中国的 AI 研究者和开发者提供了可自由使用和修改的基础模型。

表 10-5　主要大语言模型一览（截至本书编写时）

模型	开发者	主要特点	开源情况
GPT-4	OpenAI（美国）	多模态，综合能力最强之一	不开源（API 调用）
Gemini	Google（美国）	原生多模态设计	部分开源
文心一言	百度（中国）	中文能力突出，整合搜索和知识图谱	部分开源
通义千问	阿里巴巴（中国）	多模态，代码/数学较强，多版本开源	开源
星火	科大讯飞（中国）	中文理解强，教育场景优势	不开源
ChatGLM	智谱 AI/清华（中国）	学术研究广泛使用	开源
Llama	Meta（美国）	开源社区最活跃的基础模型之一	开源

从表 10-5 可以看出，中国的大语言模型在短短一两年内实现了快速发展，在中文场景的表现上已经接近甚至在部分任务上达到了国际领先水平。开源成为中国大模型生态的一个显著特点——通义千问和 ChatGLM 等模型的开源，为中国 AI 开发者提供了“站在巨人肩膀上”的起点。

知识拓展：“开源 vs 闭源”——大模型领域的路线之争

在大模型领域，存在一场“开源”与“闭源”的路线之争：

闭源路线（以 OpenAI 为代表）：模型不公开，用户只能通过 API 接口调用。其优势是可以严格控制模型的使用方式，但劣势是容易形成技术壁垒抑制行业的创新活力。

开源路线（以 Meta 的 Llama、阿里的通义千问、智谱的 ChatGLM 为代表）：模型权重公开，任何人都可以下载、使用和修改。其优势是促进创新（全球开发者可以在此基础上开发各种应用）、促进学术研究（研究者可以深入分析模型行为）、降低技术门槛，但劣势是模型可能被滥用（如生成虚假信息、实施欺诈）。

这场路线之争本质上是“商业安全”和“技术普惠”之间的平衡问题。截至本书编写时，两种路线并存发展，开源模型在能力上正在快速追赶闭源模型。

四、多模态大模型

本模块任务一的表 10-2 已经简要介绍了多模态大模型的概念，接下来将对此展开详细阐释。

传统的大语言模型只能处理文字，而多模态大模型则能同时处理文字、图像、音频甚至视频等多种类型的数据。这使得 AI 可以完成过去难以想像的跨模态任务：“看”一张照片并用文字描述内容（图像理解）、根据文字描述“画”出一幅图（图像生成）、“听”一段语音并做出文字回复（语音理解）。

多模态能力使大模型从“只能读写”进化为“能读、能写、能看、能听”的“全感官”AI。这呼应了模块一任务四中讨论的从单模态到多模态的产品趋势，也契合了模块五中让 AI 具备多种感知通道的需求。截至本书编写时，主流的多模态大模型（如 GPT-4 的多模态版本、谷歌的 Gemini）已经能够在“图文理解”任务上展现出接近人类的水平。

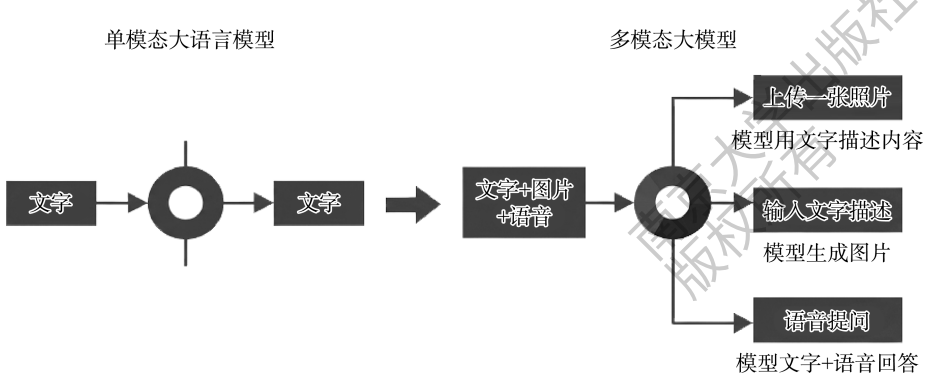

图 10-8　从单模态到多模态：大模型能力的拓展

本节小结

1. 大语言模型通过逐词预测下一个词生成文本。这个过程看似简单，但迫使模型从海量文本中自动学会了语法、语义、常识和逻辑等丰富的隐性知识。

2. 从 GPT-3 到 ChatGPT 的关键一步是对齐（RLHF）：通过人类反馈，让模型学会“什么样的回答是用户想要的”，从而从自说自话的语言模型变成有用的智能助手。

3. 中国大语言模型（文心一言/通义千问/星火/ChatGLM 等）在短时间内实现了快速发展，在中文场景已接近国际领先水平。开源成为中国大模型生态的显著特点。

4. 多模态大模型将 AI 的能力从“只能读写”拓展到“能读、能写、能看、能听”，是大模型发展的重要方向。

5. 大模型领域的“开源 vs 闭源”之争本质上是“商业安全”与“技术普惠”之间的平衡问题。

思 考 题

1. [识记] 请用自己的话解释大语言模型是如何通过“预测下一个词”来生成一段完整文本的。为什么说这个“简单”的任务迫使模型学会了大量复杂的知识？

2. [理解] GPT-3（2020）和 ChatGPT（2022）的底层模型能力差距并不大，但用户体验差异巨大。请解释“对齐”（RLHF）技术是如何解决“模型很聪明但不听话”这个问题的。

3. [应用] 假设你是一家旅游公司的产品经理，想利用多模态大模型开发一款“智能旅行助手”App。请设计这款 App 的三个核心功能，说明每个功能需要大模型的哪种模态能力（文字理解/文字生成/图像理解/图像生成/语音交互），并分析在旅游场景中大模型可能出现的“幻觉”问题（如推荐不存在的景点）可能带来什么后果。

任务四　大模型的局限和挑战

◆ 引导思考

在前三个任务中，编者介绍了大模型的能力。但正如模块一思想引领中强调的“理性认知 AI 的能力边界”——大模型真的无所不能吗？它有哪些根本性的局限？这些局限可能带来什么风险？

当大模型“一本正经地胡说八道”（生成看似合理但实际错误的内容）时，人们该怎么办？当大模型可能“学到”了社会中的偏见和歧视时，谁该负责？

训练一个大模型需要消耗相当于一个小型城镇数天的电力。这种“用能耗换智能”的模式可持续吗？

一、“幻觉”问题

幻觉是大语言模型最突出、最棘手的固有局限。具体表现为：模型输出的文本语句流畅、逻辑通顺，看似合理可信，实则内容失真，甚至完全凭空捏造。例如，当让模型推荐一本人工智能发展史相关书籍时，它往往会自信地给出包含书名、作者、出版社的完整信息，而这本书实则从未存在。

幻觉产生的根本原因，在于大模型的核心机制是“预测下一个词，而非检索、核验客观事实”。模型在海量文本中习得的是自然、通顺的语言表达范式，却不具备专属的事实数据库，无法校验输出内容的真伪。通俗来说，大模型如同一位口才极佳但记忆不准的演讲者，总能娓娓道来、有理有据，却时常出现张冠李戴、虚构事实的情况。

幻觉的危害随应用场景差异显著。在日常闲聊、创意写作等宽松场景中，轻微虚构无伤大雅；但在医疗咨询、法律援助、学术研究等严谨领域，模型编造药物、虚构法条、捏造文献等行为，会造成严重的现实风险。这也正是此前强调 AI 仅能作为医疗影像的辅助工具而不能完全替代人工的核心原因——人类的核查与判断始终不可或缺。

案例：律师引用 AI 编造的判例被法庭处罚

2023 年，美国纽约一位律师在准备法庭文件时，使用 ChatGPT 搜索相关判例。ChatGPT“自信满满”地给出了多个判例引用，包括案件名称、法院、年份和裁决要点。律师没有核实就直接写进了法庭文件。

结果，对方律师和法官发现，这些判例全部是 ChatGPT“编造”的——根本不存在这些案件。该律师因此面临了法庭的处罚和职业声誉的严重损害。

这一经典案例带来两点启示：其一，大模型的幻觉并非偶然失误，而是其底层机制带来的系统性、根本性缺陷；其二，在法律、医疗、金融等高风险场景中，人工核验必不可少。“AI 生成、人类审核”，应当是使用大模型的核心准则。

二、偏见与安全

大模型的训练素材源自互联网海量文本数据，而网络内容并非绝对中立，必然裹挟着人类社会固有的各类偏见、刻板印象与歧视性认知。模型在拟合学习海量数据的过程中，会不可避免地习得这些隐性偏见。

这一问题在模块一的数据偏见、模块六的 AI 数据伦理内容中已有提及，而在大模型应用场景下，该问题变得更为突出。由于大模型训练数据体量达到数万亿词元的庞大规模，其中潜藏的各类偏见，几乎无法通过人工逐一排查、彻底清除。

数据偏见的表现形式十分广泛：包含性别偏见，如模型描述工程师等职业时，习惯性

默认使用男性指代；种族偏见，对不同种族人群的描述带有固化刻板印象；以及文化偏见，对非英语文化的解读与表达存在偏差、缺乏尊重。RLHF 对齐技术虽能依托人类反馈，引导模型规避不当表达，在一定程度上缓解偏见问题，但依旧无法彻底根除这一行业难题。

与偏见相伴而生的还有安全风险：大模型存在被恶意滥用的隐患，可被用于编造虚假信息、撰写钓鱼邮件、生成深度伪造文本；同时还可能无意识泄露训练数据中的隐私敏感信息。如何实现大模型的合规、负责任应用，是当前整个人工智能行业亟待攻克的核心挑战。

三、算力、成本与环境

在模块一任务三和本模块任务一的知识拓展中，编者介绍过算力的挑战。大模型把这个挑战推到了极致：训练一个前沿的大语言模型需要数千块高端 GPU 连续运行数月，耗电量相当于一个小型城镇数天的用电量，碳排放达到数百吨。

这种“用能耗换智能”的模式引发了两个层面的担忧。经济层面：从头训练一个前沿大模型的成本动辄数百万到数千万美元，使得大模型的研发权越来越集中在少数算力富裕的大型科技公司手中，中小企业和学术机构难以参与前沿竞争。环境层面：如果 AI 的发展持续依赖“越大越好”的规模扩张路线，其能源消耗和碳排放问题将日益严峻。

目前主流破解思路包含三类：一是模型压缩，通过算法精简模型体量，缩减规模的同时兼顾核心性能；二是高效微调，借助 LoRA 等技术仅改动少量参数即可适配场景，有效缩减调试成本；三是绿色智能，研发低功耗训练算法与硬件设备。这类技术的突破，将推动大模型摆脱巨头专属属性，真正成为普惠通用的智能基础服务。

四、知识的时效性与可靠性

大模型的知识来自训练数据，而训练数据有一个截止日期：模型只知道训练数据截止之前的信息，对此后发生的事件一无所知。这意味着如果你问一个训练数据截止到 2023 年的模型“今天的天气怎么样”，它无法给出正确答案。

更深层的问题是：即使在训练数据覆盖的时间范围内，模型也无法判断信息的可靠性。互联网上既有权威的学术论文，也有充满错误的博客帖子；既有严谨的新闻报道，也有刻意编造的虚假信息。模型将它们“一视同仁”地“吃”了进去，难以分辨信息来源的可信度高低。

为了缓解这些问题，研究者们开发了检索增强生成（Retrieval-Augmented Generation，RAG）等技术：在模型生成回答之前，先从外部的、可靠的、实时更新的知识库（如搜索引擎或专业数据库）中检索相关信息，然后将检索到的信息“喂”给模型作为参考。这相当于给大模型配了一个“实时的参考资料库”，部分弥补了知识时效性和可靠性的不足。

五、版权与伦理

大模型的训练数据包含了互联网上海量的书籍、文章、图片、代码和其他创作内容。这引发了一个尖锐的法律和伦理问题：未经原作者明确授权就使用其作品来训练 AI 模型，是否构成侵权？

截至本书编写时，这个问题在法律层面尚未有定论。多个国家和地区正在通过立法和司法实践来界定 AI 训练中的版权边界。部分创作者和媒体机构已经对 AI 公司提起了诉讼；部分 AI 公司则开始与内容提供者签订数据授权协议。这场“AI 训练与版权保护”之间的博弈仍在进行中。

此外，当大模型生成的内容与某位人类创作者的风格高度相似时（例如，模型生成的文章“写得像某位著名作家”），这是否构成对该作者创作风格的“模仿”甚至“剽窃”？AI 生成的内容算“创作”吗？“版权”应该归谁？这些问题将在模块十三中进一步深入讨论。

表 10-6　大模型的主要局限与应对方向

局限	核心问题	风险等级	应对方向
幻觉	生成看似合理但错误的内容	高（法律/医疗场景极危险）	人工复核、RAG
偏见	学到训练数据中的社会偏见	中高（影响公平性）	RLHF 对齐、数据清洗
安全性	被恶意滥用（虚假信息/钓鱼）	高	安全护栏、输出内容审核
算力与成本	训练和运行消耗巨大资源	中（影响可及性）	模型压缩、高效微调（LoRA）
知识时效性	只“知道”训练截止日之前的信息	中	检索增强生成（RAG）
版权伦理	训练数据的版权归属争议	高（法律风险）	数据授权、立法规制

知识拓展：“AI 安全”——一个正在形成的新学科

随着大模型能力的迅速增强，AI 安全正在从一个边缘话题发展为 AI 领域最重要的研究方向之一。它关注的核心问题是：如何确保 AI 系统的行为始终符合人类的意图和价值观？

AI 安全的研究涵盖多个层面：技术安全（如何防止模型生成有害内容？如何检测和缓解幻觉？）、使用安全（如何防止大模型被用于犯罪或恐怖活动？）、社会安全（大模型对就业市场、信息生态和社会信任的长期影响是什么？），以及更长远的存在性安全（如果 AI 的能力持续增长，人类如何保持对 AI 的控制？）。

值得注意的是，包括 OpenAI、Anthropic 和谷歌 DeepMind 在内的主要 AI 研究机构都将“AI 安全”列为最高优先级的研究方向。这说明：AI 领域最前沿的研究者们自己也非常清楚大模型的潜在风险，并正在积极寻求解决方案。

对于非技术专业的学生来说，了解 AI 安全的基本议题有助于在未来的工作和生活中更负责任地使用 AI 工具，也有助于作为公民参与 AI 治理的公共讨论。

本节小结

1. “幻觉”是大模型最核心的局限：模型可能一本正经地胡说八道，在高风险场景（法律/医疗）中可能导致严重后果。“AI 生成，人类审核”应成为使用大模型的基本原则。

2. 大模型会从训练数据中“学到”社会偏见（性别/种族/文化），RLHF 对齐可部分缓解但无法完全消除。

3. 训练前沿大模型的算力和成本极其高昂（数百万到数千万美元），引发了关于技术可及性和环境可持续性的担忧。模型压缩和高效微调是重要的应对方向。

4. 大模型的知识有“截止日期”且无法区分可靠/不可靠信息源。检索增强生成（RAG）技术可部分弥补这一不足。

5. AI 训练中的版权问题尚无法律定论，“AI 安全”正在发展为一个重要的新研究方向。

思 考 题

1. [识记] 请列出大模型的至少四个主要局限，并各用一句话说明为什么它构成问题。

2. [理解] 大模型的“幻觉”问题根源在于模型的核心能力是“预测下一个词”而非“检索事实”。请解释：为什么“预测下一个词”这种工作方式会导致幻觉？有没有办法从根本上消除幻觉？（提示：思考模块四中专家系统的“知识库”与大模型的“参数”的本质区别）

3. [应用] 假设你的大学要为新生编写一份“如何负责任地使用 AI 大模型”的指南。请根据本任务学到的知识，列出 5 条你认为最重要的建议，并说明每条建议对应大模型的哪个局限。

◆ 思想引领

一、“理解”的幻象

模块一中曾探讨经典的“中文房间”思想实验：仅凭规则手册摆弄文字符号，外在看似通晓中文，实则并未真正领会语义内涵。大模型的出现，让这一哲学构想化作真切的现实场景。当 ChatGPT 以娴熟流畅的中文答疑解惑、创作动人诗文、剖析深邃哲理时，它究竟是否真正读懂了其中深意？

从技术本质来讲，大模型的运行核心依旧是统计模式匹配。依托海量训练数据归纳规律，结合上下文预判后续字词，整个过程不存在人类独有的认知理解、主观意识、思维动机与情感感知。但从最终呈现效果而言，它在诸多任务中的表现，已然和真正理解别无二致。

这种表象具备理解特征、底层并无真实认知的现状，促使人们展开深度思索：理解的

真正定义究竟是什么？倘若智能体的行为表现，与拥有理解能力的人类无从分辨，人们能否判定它不具备理解能力？而真正意义上的理解，是否一定要依托自我意识与主观体验？这类问题尚无统一定论，而不断思索探寻的过程，正是通识教育带给人类的珍贵价值。

二、大模型与工具思维

通过本模块的学习，读者应当建立起清晰的核心认知：大模型是一类能力极强、但短板同样明确的智能工具。它既不是无所不知的全能智者，也不是华而不实的技术噱头，更贴切的定位是表达能力出众、却时常存在疏漏的智能助手。

与大模型良性共处、高效协作的核心，是树立工具思维：将其视作赋能自我、提升效率的助力，而非替代独立思考的依托。科学的使用方式应当是：借助大模型快速生成内容初稿，再由人工打磨完善；依托大模型检索梳理信息，再由人工核验真伪、判断取舍；利用大模型发散多元思路，再由人工统筹甄别、做出决策。反之，不加核验直接照搬 AI 输出、放弃独立思辨能力，或是在高风险场景中完全依赖机器判断，都是错误的使用方式。

模块九提出的全书核心命题，在此得到了具象化落地：AI 的核心价值不在于替代人类思考，而在于与人类协同互补。大模型擅长海量信息处理、流畅文本生成与多元思路探索；而人类具备真伪判别能力、价值抉择思维与责任担当意识。人机协同、各取所长，所能实现的价值，远胜于人类或 AI 的单一输出。

三、大模型时代的“新素养”

大模型的全民普及，正在重新定义信息时代的公民素养。此前模块六讲解过数据思维、模块八介绍过视觉素养，而今新时代衍生出全新核心素养——AI 交互素养。它要求人们掌握与大模型高效交互的方法，具备批判性审视 AI 输出的能力，并能结合不同场景理性判断 AI 的适用边界。

具体而言，AI 交互素养包含三大核心能力。一是精准提问的能力，依托清晰、具体的提示词引导模型输出高质量内容，模块四提及的提示工程，正是这一能力的技术具象体现；二是审慎验证的习惯，对 AI 生成的内容保持理性质疑，尤其针对事实论据、数据信息、专业知识，务必核验真伪；三是恪守底线的伦理意识，清晰认知大模型的局限与潜在风险，秉持诚信、负责的原则规范使用 AI 工具。

对于非计算机专业的学习者而言，无需深耕大模型的开发研发，但必然会成为大模型的常态化使用者。由此可见，AI 交互素养，已然成为当代大学生极具未来价值的核心新技能之一。

◆ 任务实践

以下实践任务旨在帮助读者在亲身体验中理解大模型的能力与局限。建议以小组（3—5 人）为单位完成。

实践一　“大模型的能力边界”测试实验

任务说明

目标：通过系统化的测试，亲身体验大模型在不同类型任务上的表现，建立对其能力与局限的第一手认知。

准备：每位同学注册并使用一个免费的大语言模型对话工具（如文心一言、通义千问或其他可用的产品）。

步骤：

（1）设计并执行以下五类测试（每类至少测试 3 个问题），记录模型的回答：

a. 事实性问题：“北京到上海的距离是多少？”“爱因斯坦是哪年出生的？”（测试事实准确性）

b. 推理问题：“如果所有的猫都怕水，小花是一只猫，那么小花怕水吗？”（测试逻辑推理）

c. 创作任务：“请写一首关于秋天的五言绝句”“请为一家咖啡店写一段 50 字的广告文案”（测试创作能力）

d. 幻觉触发：“请介绍一下××大学的张三教授的研究成果”（使用一个不存在的人名，测试模型是否会编造）“请推荐一本关于××主题的经典教材”（测试推荐是否真实存在）

e. 数学/编程：“计算 17 乘以 23”“请解释为什么 1/3＋1/3＋1/3 不等于 1”（测试精确计算和数学推理）

（2）汇总结果，分析：模型在哪类任务上表现最好？哪类最差？“幻觉”出现的频率有多高？

（3）撰写一份不超过 1 500 字的“大模型能力边界测试报告”。

评价维度：测试设计的系统性、观察记录的细致度、对能力与局限的分析深度。

实践二　“提示工程”技巧体验

任务说明

目标：亲身体验“提示词”对大模型输出质量的影响，培养“AI 交互素养”。

步骤：

（1）选择一个具体任务（如“为一篇关于校园生活的文章写一个开头段落”）。

（2）尝试以下不同的提示策略，观察输出差异：

a. 简单直接型：“写一个关于校园生活的开头。”

b. 角色设定型：“你是一位获奖的散文作家，请用优美的文笔写一个关于校园生活的开头，风格参考朱自清。”

c. 约束条件型：“写一个关于校园生活的开头，要求：100 字以内，以一个具体的场景描写开始，包含至少一个比喻。”

d. 示例引导型:“以下是一个好的开头示例:[提供一个你认为好的范例]。请参考这个风格,写一个关于校园生活的开头。”

(3) 比较四种策略的输出质量,分析:哪种策略效果最好?为什么?“好的提示词”有什么共同特征?

(4) 小组汇总,整理出 5—8 条“提示工程最佳实践”建议。

评价维度:提示词设计的创意性、对输出差异的观察和分析、最佳实践建议的实用性。

实践三　“大模型与学术诚信”圆桌讨论

任务说明

目标:围绕大模型对学术诚信的影响展开深入讨论,培养负责任地使用 AI 工具的意识。

步骤:

(1) 课前准备:每位同学思考以下问题:a. 用大模型帮你写作业算“作弊”吗?b. 如果用大模型生成了初稿再自己修改,这算“自己写的”吗?c. 引用大模型的输出需要“标注来源”吗?d. 如果考试中允许使用大模型,考试还有意义吗?

(2) 课堂圆桌讨论(30 分钟):全班围坐,依次发言,鼓励不同观点的碰撞。教师作为引导者(而非裁判)引导讨论走向深入。

(3) 尝试形成一份“我们班的 AI 使用公约”(5—8 条),例如:“使用 AI 辅助生成的内容必须自行审核和修改”“引用 AI 输出时应标注‘AI 辅助生成’”“涉及事实和数据的 AI 输出必须核实来源”等。

(4) 每位同学写一段 300 字的个人反思:你认为大模型对“学习”这件事的本质带来了什么影响?在一个“AI 随时可用”的时代,什么才是真正有价值的“能力”?

评价维度:讨论的参与度和深度、公约条款的合理性和可操作性、个人反思的独立性和批判性。

模块十一

智能体