

# AAAR: AI 加速学术研究的方法论与经验

Zhang Yu

2026-01-20

# Table of contents

Welcome	5
Buy Today! . . . . .	5
快速导览 . . . . .	5
重点提示 . . . . .	5
关键原则 (有序列表) . . . . .	5
工具与流程 (无序列表) . . . . .	5
表格示例 . . . . .	5
图片示例 . . . . .	6
代码块示例 . . . . .	6
引用与引用框 . . . . .	7
超链接与行内代码 . . . . .	7
公式示例 . . . . .	7
任务列表 . . . . .	7
注脚与脚注引用 . . . . .	7
引用文献与参考文献 . . . . .	7
折叠块 . . . . .	8
分隔线 . . . . .	8
公式编号与引用 . . . . .	8
图注与交叉引用 . . . . .	8
代码与数据 . . . . .	8
1 序章   为什么写这本书	10
1.1 一、从一次失败的文献综述说起 . . . . .	10
1.2 二、为什么现有的“AI 科研指南”让我不满意 . . . . .	10
1.3 三、这本书想要提供什么 . . . . .	11
1.3.1 3.1 一个可持续的思维框架 . . . . .	11
1.3.2 3.2 一套可复现的工作流 . . . . .	11
1.3.3 3.3 必要的批判性 . . . . .	11
1.4 四、这本书写给谁 . . . . .	12
1.4.1 4.1 不同阶段的读者 . . . . .	12
1.4.2 4.2 不同学科的读者 . . . . .	12
1.4.3 4.3 不同技术背景的读者 . . . . .	12
1.5 五、这本书不是什么 . . . . .	12
1.6 六、本书的结构与阅读建议 . . . . .	13
1.6.1 6.1 整体结构 . . . . .	13
1.6.2 6.2 阅读路径建议 . . . . .	13
1.6.3 6.3 每章的结构 . . . . .	13
1.7 七、一些重要的声明 . . . . .	14
1.7.1 7.1 关于时效性 . . . . .	14
1.7.2 7.2 关于局限性 . . . . .	14
1.7.3 7.3 关于 AI 辅助写作 . . . . .	14
1.7.4 7.4 关于争议性观点 . . . . .	14
1.8 八、为什么选择开源 . . . . .	15
1.8.1 8.1 可复现与可验证 . . . . .	15
1.8.2 8.2 持续更新 . . . . .	15
1.8.3 8.3 社区协作 . . . . .	15
1.8.4 8.4 价值观的表达 . . . . .	15

1.9	九、我们正处于什么样的时刻 . . . . .	15
1.9.1	9.1 炒作周期的高峰与低谷 . . . . .	15
1.9.2	9.2 学术界的特殊处境 . . . . .	16
1.9.3	9.3 一场正在发生的范式转移 . . . . .	16
1.9.4	9.4 个人选择的重要性 . . . . .	16
1.10	十、致谢 . . . . .	16
1.11	十一、写在最后 . . . . .	17
<b>2</b>	<b>第 1 章 AI 能做什么，不能做什么</b>	<b>18</b>
<b>3</b>	<b>第 2 章 幻觉、偏差、泄露：三类核心风险</b>	<b>19</b>
<b>4</b>	<b>第 3 章 提示词不是方法论</b>	<b>20</b>
4.1	一、一场关于提示词的迷信 . . . . .	20
4.2	二、提示词能做什么，不能做什么 . . . . .	20
4.2.1	2.1 提示词确实有用 . . . . .	20
4.2.2	2.2 提示词不能做什么 . . . . .	20
4.2.3	2.3 核心区分：流程辅助 vs. 判断替代 . . . . .	21
4.3	三、为什么提示词被过度神化 . . . . .	21
4.3.1	3.1 简单问题的诱惑 . . . . .	21
4.3.2	3.2 商业利益的驱动 . . . . .	21
4.3.3	3.3 对 AI 能力的误解 . . . . .	22
4.3.4	3.4 可观察性偏差 . . . . .	22
4.4	四、从“提问技巧”走向“研究系统” . . . . .	22
4.4.1	4.1 什么是“研究系统” . . . . .	22
4.4.2	4.2 系统的核心特征 . . . . .	23
4.4.3	4.3 AI 在系统中的位置 . . . . .	23
4.5	五、建立自己的工作流原则 . . . . .	23
4.5.1	5.1 任务拆解原则 . . . . .	23
4.5.2	5.2 层级输出原则 . . . . .	24
4.5.3	5.3 验证机制原则 . . . . .	24
4.5.4	5.4 版本记录原则 . . . . .	24
4.5.5	5.5 边界意识原则 . . . . .	25
4.6	六、一个完整的例子 . . . . .	25
4.6.1	6.1 没有系统的做法 . . . . .	25
4.6.2	6.2 有系统的做法 . . . . .	25
4.6.3	6.3 两种做法的对比 . . . . .	26
4.7	七、写在最后 . . . . .	26
<b>5</b>	<b>第 4 章 信息压缩与文献整理</b>	<b>28</b>
5.1	引言：从信息过载到结构化理解 . . . . .	28
5.2	一、文献筛选与主题聚类 . . . . .	28
5.2.1	1.1 文献筛选的两阶段模型 . . . . .	28
5.2.2	1.2 设计有效的筛选指令 . . . . .	28
5.2.3	1.3 主题聚类的策略 . . . . .	29
5.2.4	1.4 一个实践案例 . . . . .	29
5.3	二、长上下文模型的正确用法 . . . . .	30
5.3.1	2.1 长上下文能力的本质与局限 . . . . .	30
5.3.2	2.2 分层处理策略 . . . . .	30
5.3.3	2.3 输入质量的重要性 . . . . .	30
5.3.4	2.4 输出设计与质量控制 . . . . .	30
5.3.5	2.5 一个错误使用的案例 . . . . .	31
5.4	三、可信综述生成：引用与验证 . . . . .	31
5.4.1	3.1 为什么综述生成是高风险任务 . . . . .	31
5.4.2	3.2 AI 在综述生成中的适当角色 . . . . .	31

5.4.3 3.3 引用验证的系统方法 . . . . .	32
5.4.4 3.4 建立可追溯的证据链 . . . . .	32
5.4.5 3.5 AI 辅助综述的披露与伦理 . . . . .	32
5.5 四、从工具到能力：信息素养的新维度 . . . . .	32
5.5.1 4.1 信息压缩作为研究者的核心能力 . . . . .	32
5.5.2 4.2 人机协作的最优配置 . . . . .	33
5.5.3 4.3 未来研究者的信息素养 . . . . .	33
5.6 结语 . . . . .	33
<b>6 第 5 章 结构化表达与写作</b>	<b>34</b>
<b>7 第 6 章 代码与计算能力</b>	<b>35</b>
<b>8 第 7 章 知识管理与协作</b>	<b>36</b>
<b>9 第 8 章 选题与研究问题</b>	<b>37</b>
<b>10 第 9 章 文献与理论构建</b>	<b>38</b>
<b>11 第 10 章 数据获取与构造</b>	<b>39</b>
<b>12 第 11 章 数据清洗与分析</b>	<b>40</b>
<b>13 第 12 章 写作、投稿与传播</b>	<b>41</b>
<b>14 第 13 章 RA Level：工具层</b>	<b>42</b>
<b>15 第 14 章 Supervisor Level：认知协作层</b>	<b>43</b>
<b>16 第 15 章 Domain Expert Level：推理与建模层</b>	<b>44</b>
<b>17 第 16 章 Agent Level：AI 作为行动者</b>	<b>45</b>
<b>18 第 17 章 Governance Level：制度与治理层</b>	<b>46</b>
<b>19 终章   AI 时代研究者的新能力结构</b>	<b>47</b>
<b>20 附录（可选）</b>	<b>48</b>

# Welcome

这是一个用于展示排版和视觉效果的封面页。我在这里补充了多种 Markdown 元素，方便你快速判断整体的排版质感与细节风格。

## Buy Today!

AAAR: AI 加速学术研究的方法论与经验

Open access • Print version coming soon.

[Buy from Amazon](#) [Buy from Publisher](#) [Download PDF](#)

## 快速导览

本书面向从本科生到教授的读者。你可以按“能力模块”阅读，也可以按“科研工作流”顺序阅读。

## 重点提示

AI 擅长生成，但不擅长负责。使用 AI 的关键在于验证与责任分配。

## 关键原则（有序列表）

1. 先定义问题，再让 AI 参与。
2. 先验证证据，再相信输出。
3. 先记录流程，再扩展规模。

## 工具与流程（无序列表）

- 文献：综述不是堆摘要，而是证据链。
- 数据：清洗的透明性决定可信度。
- 写作：表达是结果，逻辑是核心。

## 表格示例

模块	典型任务	主要风险
信息压缩	文献综述	幻觉引用
数据处理	清洗/结构化	规则不透明
写作表达	初稿/润色	逻辑空洞

## 图片示例



Figure 1: 封面示例

## 代码块示例

你是我的研究助理。请将下面的研究问题拆解成可检验的子问题，  
并给出潜在的数据来源与方法路径。输出为列表。

## 引用与引用框

“AI 擅长生成，但不擅长负责。”  
——写作时请把证据链放在第一位。

### **i 提示**

把 AI 当作“加速器”，不要当作“裁判”。

### **⚠ 警告**

不要把未发表的数据或敏感信息直接交给模型。

## 超链接与行内代码

- 官网链接：<https://yuzhangsjtu.github.io/AAAR/>
- 仓库链接：<https://github.com/yuzhangsjtu/AAAR>
- 行内代码示例：使用 `quarto render` 重新生成站点

## 公式示例

这是一条行内公式： $E[X] = \sum_i x_i p_i$ ，以及一个块级公式：

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N (Y_i^1 - Y_i^0)$$

## 任务列表

- 完成大纲
- 生成示例页面
- 补充案例
- 进一步润色正文

## 注脚与脚注引用

这是一个脚注示例。<sup>1</sup>

## 引用文献与参考文献

这是一条文献引用示例 (Knuth 1984)，引用会自动生成参考文献列表。

---

<sup>1</sup>这是脚注内容，用于测试排版与字号。

## 折叠块

### 折叠内容

这里是折叠块的内容，用来展示 Anthropic 风格下的折叠区域样式。

## 分隔线

---

## 公式编号与引用

带编号的公式如下：

$$\tau = E[Y(1) - Y(0)] \tag{0.1}$$

在正文中引用该公式：见公式 Equation 0.1。

## 图注与交叉引用

在正文中引用该图：见图 Figure 2。

## 代码与数据

本书所有内容均开源：

- 代码仓库：<https://github.com/yuzhangsjtu/AAAR>
- 在线阅读：<https://yuzhangsjtu.github.io/AAAR/>



Figure 2: 示意图：AI 加速研究的简单流程

# 1 序章 | 为什么写这本书

## 1.1 一、从一次失败的文献综述说起

2023年初的某个深夜，我坐在办公室里盯着屏幕，面前是一份用ChatGPT生成的文献综述初稿。那时候，GPT-4刚刚发布不久，整个学术圈都在讨论“AI会不会取代研究者”。我也不能免俗，决定亲自试一试。

那篇综述的主题是关于社会网络对信息传播的影响——一个我研究了好几年的领域。我给模型提供了详细的背景、明确的问题、甚至列出了几篇我认为重要的文献，然后让它“帮我写一篇文献综述”。

结果出乎意料地“好”。结构清晰，语言流畅，甚至引用格式都很规范。如果只看表面，这完全可以当作一份合格的课程作业。但当我开始逐条核对引用时，问题出现了：十二条引用中，有四条是完全捏造的——作者名对不上，期刊名不存在，DOI链接指向404。更糟糕的是，另外三条虽然文献确实存在，但模型对它们的概括与原文观点完全相反。

那一刻我意识到，这不仅仅是“工具有bug”的问题。模型并没有“理解”什么是文献综述，它只是在模仿文献综述的形式——用流畅的语言把看起来相关的东西串起来。它不知道证据链是什么，不知道学术引用的意义在于可追溯和可验证，更不知道一条错误的引用可能导致整个论证链条的崩塌。

这次经历让我开始思考一个更深层的问题：AI究竟能在学术研究中扮演什么角色？它显然可以提高效率，但效率的提高如果建立在可靠性的下降之上，那这种“加速”的代价是什么？

## 1.2 二、为什么现有的“AI科研指南”让我不满意

在那次失败之后，我开始系统地阅读市面上关于“如何在学术研究中使用AI”的书籍、文章和教程。令我失望的是，绝大多数内容都存在以下几个问题。

**第一，工具堆砌，缺乏方法论。**很多指南的逻辑是：这里有一个AI工具，它能做X，所以你应该用它做X。这种“功能清单”式的写作完全忽略了一个核心问题——为什么你需要做X？X在你的研究流程中扮演什么角色？用AI做X和用传统方法做X有什么本质区别？没有方法论的工具介绍，最多只能培养“熟练的操作员”，而不是“有判断力的研究者”。

**第二，回避风险，夸大功用。**或许是出于推广的目的，很多内容只讲AI能做什么，不讲它不能做什么；只讲成功案例，不讲失败教训。幻觉(hallucination)被轻描淡写地提一句，然后就被“所以记得核实”这种空洞的建议带过。但问题是，核实需要成本，而且很多时候你根本不知道哪些地方需要核实。如果一个工具需要你“处处核实”才能安全使用，那它到底是帮你节省时间，还是把风险转嫁给你？

**第三，混淆“更快”与“更好”。**我见过太多的句式：“以前需要三天的工作，现在只需要三小时。”这种表述隐含了一个危险的假设——原来三天的工作和现在三小时的工作是等价的。但真的等价吗？三天的文献阅读和三小时的模型总结，产出的是同样的理解吗？三天的数据清洗和三小时的自动化处理，留下的是同样的审计痕迹吗？效率的提升往往伴随着某些东西的损失，但这些损失很少被讨论。

**第四，对“责任”的回避。**当AI参与学术生产，责任如何分配？如果模型生成的内容包含错误，责任在使用者还是开发者？如果审稿人无法分辨AI生成的文本，学术诚信的边界在哪里？这些问题极少被正面回应。很多指南的态度是“工具是中性的，看你怎么用”——这当然是正确的废话，但它回避了真正的难题：在实践中，我们如何具体地、可操作地分配责任？

**第五，缺乏批判性反思。**AI的广泛使用可能对学术生态产生什么长期影响？如果所有人都用同样的模型来“加速”研究，学术多样性会不会下降？如果AI让写作变得更容易，会不会有更多的论文但更少的发现？如果模型的训练数据主要来自英文文献，非英语学术传统会不会被进一步边缘化？这些问题几乎没有讨论。

正是这些不满意，推动我写这本书。我想提供一种不同的视角——不是“AI能帮你做什么”，而是“你应该如何与AI协作，同时保持研究的严谨性”。

## 1.3 三、这本书想要提供什么

基于上述反思，我给这本书设定了三个核心目标。

### 1.3.1 3.1 一个可持续的思维框架

第一个目标是提供一个可持续的思维框架。所谓“可持续”，是指这个框架不会因为某个具体工具的更新而过时。GPT-5 出来了，这个框架依然适用；换一个模型，框架依然成立。

这个框架的核心是三个问题：

1. AI 在这个任务中的能力边界在哪里？它擅长什么？不擅长什么？会犯什么类型的错误？
2. 这个任务的质量标准是什么？我们用什么指标判断任务完成得好不好？这些指标中哪些是 AI 容易满足的，哪些是困难的？
3. 人机协作的最优分工是什么？哪些环节交给 AI 更高效？哪些环节必须由人类来做？交接点在哪里？如何确保交接不丢失关键信息？

这三个问题不是一次性回答的，而是在每一个具体任务中反复追问的。我希望读者在读完这本书之后，能够在面对任何新工具、新场景时，自动地问出这三个问题，并且有能力给出自己的回答。

### 1.3.2 3.2 一套可复现的工作流

第二个目标是提供一套可复现的工作流。“可复现”是学术研究的基石之一，但在 AI 辅助的研究中，可复现性面临新的挑战。

传统研究的可复现性主要依赖于方法描述的精确性：你告诉别人你用了什么数据、什么模型、什么参数，别人就能重复你的分析。但当 AI 参与之后，问题变得复杂：

- 模型的版本会更新，同样的提示词在不同版本下可能产出完全不同的结果。
- 模型的输出有随机性，即使版本相同、提示词相同，两次运行的结果也可能不一样。
- 很多 AI 工具是黑箱，你无法精确描述它内部做了什么。

面对这些挑战，我在本书中提出的工作流强调三个原则：

1. **记录一切。**不仅记录结果，还记录过程：你用了什么提示词？模型返回了什么？你做了哪些筛选和修改？为什么做这些修改？
2. **保留原始数据。**AI 处理过的数据应该与原始数据分开存放，确保任何时候都能回溯到 AI 介入之前的状态。
3. **人工复核关键节点。**识别工作流中的“高风险节点”——一旦出错，代价很高的环节——并在这些节点强制进行人工复核。

这套工作流会贯穿全书，在每一章具体任务的讨论中都会体现。

### 1.3.3 3.3 必要的批判性

第三个目标是保持必要的批判性。这不是说我要“反对 AI”——恰恰相反，我是 AI 的重度用户，我的日常研究工作中大量使用各种 AI 工具。但正因为我是用户，我更清楚它的问题在哪里。

批判性体现在几个层面：

**对工具本身的批判。**大语言模型不是魔法，它有非常具体的能力边界和失败模式。理解这些边界和模式，是安全使用的前提。

**对使用方式的批判。**很多“最佳实践”其实是未经检验的假设。比如“提示词越详细越好”——真的吗？在什么条件下这个结论成立？有没有反例？

**对系统效应的批判。**当一项技术被广泛采用，它会改变整个系统的激励结构。AI 让写作更容易，但这会导致更多好研究，还是更多坏研究？投稿数量上升，审稿负担加重，审稿质量会不会下降？这些系统层面的问题，虽然个体研究者难以改变，但至少应该意识到。

## 1.4 四、这本书写给谁

在动笔之前，我认真思考过目标读者是谁。最终的定位是：所有在学术工作中与 AI 打交道的人。这个范围听起来很宽，但实际上 是经过取舍的。

### 1.4.1 4.1 不同阶段的读者

**本科生和硕士研究生**可能是最直接的受益者。你们正处于学术训练的关键阶段，既要学习传统的研究方法，又要适应新工具的冲击。这本书希望帮你建立一个平衡的视角：AI 是有用的工具，但学术能力的核心——提出好问题、设计好研究、做出好判断——不会被工具取代。

**博士研究生和青年研究者**面临的挑战更具体：你们需要在有限的时间内产出高质量的研究，同时还要应对发表压力、资金压力、职业压力。AI 看起来像是缓解压力的捷径，但如果使用不当，它可能成为新的陷阱——你可能产出更多，但质量下降；你可能写得更快，但想得更浅。这本书希望帮你找到真正能提高研究质量的使用方式，而不是只追求表面的效率。

**成熟的研究者和教授**可能对 AI 工具本身已经有了自己的判断，但你们面临另一个问题：如何指导学生？如何制定团队的使用规范？如何评估 AI 辅助产出的研究成果？这本书的“五层次框架”和“团队协作”章节可能对你们特别有用。

**研究团队的管理者**——无论是实验室主任、项目负责人还是研究机构的领导——需要从更宏观的层面考虑 AI 的影响：如何制定政策？如何分配责任？如何平衡效率与风险？本书后半部分的讨论会涉及这些议题。

### 1.4.2 4.2 不同学科的读者

虽然我的主场是社会科学，但本书并不局限于社科读者。

**社会科学研究者**会发现这本书最贴近你的工作场景。从选题到文献综述，从数据收集到统计分析，从写作到投稿，每个环节都有具体的讨论。社科研究的特殊挑战——比如概念的模糊性、理论的多元性、数据的复杂性——也会被专门讨论。

**人文学科研究者**可能会觉得某些章节（比如统计分析）与你的工作距离较远，但其他章节——比如文献整理、概念辨析、写作表达——应该同样适用。AI 在人文研究中的角色可能与社科有所不同，但核心问题是相似的：如何利用工具而不被工具绑架？如何保持批判性而不是盲目接受？

**自然科学和工程学科的研究者**可能会发现，虽然具体的例子和场景与你的领域有差异，但底层的方法论讨论是通用的。特别是关于可复现性、数据管理、代码规范的讨论，对任何经验研究都有参考价值。

### 1.4.3 4.3 不同技术背景的读者

**对 AI 技术有深入了解的读者**可能会觉得某些技术解释过于基础。我的建议是跳过那些你已经熟悉的部分，直接进入方法论讨论。这本书的价值不在于教你“如何使用 ChatGPT”，而在于探讨“在学术研究中应该如何使用 AI”。

**对 AI 技术了解有限的读者**不用担心。本书不假设任何编程知识或技术背景。所有技术概念都会用通俗的语言解释，所有操作都会给出具体的步骤。如果你能用 Word 写论文、用浏览器搜文献，你就能跟上本书的节奏。

## 1.5 五、这本书不是什么

在说明这本书是什么的同时，我想明确它不是什么。

**这不是工具手册。**我不会逐一介绍市面上的 AI 工具，不会比较哪个工具“更好用”，也不会提供“N 个提示词模板让你效率翻倍”。工具更新太快，今天的教程明天就可能过时。我更关心的是那些不会过时的东西：思维方式、方法论、判断标准。

**这不是提示词大全。**提示词（prompt）确实很重要，但它被过度神秘化了。很多所谓的“高级提示词技巧”其实只是常识的重新包装。真正决定输出质量的，是你对问题的理解、对任务的分解、对结果的评估——这些都发生在提示词之外。

**这不是“AI 取代研究者”的宣言。**我不认为 AI 会取代研究者，至少在可预见的将来不会。但我也不认为 AI 只是“更快的搜索引擎”或“更智能的拼写检查器”。它是一种新型的工具，需要新型的使用方式。本书探讨的就是这种新型使用方式应该是什么样的。

**这不是中立的技术介绍。**我有自己的立场和偏好。我认为研究的核心是发现而不是写作，我认为可复现性比效率更重要，我认为批判性思考不能外包给机器。这些立场会影响本书的内容和论调。如果你不同意这些立场，你可能不会喜欢这本书——但我依然希望你读完它，因为不同意也是一种对话。

## 1.6 六、本书的结构与阅读建议

本书采用“多入口”设计，不同读者可以根据自己的需求选择不同的阅读路径。

### 1.6.1 6.1 整体结构

全书分为四个部分：

**第一部分（第 1–3 章）：基础认知。**这部分回答最基本的问题：AI 是什么？它能做什么、不能做什么？使用它有什么风险？如何建立正确的心理预期？建议所有读者都从这部分开始，即使你觉得“自己已经”很了解 AI 了——很多“了解”其实是误解。

**第二部分（第 4–7 章）：能力模块。**这部分按照“能力”来组织内容：信息压缩、结构化表达、代码与计算、知识管理。每种能力都是学术研究需要的，每种能力也都可以借助 AI 来增强。如果你对某种能力特别感兴趣，可以直接跳到对应章节。

**第三部分（第 8–12 章）：研究工作流。**这部分按照研究的“阶段”来组织内容：选题、文献、数据、分析、写作。如果你正处于研究的某个阶段，想知道“在这个阶段可以怎么用 AI”，可以直接跳到对应章节。

**第四部分（第 13–18 章）：层次与反思。**这部分提出“五层次框架”，讨论从个人使用到团队协作再到制度设计的不同层面。最后一章是全书的总结与反思。如果你是团队负责人或对 AI 政策感兴趣，这部分值得仔细阅读。

### 1.6.2 6.2 阅读路径建议

**如果你是刚开始使用 AI 的新手：**建议按顺序阅读，不要跳章。第一部分的基础认知特别重要——很多后面的讨论都建立在这些基础概念之上。

**如果你已经有丰富的 AI 使用经验：**可以快速浏览第一部分，然后根据自己的需求选择第二部分或第三部分的具体章节。第四部分的“五层次框架”可能会给你新的视角。

**如果你是带团队的负责人：**建议重点阅读第四部分，然后根据团队的具体需求选读其他章节。你可能需要的不是自己学会所有技巧，而是建立一套团队可以遵循的规范和标准。

**如果你只是想快速了解本书的核心观点：**可以只读第 1 章（基本原则）、第 3 章（提示词不是方法论）和第 18 章（终章）。这三章浓缩了全书最重要的思想。

### 1.6.3 6.3 每章的结构

每一章都采用相似的结构：

1. 问题引入：这一章要解决什么问题？为什么这个问题重要？
2. 核心讨论：理论分析、方法论讨论、常见误区剖析
3. 实践指南：具体的操作步骤、工作流示例、提示词参考
4. 案例分析：真实或仿真的案例，展示如何在实践中应用
5. 反思与延伸：这一章的局限性、开放问题、进一步阅读建议

你不必每次都按顺序读完一整章。如果你时间有限，可以先读“问题引入”和“反思与延伸”，快速把握要点；如果你需要立刻开始工作，可以直接跳到“实践指南”；如果你想深入思考，可以专注于“核心讨论”。

## 1.7 七、一些重要的声明

在正式开始之前，我想做几个声明。

### 1.7.1 7.1 关于时效性

AI 领域变化极快。这本书写作于 2024 年末到 2025 年初，其中提到的具体工具、功能、政策都可能已经更新。我会尽量确保书中的信息在出版时是准确的，但无法保证你阅读时依然准确。

不过，这也是我强调“方法论”而非“工具介绍”的原因。具体工具会变，但思考问题的方式不会轻易过时。如何评估工具的可靠性、如何设计人机协作的流程、如何在效率与质量之间取得平衡——这些问题在任何时候都是相关的。

### 1.7.2 7.2 关于局限性

我不是 AI 专家，不是计算机科学家，也不是技术哲学家。我的背景是社会科学研究。这意味着本书的视角必然有局限：

- 技术细节可能不够深入。如果你想了解大语言模型的内部原理，应该去读专业的技术文献。
- 社科以外的学科可能覆盖不够。我会尽量涵盖不同学科的例子，但我对某些领域的了解确实有限。
- 文化语境可能有偏差。我主要在中文和英文学术环境中工作，对其他语言和文化的学术传统了解有限。

我认为诚实地承认局限性是学术写作的基本态度。如果我假装自己什么都懂，读者反而无法判断哪些内容可信、哪些需要进一步验证。

### 1.7.3 7.3 关于 AI 辅助写作

这是一个无法回避的问题：这本书本身是不是用 AI 写的？

答案是：部分是。

更准确地说，我在写作过程中使用 AI 来做以下事情：

- 初稿的部分段落由 AI 生成，然后我进行大幅修改、重组、补充。
- 一些技术概念的解释参考了 AI 的输出，但核心观点和论证是我自己的。
- 文字润色和错别字检查使用了 AI 辅助。
- 某些案例的框架由 AI 生成，但具体细节是我补充或编造的。

我不认为这是值得隐瞒的事情。首先，这是我在书中倡导的“人机协作”的实践——如果我自己都不用 AI 来写作，这本书的建议就缺乏实践基础。其次，这也是对可复现性承诺的延伸——我愿意披露我的写作过程，让读者知道他们在读什么。

但我想强调两点：第一，本书的核心论点、整体结构、关键判断都是我的，不是 AI 的。AI 可以帮助生成文字，但不能帮我决定“什么值得说”。第二，所有重要的事实陈述和引用都经过我的人工核实。我前面讲的那个“失败的文献综述”的教训，我没有忘记。

### 1.7.4 7.4 关于争议性观点

本书包含一些可能引发争议的观点，比如：

- 很多所谓的“提示词工程”是被过度包装的简单技巧
- “效率提升”不应该是使用 AI 的主要理由
- 学术界对 AI 的某些恐惧是合理的，不应该被嘲笑为“守旧”
- 当前的很多“AI 科研工具”在商业利益驱动下夸大了自己的能力

这些观点不是随意发表的情绪，而是基于我对文献、实践和逻辑的综合判断。但判断可能是错的。如果你不同意某个观点，我欢迎你通过本书的 GitHub 仓库提出 issue，我们可以进行公开讨论。

## 1.8 八、为什么选择开源

这本书从一开始就是以开源项目的形式写作的。这个选择是刻意的，原因有几个。

### 1.8.1 8.1 可复现与可验证

学术研究的核心价值之一是可复现性：别人可以重复你的工作，验证你的结论。传统的学术出版物在这方面做得还可以——你可以看到论文的方法描述、数据来源、分析代码。但书籍出版往往是不透明的：你看到的只是最终成品，不知道作者是如何得出那些结论的。

开源改变了这一点。在本书的 GitHub 仓库里，你可以看到：

- 每一章的写作历史和修改记录
- 我使用的提示词和 AI 输出的原文
- 读者提出的质疑和我的回应
- 错误的勘误和更正

这不仅是对读者的尊重，也是对自己的约束。当你知道自己的一切都会被公开审视，你会更加谨慎。

### 1.8.2 8.2 持续更新

AI 领域变化太快，一本传统出版的书籍在出版时可能已经过时。开源允许持续更新：当新工具出现、旧工具下架、某个建议被证明有问题，我可以及时修改。读者永远可以访问最新版本。

当然，这也带来版本管理的问题。我会维护清晰的版本号和更新日志，确保读者知道自己读的是哪个版本，以及这个版本与之前版本有什么区别。

### 1.8.3 8.3 社区协作

一个人的知识和经验总是有限的。开源允许社区协作：如果你发现某个错误、有更好的案例、想补充某个领域的视角，你可以直接提交贡献。这本书不仅仅是我的，也是所有贡献者共同创作的。

我承诺会认真对待每一个贡献，无论是批评还是建议。当然，最终的编辑权在我手里——这是为了保持全书的一致性和质量——但贡献者会在书中得到致谢。

### 1.8.4 8.4 价值观的表达

最后，选择开源也是一种价值观的表达。我相信知识应该是开放的，尤其是关于如何更好地生产知识的知识。如果我写一本关于“AI 如何让学术研究更透明、更可复现”的书，然后把它锁在付费墙后面，这在逻辑上是自相矛盾的。

当然，开源不等于“没有成本”。维护这个项目需要时间和精力，我也需要生存。如果你觉得这本书对你有价值，可以通过购买实体书（如果出版的话）、赞助 GitHub 项目、或者简单地在社交媒体上分享来支持我。

## 1.9 九、我们正处于什么样的时刻

在结束这篇序章之前，我想谈谈我对当前时刻的理解。这不是技术分析，而是一种历史感——一种对“我们站在哪里”的判断。

### 1.9.1 9.1 炒作周期的高峰与低谷

如果你熟悉 Gartner 的技术炒作周期（Hype Cycle），你会知道每一项新技术都会经历类似的过程：先是过度乐观的炒作期，然后是幻灭期的低谷，最后才是稳定的应用期。大语言模型显然还处于炒作期的尾声——铺天盖地的报道开始减少，“AI 会取代一切”的论调也不如以前那么响亮。

但这不意味着 AI 不重要了。恰恰相反，真正的影响往往在炒作退潮之后才开始显现。互联网的炒作期在 2000 年泡沫破灭后结束，但互联网对社会的真正改变发生在之后的二十年。我预感大语言模型也会是类似的轨迹：当媒体不再每天报道它，当人们不再把它当作神奇的新玩具，它才会真正融入我们的工作流程，产生深远而隐秘的影响。

这也是我选择在这个时候写这本书的原因。炒作期的内容往往过于狂热或过于恐惧，都难以提供冷静的分析。而在炒作退潮、实际应用逐渐铺开的阶段，才是最需要方法论指导的时候。

## 1.9.2 9.2 学术界的特殊处境

学术界对 AI 的态度一直很矛盾。一方面，很多研究者是 AI 技术的直接创造者和研究者——如果没有学术界的贡献，今天的大语言模型根本不会存在。另一方面，学术界又是最可能被 AI 冲击的领域之一——阅读、写作、分析这些 AI 擅长的任务，恰恰是学术工作的核心。

这种矛盾导致了两种极端的反应。一些人完全拥抱 AI，把它当作解放生产力的神器，迫不及待地用它来加速每一个环节。另一些人则极度警惕，视 AI 为学术诚信的威胁，呼吁全面禁止。

我认为这两种极端都是有问题的。完全拥抱忽视了风险，而完全禁止则忽视了现实——学生和研究者已经在用 AI 了，禁止只会把使用推向地下，让问题更难管理。我们需要的是第三条路：承认 AI 会被使用，然后认真思考“如何使用才是负责任的”。

## 1.9.3 9.3 一场正在发生的范式转移

回顾科学史，每一次重大工具的引入都会带来研究范式的转移。显微镜让我们看到了微观世界，计算机让我们能处理海量数据，互联网让我们能即时获取全球文献。每一次转移都伴随着担忧和争议，但最终都被吸收进了“正常科学”的实践中。

大语言模型可能正在引发又一次这样的转移。它不只是“更快的工具”，而是一种新型的认知伙伴——它能理解（某种意义上的）语言，能生成（某种意义上的）知识，能与研究者进行（某种意义上的）对话。这是之前任何工具都做不到的。

当然，这种“理解”“知识”“对话”都需要打引号。模型是否真的理解任何东西，是一个哲学上有争议的问题。但无论如何，它的行为——它能做什么——是实实在在的。而正是这些行为，正在改变学术研究的实践。

这本书的写作，正是处于这场范式转移的早期阶段。我不能告诉你转移的终点在哪里，但我希望能帮你更清醒地认识到：转移正在发生，而你需要为此做好准备。

## 1.9.4 9.4 个人选择的重要性

面对技术变革，个体往往感到无力：趋势是不可阻挡的，个人能做什么呢？

但我想强调的是，个人选择依然重要。技术不是自动生效的，它需要被具体的人在具体的场景中采用。你选择如何使用 AI，会影响到你的研究质量；你的选择汇聚起来，会影响到整个学术社区的规范。

如果每个人都选择“快速产出”而忽视“仔细核实”，学术界就会充斥着低质量的研究。如果每个人都选择“隐藏使用”而不是“透明披露”，学术诚信的边界就会变得模糊。但反过来，如果足够多的人选择负责任的使用方式，并且公开倡导这种方式，新的规范就有可能形成。

这本书，从某种意义上说，就是我的选择——我选择花时间思考这些问题，并且把思考的结果分享出来。我希望它能帮助你做出你自己的选择。

# 1.10 十、致谢

这本书的写作得到了很多人的帮助，我想在这里表达感谢。

感谢我的同事和学生，他们在日常工作中与我分享了各种 AI 使用的经验和困惑。很多章节的灵感直接来自于我们的对话。

感谢那些在 GitHub 上提交 issue 和建议的读者（你们的名字会出现在贡献者列表中）。开源写作的意义就在于集体智慧的汇聚。

感谢那些写过”AI 科研指南”的前辈。虽然我在本章中批评了很多现有内容的不足，但我并不否认它们的价值——正是因为有了它们，我才能知道还缺什么，才能尝试填补空白。

感谢我的家人，他们容忍了我在电脑前度过的无数个夜晚。

最后，感谢你，读者。一本书只有在被阅读的时候才真正存在。你选择花时间读这本书，就是对我工作的最大肯定。

## 1.11 十一、写在最后

动笔之前，我犹豫了很久。

犹豫不是因为不知道写什么——相反，我有太多想说的。犹豫是因为担心自己没有资格说。我不是 AI 领域的权威，不是有几十年经验的老教授，甚至不是一个特别成功的学者。我凭什么写这本书？

后来我想通了：正因为我不是权威，我才可能写一本真正有用的书。权威往往离实践太远，他们的建议虽然”正确”，但普通研究者很难执行。而我，就是一个普通研究者。我每天都在用 AI 做研究，每天都在踩坑和填坑。我写的不是”应该怎么做”，而是”我是怎么做的，以及我从错误中学到了什么”。

这本书不会解决所有问题。AI 与学术研究的关系还在快速演变，很多问题现在还没有答案。但我希望它能提供一个起点：一个思考的起点，一个对话的起点，一个改进的起点。

如果你读完这本书，产生了一些想法——无论是同意、反对还是困惑——我希望你能告诉我。这本书是开源的，意味着它永远是未完成的。你的反馈会让它变得更好。

感谢你选择这本书。让我们开始吧。

## 2 第1章 AI 能做什么，不能做什么

AI 最擅长的不是“正确”，而是“像”。它能在语言上高度拟合学术表达的样子，能快速压缩文本、重写段落、生成结构化输出，但这并不等于它理解或验证了事实。研究者要先看清这一点：AI 是生成式系统，不是知识库、也不是裁判。

一个实用的判断方式是把 AI 能力分成四类：压缩、改写、重组、推理。压缩指摘要与信息提取；改写指语言润色与风格转换；重组指把散乱信息结构化；推理则是它最容易被高估的能力，因为模型擅长“构造合理解释”，却未必能承担“正确推导”的责任。

因此，AI“很强但不可靠”的根源在于：它更像一个高水平写作者，而不是可靠的事实校验器。你可以让它生成思路、提出假设、搭建框架，但关键环节必须回到人类研究者的验证流程里。把 AI 当成“高效助手”，而不是“权威来源”，这是这本书的首要原则。

## 3 第 2 章 幻觉、偏差、泄露：三类核心风险

AI 带来的最大问题不是“效率”，而是“错误的效率”。三类风险贯穿所有研究流程：幻觉、偏差、泄露。它们不只是技术问题，更是方法论和伦理问题。

幻觉是模型在不确定时仍然给出流畅答案的倾向。它在科研中的危害不在于“出现错误”，而在于“错误看起来合理”。这会侵蚀研究者的判断力，尤其在文献综述、理论归纳、背景介绍等环节，幻觉常以“自信口吻”呈现，导致难以识别。

偏差来自模型训练数据与对齐机制，会在不知不觉中渗入研究。比如对某些群体的刻板印象、对某些研究范式的偏好，都会影响输出。研究者必须意识到：AI 的“中立语气”并不意味着“中立立场”。

泄露则是更现实的风险。把未发表的研究数据、敏感信息或受限数据直接输入模型，会带来合规与伦理问题。即便使用本地模型，也要考虑数据治理与权限边界。使用 AI 前先回答一个问题：这份数据能否被任何人看到？如果答案是否定的，就不应该直接交给模型处理。

# 4 第3章 提示词不是方法论

## 4.1 一、一场关于提示词的迷信

2024年春天，我参加了一个AI辅助科研的工作坊。主讲人是一位在社交媒体上颇有影响力的“AI效率专家”，他用了整整两个小时讲解“学术研究的终极提示词模板”。

他的演示确实令人印象深刻：输入一个精心设计的提示词，模型就能输出结构完整的文献综述框架；换一个模板，就能生成看起来专业的研究假设；再换一个，还能产出像模像样的方法论描述。现场的研究生们兴奋地拍照记录，仿佛找到了学术写作的捷径。

但当我仔细审视那些输出时，一种熟悉的不安感涌上心头。那些文献综述框架虽然结构漂亮，但核心论点是空洞的；那些研究假设虽然表述规范，但缺乏对真实研究问题的深入理解；那些方法论描述虽然术语正确，但与具体研究场景脱节。

更让我担忧的是提问环节。一位博士生问：“我的研究是关于社交媒体对青少年心理健康的影响，能不能给我一个更具体的提示词？”主讲人的回答是：“你可以在我的模板里把‘研究主题’替换成你的具体领域就行了。”

这个回答让我意识到，我们正在见证一场关于提示词的集体迷信。人们把提示词当成了某种魔法咒语——只要念对了咒语，研究的难题就会迎刃而解。但真正的问题是：如果研究者本身不清楚自己要研究什么、为什么研究、如何研究，再精妙的提示词也只能产出精致的空壳。

## 4.2 二、提示词能做什么，不能做什么

让我们先诚实地面对提示词的能力边界。

### 4.2.1 2.1 提示词确实有用

我不是在说提示词毫无价值。恰恰相反，好的提示词确实能显著提升AI输出的质量。它的作用主要体现在三个方面。

**第一，约束输出格式。**如果你需要模型输出一个表格、一份大纲、或者按照特定结构组织的文本，提示词可以明确这些要求。比如“请以 markdown 表格形式呈现，包含作者、年份、主要发现、方法论四列”——这种格式约束是有效的，因为它本质上是在给模型一个模板。

**第二，设定角色与语境。**告诉模型“你是一位熟悉社会科学方法论的研究者”或“请以学术论文的语言风格回答”，可以帮助模型调整输出的语气和专业程度。这类似于告诉一个通才“现在请用专家的方式来回答”——虽然不能让通才变成真正的专家，但至少能让回答更像是专家会说的话。

**第三，分解复杂任务。**把一个大问题拆成多个小问题，每次只让模型处理一个环节，这确实能提高输出质量。比如先让模型“列出这篇论文的核心论点”，再让它“分析每个论点的证据支撑”，最后让它“评估整体论证的逻辑强度”——这种分步处理比一次性要求“请全面分析这篇论文”效果更好。

这三种作用是真实的，也是值得学习的。问题在于，很多提示词教程把这些基本技巧包装成了某种神秘的“prompt engineering”，仿佛只要掌握了正确的咒语，就能解锁AI的隐藏能力。

### 4.2.2 2.2 提示词不能做什么

更重要的是认识到提示词的边界。有些东西是再好的提示词也无法弥补的。

**第一，提示词无法提供研究者自己没有的判断力。**如果你不知道什么是好的研究问题，提示词不会帮你识别出好的研究问题。如果你不理解某个理论框架的核心逻辑，提示词不会让模型的解释变得对你有意义。AI可以生成看起来像是好问题、好框架的东西，但判断它们是否真的好，仍然需要研究者自己的专业素养。

我见过太多这样的情况：研究生用提示词生成了一堆“研究问题选项”，然后问导师“您觉得哪个好”。这个问题本身就暴露了问题所在——如果你需要别人来判断哪个研究问题更好，那么使用提示词生成这些问题并没有让你在研究能力上有任何进步。

**第二，提示词无法保证输出的正确性。**无论你的提示词写得多么精确，模型仍然可能产生幻觉、犯事实错误、或者给出逻辑上有漏洞的论证。很多提示词模板声称能“减少幻觉”或“提高准确性”，但这些说法大多缺乏严格验证。即使某个提示词在特定场景下确实降低了错误率，你也无法事先知道当前这次输出是否属于那“降低的”部分还是“剩余的”错误部分。

**第三，提示词无法替代研究过程本身。**学术研究不是输入问题、输出答案的过程。它涉及反复的阅读与思考、假设的形成与修正、数据的收集与分析、论证的构建与检验。这些过程需要时间，需要投入，需要走弯路然后纠正方向。提示词可以在某些环节提供辅助，但它无法压缩研究过程本身所认知劳动。

有一次，一位学生得意地告诉我，他用一个精心设计的提示词，让 ChatGPT 帮他完成了整个文献综述”。我让他随机挑选其中引用的三篇论文，说说每篇论文的核心贡献和方法论特点。他答不上来。这不是因为提示词不够好，而是因为提示词从根本上不可能替代真正阅读和理解文献的过程。

### 4.2.3 2.3 核心区分：流程辅助 vs. 判断替代

理解提示词的一个关键是区分两类不同的需求：流程辅助和判断替代。

**流程辅助**指的是那些规则明确、可重复、不需要太多专业判断的任务。比如：把一份英文摘要翻译成中文，按照特定格式整理参考文献，从一段文本中提取关键词，把会议记录整理成结构化的要点。这些任务的“正确答案”相对客观，可以被验证，而且验证的成本较低。在这类任务上，提示词确实能发挥显著作用。

**判断替代**指的是那些需要专业知识、涉及价值判断、或者依赖语境理解的任务。比如：评估一个研究假设是否有价值，判断一篇论文的方法论是否合理，决定某个论点是否站得住脚，选择研究设计中的关键参数。这些任务没有“标准答案”，正确性高度依赖于研究者的专业素养和对具体情境的理解。在这类任务上，提示词能做的非常有限——它可以让模型输出“像”是专业判断的东西，但不能保证那真的是正确的判断。

很多对提示词的误解，源于没有认清这个区分。人们把在“流程辅助”任务上看到的效果，错误地外推到了“判断替代”任务上。他们想：既然提示词能让翻译质量提高那么多，那应该也能让研究设计的质量提高吧？但这两类任务有本质区别。

## 4.3 三、为什么提示词被过度神化

既然提示词的边界如此明显，为什么关于它的迷信如此流行？我认为有几个原因。

### 4.3.1 3.1 简单问题的诱惑

人类天生偏好简单的解决方案。面对“如何做好学术研究”这个复杂问题，“学会写好提示词”是一个极具吸引力的答案——它具体、可操作、而且见效快。相比之下，“深入理解你的研究领域”“培养批判性思维”“在反反复试中积累经验”这些答案虽然更接近真相，但听起来既抽象又费力。

提示词模板满足了人们对“速成攻略”的渴望。收藏一个“万能提示词”比花三个月读透一个研究领域的文献要轻松得多。更诱人的是，模型确实会给出看起来不错的回应——它不会说“这个问题太复杂，我回答不了”，它会给你一个结构完整、语言流畅的输出。这种即时的正反馈强化了“提示词就是答案”的错觉。

### 4.3.2 3.2 商业利益的驱动

我们不能忽视围绕“提示词工程”已经形成的商业生态。有人卖课程，有人卖模板，有人靠分享“提示词技巧”积累社交媒体影响力。这些商业模式都依赖于一个前提：提示词是一种需要专门学习的技能，而且学会了就能获得显著回报。

这不是说所有提示词教程都是骗人的。有些确实提供了有价值的思路和技巧。但商业激励确实会导致系统的夸大。“教你”这个提示词能让你的论文写作效率提升 300% “比教你”这个技巧能在某些特定场景下略微改善输出质量”更容易吸引点击和付费。

### 4.3.3 3.3 对 AI 能力的误解

很多对提示词的迷信，根源在于对大语言模型本身能力的误解。

有一种常见的心智模型是：AI“知道”很多东西，问题只是如何“问对问题”把这些知识引出来。按照这个逻辑，提示词就像是打开宝库的钥匙——只要找到正确的钥匙，就能获取里面的宝藏。

但这个心智模型是错误的。大语言模型不是一个装满知识等待被提取的容器，而是一个被训练来预测“给定前文，下一个词最可能是什么”的系统。它在生成输出时，并没有在“搜索”某个内部数据库，而是在基于统计规律“构建”一个最可能的延续。

这意味着什么？意味着模型的输出质量不仅取决于你怎么问，更取决于它在训练时见过什么样的数据、这些数据的质量和偏向是什么、以及你的问题在多大程度上落在它的“能力范围”之内。改变提示词可以影响输出，但无法突破这些更根本的限制。

### 4.3.4 3.4 可观察性偏差

还有一个更微妙的原因：提示词是可见的，而研究能力是不可见的。

当一个研究生使用同样的提示词却得到比你更好的结果时，你看不到的是：他在使用提示词之前已经花了多少时间理解那个研究领域，他在评估输出时运用了怎样的专业判断，他如何在多轮对话中逐步修正和深化最初的输出。你能看到的只是那个“提示词”，于是你会误认为差别在于提示词本身。

真正起作用的往往是那些不可见的东西：对问题的清晰理解、对质量的准确判断、对输出的批判性评估。但这些东西难以传授、难以包装、也难以售卖。提示词之所以成为焦点，部分原因是它恰好是整个过程中最容易观察和复制的部分。

## 4.4 四、从“提问技巧”走向“研究系统”

如果提示词不是答案，那什么才是？我的建议是：把注意力从“如何问好一个问题”转向“如何构建一个可靠的研究系统”。

### 4.4.1 4.1 什么是“研究系统”

所谓研究系统，是指一套可重复、可检验、可追溯的工作流程，它规定了：

- 输入什么材料：你用什么数据、什么文献、什么信息作为起点
- 执行什么任务：你要完成哪些具体的处理和分析步骤
- 如何校验输出：你用什么标准和方法来检查结果的质量
- 怎样记录过程：你如何保留足够的信息以便日后追溯和复现

注意，这个定义中没有提到“提示词”。提示词可以是这个系统的一部分，但只是很小的一部分。系统的核心不在于你和 AI 说了什么话，而在于这些对话嵌入在怎样的工作流程中。

让我用一个具体例子来说明。假设你的任务是“整理某个领域的核心文献”。

**没有系统的做法**是：打开 ChatGPT，输入“请推荐关于 XX 领域的重要文献”，然后把输出保存下来。这个做法的问题是显而易见的：你不知道模型推荐这些文献的依据是什么，你无法验证这些推荐的质量，一个月后你可能完全记不起这份列表是怎么来的。

**有系统的做法**可能是这样的：

1. 首先，确定文献来源（比如 Web of Science、Scopus、Google Scholar），并记录检索策略（关键词、时间范围、筛选条件）。
2. 其次，下载初筛结果（比如 100 篇论文的题目和摘要），存档为原始数据。
3. 然后，用 AI 辅助进行第一轮分类：把论文按主题聚类，识别高频引用的关键作者和概念。在这个步骤中，你可以使用提示词，但关键是记录你使用的具体提示词和模型版本。
4. 接着，人工审核 AI 的分类结果，修正错误，补充遗漏。这个步骤不能省略，因为它是你真正建立对领域理解的过程。

5. 然后，针对每个主题聚类，精读 3-5 篇代表性论文，记录你自己的理解和评价。
6. 最后，整理输出一份文献地图，包括主要主题、核心论文、关键争议、研究空白。
7. 全程保留版本记录：每次修改都有时间戳，可以追溯任何结论是怎么得出的。

在这个系统中，AI 确实参与了某些环节（比如第 3 步的初步分类），但它的角色是明确的、有限的、可审计的。更重要的是，即使完全不用 AI，这个系统依然成立——AI 只是让某些步骤更高效，但没有改变系统的基本结构。

#### 4.4.2 4.2 系统的核心特征

一个好的研究系统应该具备几个核心特征。

**可重复**：如果你或你的同事遵循同样的流程，应该能够得到相似（虽然不必完全相同）的结果。这意味着流程需要足够具体，关键参数需要被记录，随机性需要被控制或至少被记录。

**可检验**：系统的每个环节都应该有某种质量检查的方式。不是等到最后才发现问题，而是在过程中就能发现和纠正问题。比如，AI 生成的文献分类应该通过抽查来验证，而不是默认接受。

**可追溯**：事后应该能够回答“这个结论是怎么得出的”。这需要充分的记录：使用了什么数据，经过了哪些处理，做了哪些决策，依据是什么。当 AI 参与时，这一点尤其重要——你需要记录使用的模型、版本、提示词、输出。

**容错**：系统应该假设会出错，并在设计中考虑如何发现和纠正错误。冗余检查、交叉验证、人工复核——这些机制不是“额外工作”，而是保证质量的必要投入。

#### 4.4.3 4.3 AI 在系统中的位置

在一个设计良好的研究系统中，AI 的位置应该是明确的和受限的。

**明确**意味着：你清楚 AI 在哪些环节参与、执行什么任务、产出什么结果。不是“我用 AI 帮忙做了一些事情”，而是“AI 在步骤 3 执行了初步分类，在步骤 5 协助了文本润色”。

**受限**意味着：AI 不应该出现在那些需要核心判断的关键节点上，或者即使出现也必须有严格的人工复核。识别哪些节点是“关键的”，本身就是设计系统的重要工作。

一个有用的问题是：如果 AI 在这个环节犯了错，后果是什么？如果后果严重且难以通过后续步骤发现，那这个环节就需要额外的保护——要么不使用 AI，要么必须有独立的验证机制。

另一个有用的问题是：这个环节的质量，我能通过什么方式检验？如果你无法独立验证 AI 的输出质量（比如因为你自己不具备相关知识），那就不应该在这个环节依赖 AI——因为你没有办法知道它是否做对了。

### 4.5 五、建立自己的工作流原则

基于以上讨论，我想提出几个构建 AI 辅助研究工作流的核心原则。这些原则不是具体的提示词模板，而是设计工作流程时应该遵循的指导思想。

#### 4.5.1 5.1 任务拆解原则

**核心思想**：把大任务分解成小任务，把模糊任务转化为具体任务。

大语言模型在处理边界清晰的小任务时表现更好。“帮我写一篇关于气候变化政策的综述”是一个糟糕的任务描述，因为它包含太多隐含的子任务：确定综述范围、检索相关文献、筛选核心论文、归纳主要观点、识别研究空白、组织成文……每一个子任务都有自己的复杂性。

更好的做法是显式地拆解：

- 任务 1：明确综述的具体问题（关于气候变化政策的什么？碳税的经济效应？还是国际协调机制的有效性？）
- 任务 2：确定检索策略（在哪些数据库检索？用什么关键词？时间范围是多少？）

- 任务 3：初步筛选（根据题目和摘要，排除明显不相关的文献）
- 任务 4：深入阅读核心文献（精读 20–30 篇最相关的论文）
- 任务 5：归纳和综合（提取主要发现，识别共识和争议）
- .....

在这个拆解中，有些任务适合 AI 辅助（比如任务 3 的初步筛选），有些则必须由人类主导（比如任务 1 的问题界定和任务 4 的深入阅读）。拆解本身就是在明确 AI 的角色边界。

**实践建议：**在使用 AI 之前，先用纸和笔（或者简单的文本文件）把你的任务拆解成尽可能小的单元。问自己：这个单元的输入是什么？期望的输出是什么？我如何判断输出的质量？

### 4.5.2 5.2 层级输出原则

**核心思想：**不要期望一次性得到完美输出，而是通过多轮迭代逐步精化。

很多人使用 AI 的方式是：输入一个提示词，得到一个输出，如果不满意就换一个提示词再试。这种方式把每次交互都当作独立的事件，没有利用对话的累积性。

更有效的方式是层级式的：

- **第一层：粗略轮廓。**先让 AI 给出一个粗略的框架或方向。这个阶段的目标不是获得可用的输出，而是快速探索可能性。
- **第二层：细化某个方向。**在粗略轮廓中选择一个值得深入的方向，让 AI 进一步展开。这时可以提供更多具体信息和约束条件。
- **第三层：具体化和验证。**对细化后的内容进行具体化，同时开始验证关键信息的准确性。这个阶段经常需要跳出 AI 对话，去查阅原始资料。
- **第四层：整合和润色。**把经过验证的内容整合成最终输出，进行语言和格式上的润色。

这个过程不是线性的。你可能在第三层发现第一层的方向选错了，需要回退重来。但层级化的思路保证了每一步都有明确的目标，错误能够被较早发现。

**实践建议：**在每一层结束时，暂停下来问自己：这个输出是否足够准确，值得在它的基础上继续深入？如果你发现自己在第三层才发现第一层就有根本性问题，下次就要考虑在第一层投入更多验证工作。

### 4.5.3 5.3 验证机制原则

**核心思想：**任何 AI 输出都需要验证，验证的力度应该与错误的后果成正比。

验证不是可选的附加步骤，而是使用 AI 的必要组成部分。问题是验证需要成本，所以需要根据风险来分配验证资源。

**低风险任务**（错误易发现，后果可逆，成本低）：可以采用抽查式验证。比如 AI 帮你整理了一份 100 条的引用列表，你可以随机抽查 10 条来判断整体质量。

**中风险任务**（错误较隐蔽，后果有影响，成本中等）：需要系统性验证。比如 AI 帮你总结了 10 篇论文的核心论点，你应该每一篇都回到原文核对，确保概括准确。

**高风险任务**（错误难发现，后果严重，成本高）：需要独立验证。比如 AI 给出了某个统计方法的解释，你应该查阅教科书或权威资料独立确认，而不是仅仅让同一个 AI “再解释一遍”。

有一种常见的错误是用 AI 来验证 AI 的输出——让同一个模型检查自己之前的回答，或者换一个提示词让它“确认”之前的结论。这种做法几乎没有验证价值，因为模型的偏差是系统性的，它不会通过重复访问而消失。

**实践建议：**在开始一个任务之前，先问自己：如果 AI 在这里犯错，我会怎么发现？如果你答不上来，说明这个任务可能不适合交给 AI，或者你需要先设计一个验证机制。

### 4.5.4 5.4 版本记录原则

**核心思想：**保留足够的信息，使得任何结论都可以追溯其来源。

这个原则在传统研究方法论中已经很重要（记录数据来源、分析代码、决策理由），在 AI 辅助研究中更加重要，因为 AI 增加了一个不透明的环节。

需要记录的信息包括：

- **模型信息**: 使用的模型名称和版本（比如 GPT-4-turbo-2024-04-09），使用的参数设置（比如 temperature）
- **输入信息**: 完整的提示词，提供给模型的上下文材料
- **输出信息**: 模型的完整输出（不仅仅是你使用的一部分）
- **处理信息**: 你对输出进行了哪些修改、筛选、整合
- **验证信息**: 你进行了哪些验证，验证的结果是什么

这听起来像是大量额外工作，但实际上可以通过工具和习惯来简化。很多 AI 对话工具允许导出完整历史；你可以建立标准化的文件夹结构来组织这些记录；关键决策点可以用简单的笔记来记录理由。

**实践建议**: 至少做到这一点——每次使用 AI 完成一个重要任务，“事后能够回答”这个结论是基于什么输入、经过什么处理、做过什么验证”。如果你发现自己答不上来，说明记录工作需要加强。

#### 4.5.5 5.5 边界意识原则

**核心思想**: 清晰地认识到哪些任务适合 AI 辅助，哪些不适合。

这个原则是前面所有原则的基础。如果你不知道 AI 的能力边界在哪里，就无法正确地设计任务拆解、输出层级、验证机制。

一些经验性的指导：

**适合 AI 辅助的任务特征**: – 规则明确，“正确答案”相对客观 – 验证成本低（你可以较容易地检查输出质量）– 错误后果可控（即使出错也可以补救）– 任务可分解，每个子任务边界清晰

**不适合 AI 辅助的任务特征**: – 需要深度专业判断 – 你自己无法独立验证输出质量 – 错误后果严重且难以察觉 – 任务高度依赖特定语境和隐性知识

当然，这不是非黑即白的分类。很多任务介于两者之间，需要具体分析。关键是培养这种分析习惯：在把任务交给 AI 之前，先思考这个任务的特征，评估 AI 可能带来的价值和风险。

**实践建议**: 保持一个“AI 使用日志”，记录你使用 AI 的场景、效果、遇到的问题。一段时间后，你会对自己的使用模式有更清晰的认识，知道哪些场景值得继续探索，哪些应该避免。

### 4.6 六、一个完整的例子

让我用一个具体的例子来说明如何应用这些原则。假设你是一位社会科学研究生，你的导师让你“调研一下最近五年关于算法歧视的实证研究”。

#### 4.6.1 6.1 没有系统的设计

你打开 ChatGPT，输入：“请帮我整理最近五年关于算法歧视的实证研究，包括主要发现和方法论。”

模型给出了一份看起来不错的列表，包含了十几项研究的简要描述。你把这份列表整理一下，交给了导师。

两周后，导师问你：“你提到的那个 ProPublica 关于 COMPAS 的研究，他们具体的分析方法是什么？有什么局限性？”你答不上来，因为你从来没有读过那篇原始报告。

#### 4.6.2 6.2 有系统的设计

##### 第一步：明确任务边界（任务拆解）

你首先问自己几个问题：– “最近五年”是指哪个时间范围？（确定：2019–2024）– “算法歧视”的定义是什么？包括哪些类型？（确定：主要聚焦于机器学习算法在信贷、招聘、刑事司法等领域的歧视性结果）– “实证研究”是指什么？（确定：包括真实数据分析、审计研究、实验研究，不包括纯理论论文）– 导师期待的产出是什么？（确定：一份综述报告，约 3000 字，需要包括研究方法的比较分析）

##### 第二步：设计检索策略

你决定从两个来源检索文献：– 学术数据库：Web of Science，使用关键词“algorithmic bias” OR “algorithmic discrimination” AND “empirical” – 补充来源：Google Scholar，用于捕捉可能不在 WoS 中的跨学科研究和工作论文

你记录了检索时间、检索词、检索条件、结果数量。

#### 第三步：初步筛选（AI 辅助）

从检索中获得约 200 篇文献的题目和摘要。你把这些导出为一个文件，然后使用 AI 进行第一轮筛选。

提示词：“以下是 200 篇关于算法歧视的论文摘要。请帮我识别其中属于‘实证研究’的论文（即使用真实数据或实验方法研究算法歧视现象的研究），并按照研究领域（信贷、招聘、刑事司法、医疗、其他）分类。”

AI 给出了一个分类列表，标注了约 80 篇可能是实证研究的论文。

#### 第四步：验证和调整（验证机制）

你随机抽取 10 篇 AI 标注为“实证研究”的论文，快速浏览摘要验证分类是否正确。发现 8 篇分类准确，2 篇其实是综述性文章被误标。

你又抽取 5 篇 AI 标注为“非实证研究”的论文，检查是否有遗漏。发现 1 篇其实是实证研究但被遗漏了。

基于这个抽查结果，你估计 AI 的分类大致可靠，但存在约 20% 的错误率。你决定对所有 80 篇“实证研究”的摘要进行人工快速审核，最终确定约 65 篇进入下一轮。

#### 第五步：深入阅读（不使用 AI）

你从 65 篇中选择 20 篇代表性论文进行精读。选择标准包括：高被引、方法论创新、涵盖不同领域。

这个步骤没有使用 AI。你阅读每篇论文的全文，记录研究问题、数据来源、分析方法、主要发现、局限性。这是建立真正理解的过程，不可压缩。

#### 第六步：归纳综合

你基于深度阅读的笔记，开始归纳主要发现和方法论趋势。在这个阶段，你可以用 AI 辅助整理思路：

“我阅读了 20 篇关于算法歧视的实证研究，以下是我的笔记摘要。请帮我识别这些研究在方法论上的共同点和差异，以及主要发现的共识和争议。”

AI 给出了一个初步的综合，你在此基础上修改和补充，加入你自己的分析和判断。

#### 第七步：撰写报告（迭代润色）

你先手写一个粗略的提纲，然后逐节展开写作。写完初稿后，可以用 AI 帮助检查语言流畅度、识别论述空白。

但最终的判断——这篇综述是否准确反映了文献的实际状况——只有你自己能做出，因为只有你真正读过那些原始论文。

#### 第八步：保留记录（版本记录）

整个过程中，你保留了：  
– 检索记录（日期、数据库、检索词、结果数量）  
– 原始文献列表（200 篇的题目、摘要、来源）  
– AI 对话记录（每次使用 AI 的输入和输出）  
– 阅读笔记（20 篇精读论文的详细笔记）  
– 写作版本（综述报告的多个修订版本）

当导师问你任何问题，你都可以追溯到原始证据。

### 4.6.3 6.3 两种做法的对比

两种做法的时间投入差别很大——第二种可能需要两周，第一种可能只需要两小时。但产出的质量完全不同。

更重要的是，第二种做法在过程中建立了真正的理解。你读完那 20 篇论文后，你对这个领域有了真正的认识，能够参与专业讨论，能够识别新研究的价值和问题。第一种做法什么都没有留下——既没有知识的积累，也没有可追溯的记录。

AI 在第二种做法中确实发挥了作用：它帮助你进行了初步分类，节省了大量筛选时间；它帮助你综合笔记，加速了写作过程。但 AI 的角色是明确的、有限的、受控的。它是研究系统中的一个组件，而不是整个系统的替代品。

## 4.7 七、写在最后

让我回到本章开始时那个 AI 工作坊的场景。

那位主讲人教授的提示词技巧，严格来说并没有错。格式约束、角色设定、任务分解——这些确实是有用的技巧。问题在于，他把这些技巧包装成了某种“终极解决方案”，暗示掌握了这些提示词就能解决学术研究的核心难题。

但学术研究的核心难题——提出有价值的问题、设计可靠的方法、做出准确的判断、构建有说服力的论

证——不是任何提示词能够解决的。这些能力需要长期积累，需要深入阅读，需要反复实践，需要失败和纠错。提示词可以在这个过程中提供一些辅助，但它不是捷径，因为根本就没有捷径。

这就是为什么我说“提示词不是方法论”。方法论是关于“如何做好研究”的系统性思考，它涉及研究设计、证据标准、推理规则、质量控制。提示词只是与 AI 交互的界面，它是技术层面的，而不是方法论层面的。把提示词当作方法论来学习，就像把“如何握笔”当作“如何写好文章”来学习——不是说握笔不重要，但它解决不了写作的真正难题。

我希望这一章能够帮助读者建立一个更健康的与 AI 的关系。这个关系的核心是：**AI 是工具，不是导师；是助手，不是替代；是流程中的一个环节，不是整个流程。**

理解了这一点，你就会知道应该把精力放在哪里。不是去收集更多的提示词模板，而是去培养自己的研究能力。不是去追求“一键生成”的幻觉，而是去构建可靠的工作系统。不是去问“有没有更好的提示词”，而是去问“我的研究流程哪里可以改进”。

提示词可以优化，但方法论需要建构。前者是技巧，后者是能力。这本书的目的，正是帮助你从技巧走向能力。

# 5 第4章 信息压缩与文献整理

## 5.1 引言：从信息过载到结构化理解

学术研究者面临的核心挑战之一是信息过载。一个活跃的研究领域每年可能产出数千篇论文，而研究者的阅读时间是有限的。传统的文献整理方法——逐篇阅读、手工分类、人工综合——在信息爆炸的时代显得力不从心。大语言模型的出现为这一困境提供了新的可能性：它们能够快速处理大量文本，提取关键信息，生成结构化的摘要和分类。

然而，这种可能性伴随着显著的风险。模型可能遗漏重要文献，可能错误归纳论点，可能生成不存在的引用。如果研究者不加辨别地接受 AI 的输出，信息压缩就会变成信息扭曲。本章的目标是探讨如何在文献整理工作中有效利用 AI 的信息压缩能力，同时建立必要的质量保障机制。

本章将依次讨论三个核心议题：文献筛选与主题聚类的方法，长上下文模型的正确使用方式，以及可信综述生成的验证策略。贯穿这些讨论的核心原则是：AI 的输出应被视为“索引”而非“结论”，是研究过程的起点而非终点。

## 5.2 一、文献筛选与主题聚类

### 5.2.1 1.1 文献筛选的两阶段模型

文献筛选是任何研究项目的基础工作。传统的筛选流程通常包括：确定检索策略、执行数据库检索、根据题目和摘要进行初筛、根据全文进行精筛。这个流程的瓶颈在于初筛阶段——面对数百甚至数千条检索结果，研究者需要逐条判断相关性，这是一项耗时且认知负担沉重的工作。

AI 辅助可以显著提高初筛阶段的效率，但前提是研究者理解这种辅助的性质和边界。我建议采用“两阶段模型”来组织 AI 辅助的文献筛选工作。

第一阶段是 AI 辅助的粗筛。在这个阶段，研究者将检索结果（通常是题目和摘要的列表）提交给模型，要求模型根据预设的纳入和排除标准进行初步分类。模型的任务是识别“明显相关”“明显不相关”和“需要进一步判断”三类文献。这个阶段的目标不是获得最终的筛选结果，而是快速缩小需要人工审阅的范围。

第二阶段是人工复核与精筛。研究者对 AI 标记为“明显相关”和“需要进一步判断”的文献进行人工审阅，做出最终的纳入决策。同时，研究者应当抽样检查 AI 标记为“明显不相关”的文献，评估 AI 的假阴性率（即被错误排除的相关文献比例）。如果假阴性率过高，需要调整第一阶段的指令或标准。

这个两阶段模型的核心逻辑是：AI 负责处理“容易”的决策（明显相关或明显不相关的文献），人类负责处理“困难”的决策（边界情况和最终判断）。这种分工既利用了 AI 的处理速度，又保留了人类的判断权威。

### 5.2.2 1.2 设计有效的筛选指令

AI 辅助筛选的效果高度依赖于指令的质量。一个有效的筛选指令应当包含以下要素。

首先是明确的研究问题陈述。研究者需要清晰地告诉模型，这次文献检索要回答什么问题。模糊的问题陈述会导致模糊的筛选结果。例如，“关于气候变化的研究”是一个过于宽泛的陈述，而“关于碳税政策对企业减排行为影响的实证研究”则提供了明确的边界。

其次是具体的纳入标准。纳入标准应当是可操作的、可判断的。例如：“纳入标准：(1) 研究对象为企业或行业层面的减排行为；(2) 研究方法为实证分析（包括定量和定性）；(3) 研究涉及碳税或碳定价政策；(4) 发表时间为 2015 年至今。”这些标准使得模型能够对每一篇文献做出相对客观的判断。

第三是明确的排除标准。排除标准帮助模型识别那些表面相关但实际不符合需求的文献。例如：“排除标准：(1) 纯理论或模拟研究，无实证数据；(2) 研究对象为个人消费者行为而非企业行为；(3) 仅讨论碳交易而不涉及碳税；(4) 会议摘要、书评、社论等非研究性文献。”

第四是输出格式的规定。研究者应当明确告诉模型如何呈现筛选结果。例如：“请将每篇文献分类为‘纳入’‘排除’或‘待定’，并简要说明分类理由（一句话）。”结构化的输出便于后续的人工复核和记录。

### 5.2.3 1.3 主题聚类的策略

主题聚类是文献整理的另一项核心任务。当研究者面对一个新的研究领域时，往往需要首先了解这个领域的的主要研究主题、核心争论和知识结构。传统的做法是通过广泛阅读逐步建立这种理解，但这需要大量时间。AI 可以帮助研究者快速获得一个初步的领域地图。

有效的主题聚类需要遵循几个原则。

第一个原则是层次化聚类。不要试图一次性获得一个完美的分类体系，而是采用从粗到细的迭代策略。第一轮聚类可以识别 3–5 个大的主题领域；第二轮在每个大主题下进一步细分子主题；第三轮识别子主题之间的交叉和联系。这种层次化的方法既便于理解，也便于发现聚类中的问题。

第二个原则是多维度聚类。同一批文献可以从不同维度进行分类。例如，按研究主题分类、按研究方法分类、按研究对象分类、按理论框架分类。不同的分类维度揭示领域的不同面向。研究者应当根据自己的研究需求选择最相关的分类维度，或者综合使用多个维度。

第三个原则是保留边界案例。在聚类过程中，总会遇到难以归类的文献——它们可能跨越多个主题，或者不完全符合任何现有类别。这些边界案例往往具有特殊的价值：它们可能代表新兴的研究方向，或者揭示现有分类体系的局限。研究者不应强行将这些文献塞入某个类别，而应单独标记，作为进一步探索的线索。

第四个原则是验证聚类结果。AI 生成的聚类结果需要人工验证。验证的方法包括：从每个聚类中随机抽取几篇文献，检查它们是否真的属于同一主题；检查聚类的标签是否准确反映了聚类内容；寻找是否有明显的误分类。如果发现系统性的问题，需要调整聚类指令并重新执行。

### 5.2.4 1.4 一个实践案例

为了说明上述原则的应用，我将描述一个具体的文献筛选与聚类案例。

假设研究者正在开展一项关于“社交媒体对学术传播影响”的研究。初步的数据库检索返回了 350 篇相关文献的题目和摘要。研究者的任务是从中筛选出与研究问题直接相关的文献，并对这些文献进行主题分类。

第一步，研究者设计了筛选指令。纳入标准包括：实证研究社交媒体（如 Twitter/X、ResearchGate、Academia.edu）对学术论文传播的影响；研究涉及可测量的传播指标（如引用、阅读量、讨论度）；研究对象为学术论文或学术成果。排除标准包括：纯描述性的社交媒体使用调查；仅讨论社交媒体对教学的影响；技术性的平台分析而非传播效果研究；非英文文献。

第二步，研究者将 350 篇文献的题目和摘要提交给模型，要求模型按照上述标准进行分类。模型返回了以下结果：92 篇标记为“纳入”，198 篇标记为“排除”，60 篇标记为“待定”。

第三步，研究者对结果进行人工复核。首先，研究者审阅了 60 篇“待定”文献，最终纳入其中的 35 篇，排除 25 篇。然后，研究者从 198 篇“排除”文献中随机抽取 20 篇进行检查，发现其中 3 篇实际上符合纳入标准（假阴性率约 15%）。基于这个发现，研究者决定扩大抽查范围，又检查了 30 篇“排除”文献，额外发现了 4 篇应当纳入的文献。最终的纳入文献数量为 134 篇。

第四步，研究者对 134 篇纳入文献进行主题聚类。第一轮聚类识别出四个大主题：(1) 社交媒体与引用影响；(2) 社交媒体与公众传播；(3) 学术社交网络平台研究；(4) 社交媒体使用行为研究。第二轮聚类在每个大主题下进一步细分。例如，“社交媒体与引用影响”下细分为“Twitter 讨论与引用相关性”“Altmetrics 指标研究”“因果效应估计”三个子主题。

第五步，研究者验证聚类结果。从每个子主题中抽取 3 篇文献进行审阅，确认分类的准确性。在验证过程中发现，“Altmetrics 指标研究”这个子主题过于宽泛，包含了指标本身的研究和指标应用的研究两类不同性质的文献，需要进一步拆分。

通过这个案例可以看到，AI 在筛选和聚类过程中发挥了重要的辅助作用，显著减少了人工处理的工作量。但每个关键节点都有人工的参与和验证，确保了最终结果的可靠性。

## 5.3 二、长上下文模型的正确用法

### 5.3.1 2.1 长上下文能力的本质与局限

近年来，大语言模型的上下文窗口经历了显著扩展。早期的模型只能处理几千个 token，而最新的模型已经能够处理数十万甚至上百万个 token。这意味着研究者理论上可以将几十篇论文的全文一次性输入模型，要求模型进行综合分析。

然而，“能够处理”不等于“能够有效处理”。研究表明，即使是长上下文模型，其在不同位置的信息利用效率也存在显著差异。一般而言，模型对上下文开头和结尾的信息利用较好，而对中间部分的信息利用较差——这被称为“中间丢失”现象。此外，当上下文过长时，模型生成的回答往往变得更加泛化和模糊，丢失了对具体细节的把握。

这些发现对学术研究有重要启示。将大量论文一次性输入模型，期望获得一个全面准确的综合分析，这种期望是不现实的。更有效的策略是理解长上下文的局限，设计相应的使用方法。

### 5.3.2 2.2 分层处理策略

面对长上下文的局限，我建议采用分层处理策略。这个策略的核心思想是：不要试图让模型一次完成所有工作，而是将任务分解为多个层次，每个层次处理适量的信息。

第一层是单篇论文的信息提取。对于每一篇需要纳入分析的论文，首先单独提取关键信息。提取的内容可能包括：研究问题、理论框架、研究方法、核心发现、主要结论、局限性。这个层次的任务相对简单，模型通常能够较好地完成。研究者可以设计一个标准化的提取模板，确保从每篇论文中获取可比较的信息。

第二层是小组论文的比较分析。将具有相似主题或方法的 3-5 篇论文组成一个小组，要求模型对这个小组进行比较分析。比较的维度可能包括：研究发现的一致性与差异、方法论的异同、理论框架的关系。这个层次的任务开始涉及综合和推理，模型的表现会有更大的不确定性，需要研究者的仔细审核。

第三层是跨组的综合归纳。在完成所有小组的分析之后，将各组的分析结果作为输入，要求模型进行更高层次的综合。这个层次的输入是已经经过压缩和结构化的信息，而非原始的论文全文，因此更适合模型处理。

这种分层策略的优势在于：每一层的任务都保持在模型能够有效处理的范围内；每一层都有明确的输出，便于人工审核和质量控制；错误可以在较早的层次被发现和纠正，而不会传递到最终输出。

### 5.3.3 2.3 输入质量的重要性

长上下文模型的输出质量高度依赖于输入的质量。“垃圾进，垃圾出”的原则在这里尤其适用。研究者在准备输入材料时，应当注意以下几点。

首先是格式的一致性。如果输入的材料格式混乱——有的是纯文本，有的是 PDF 提取的乱码，有的包含大量表格和图片描述——模型的处理效果会大打折扣。研究者应当在输入之前对材料进行预处理，确保格式的统一和清洁。

其次是信息的相关性。不是论文的所有部分都与研究者的分析需求相关。如果研究者只关心研究方法和发现，那么将论文的致谢、参考文献列表、补充材料等全部输入模型，不仅浪费了上下文空间，还可能引入噪音。研究者应当有选择地提取和输入最相关的部分。

第三是结构的清晰性。在输入多篇论文时，应当使用清晰的分隔符和标识，让模型能够区分不同论文的内容。例如，可以在每篇论文的开头添加标识：“[论文 1] 作者：xxx，标题：xxx，年份：xxx”。这种结构化的输入有助于模型准确地引用和归因。

### 5.3.4 2.4 输出设计与质量控制

除了优化输入，研究者还应当精心设计期望的输出格式，并建立相应的质量控制机制。

在输出设计方面，研究者应当明确告诉模型输出的结构和详细程度。例如：“请按以下格式输出：(1) 主要发现综述（200 字以内）；(2) 研究方法比较表格；(3) 发现一致性分析；(4) 发现差异分析；(5) 研究空白识别。”结构化的输出要求不仅便于后续使用，也使得质量检查更加系统化。

在质量控制方面，最重要的原则是可追溯性。模型的每一个陈述都应当能够追溯到具体的原始来源。研究者可以要求模型在输出中标注来源，例如：“Smith 等（2020）发现社交媒体讨论与引用存在正相关 [论

文 3]”。然后研究者需要回到原始论文验证这个陈述是否准确。

另一个重要的质量控制方法是交叉验证。对于同一批论文，可以从不同的角度或使用不同的指令让模型进行分析，然后比较结果的一致性。如果两次分析得出截然不同的结论，说明至少有一次分析存在问题，需要人工介入检查。

### 5.3.5 2.5 一个错误使用的案例

为了说明不当使用长上下文模型的风险，我将描述一个反面案例。

一位研究生正在撰写一篇关于数字鸿沟的文献综述。他收集了 45 篇相关论文的 PDF 文件，使用工具将这些 PDF 转换为文本，然后一次性将所有文本（约 50 万字）输入一个长上下文模型，要求模型“写一篇全面的文献综述，总结这 45 篇论文的主要发现和研究趋势”。

模型生成了一篇约 3000 字的综述，结构完整，语言流畅。这位研究生非常满意，稍作修改后就准备使用。但他的导师要求他核对其中几个关键陈述的来源。

核对的结果令人担忧。综述中的一个陈述是：“Chen 等（2019）的元分析发现，数字鸿沟对教育成就的影响效应量为  $d=0.42$ 。”但当研究生查找这篇论文时，发现 Chen 等（2019）的研究根本不是元分析，而是一项单一的调查研究，也没有报告过  $d=0.42$  这个效应量。另一个陈述是：“近年来的研究趋势从关注接入差距转向关注使用差距和技能差距。”这个陈述本身可能是正确的，但模型没有提供任何具体证据，研究生也无法追溯这个“趋势”判断的来源。

这个案例说明了几个问题。第一，将大量未经预处理的材料一次性输入模型，模型无法有效利用所有信息。第二，期望模型一次性完成复杂的综合任务，超出了模型的可靠能力范围。第三，没有建立验证机制，导致错误直到被要求核对时才被发现。第四，模型生成的流畅文本给研究者造成了虚假的信心，掩盖了内容上的问题。

正确的做法应该是采用前文描述的分层处理策略：先对每篇论文单独提取信息，然后分组比较，最后综合归纳，每一步都进行验证。

## 5.4 三、可信综述生成：引用与验证

### 5.4.1 3.1 为什么综述生成是高风险任务

文献综述是学术写作中最常见也是最容易被 AI 辅助的任务之一。但它同时也是风险最高的任务之一。这种高风险源于几个因素。

第一，综述涉及对多个来源的准确归纳。模型不仅需要理解每一篇论文的内容，还需要准确地表述每篇论文说了什么、没说什么。任何一个错误归纳都可能误导读者，损害综述的可信度。

第二，综述需要准确的引用。学术写作的核心规范之一是“言必有据”——每一个关于先前研究的陈述都应当有明确的来源。模型生成虚假引用的问题在学术界已经广为人知。即使引用的论文确实存在，模型对其内容的概括也可能不准确。

第三，综述涉及判断和评价。好的综述不仅仅是论文的罗列，还需要对研究进行评价、识别研究空白、提出未来方向。这些任务需要深入的领域知识和批判性思维，超出了当前模型的可靠能力范围。

第四，综述的错误难以被非专家发现。一篇语言流畅、结构合理的综述，即使内容有问题，读者也很难察觉。尤其是当读者没有阅读过原始论文时，他们无法判断综述的归纳是否准确。

基于这些风险，我建议将综述生成视为“高风险任务”，采用最严格的验证机制。

### 5.4.2 3.2 AI 在综述生成中的适当角色

尽管风险很高，AI 在综述生成中仍然可以发挥有价值的辅助作用，只要研究者正确定位 AI 的角色。

AI 适合承担的任务包括：生成初步的组织框架、识别可能的主题分类、对单篇论文进行摘要、提示可能被遗漏的角度。这些任务的共同特点是：它们是“启发性”的而非“结论性”的——AI 的输出是供研究者参考和修改的草稿，而非最终产品。

AI 不适合承担的任务包括：对论文观点的权威性概括、对研究质量的评价、对研究领域的趋势判断、任何需要作为最终产品使用的内容。这些任务需要研究者自己完成，或者在 AI 辅助的基础上进行充分的人工审核和修改。

一个有用的思维框架是区分“生成”和“使用”两个阶段。在生成阶段，研究者可以自由地使用 AI 来探索想法、生成草稿、获取灵感。在使用阶段——也就是将内容放入最终的学术产品中——每一个陈述都必须经过验证，确保其准确性和可追溯性。

### 5.4.3 3.3 引用验证的系统方法

引用验证是综述生成中最关键的质量控制环节。我建议采用系统化的验证方法。

第一步是列出所有引用。将 AI 生成的综述中所有的文献引用提取出来，形成一个完整的列表。每个引用应当记录：作者、年份、被引用的具体观点或发现。

第二步是验证引用的存在性。确认每一篇被引用的文献确实存在。检查作者姓名、发表年份、期刊或会议名称是否正确。使用 Google Scholar、Web of Science 或其他数据库进行验证。如果发现一篇文献不存在或信息有误，这是一个明确的红旗，需要删除该引用并检查 AI 输出中的其他引用。

第三步是验证引用的准确性。这是最耗时但也是最重要的步骤。对于每一个引用，回到原始论文，确认综述中的概括是否准确反映了原文的观点。常见的问题包括：夸大了原文的发现、忽略了原文的限定条件、将原文的假设表述为结论、混淆了不同论文的内容。

第四步是标记验证状态。为每个引用标记验证状态：已验证、待验证、存疑、已删除。这个标记帮助研究者追踪验证进度，确保没有遗漏。

第五步是迭代修正。根据验证结果修正综述内容。对于发现的错误，不仅要修正具体的陈述，还要反思错误的模式——是某一类论文的概括容易出错，还是某一种陈述类型容易出错？这种反思有助于改进后续的工作流程。

### 5.4.4 3.4 建立可追溯的证据链

可信的综述不仅需要引用准确，还需要建立清晰的证据链，使读者能够追溯任何陈述的来源。

证据链的建立始于数据管理。研究者应当建立一个结构化的文献数据库，记录每篇论文的基本信息、核心内容摘要、以及在综述中如何被使用。文献管理软件如 Zotero、EndNote、Mendeley 可以支持这一工作，但研究者需要超越简单的书目管理，为每篇文献添加结构化的笔记和标签。

证据链的维护贯穿写作过程。在综述写作的每一个阶段，研究者都应当能够回答：这个陈述的依据是什么？来自哪篇文献的哪个部分？如果无法回答这些问题，说明证据链存在断裂，需要补充或修正。

证据链的最终呈现体现在综述文本中。读者应当能够通过引用信息找到原始来源，并验证综述中的陈述。这要求引用必须具体和准确——引用到具体的页码或章节，而非笼统地引用整篇论文。

### 5.4.5 3.5 AI 辅助综述的披露与伦理

使用 AI 辅助生成学术综述涉及伦理问题，研究者有责任进行适当的披露。

不同的学术机构和期刊对 AI 使用的披露要求不同。有些要求详细说明 AI 在写作中的角色，有些只要求声明是否使用了 AI，有些尚未制定明确的政策。研究者应当了解并遵守相关的规定。

即使没有明确的披露要求，透明度也是学术诚信的基本原则。我建议在使用 AI 辅助综述生成时，至少在方法部分说明：使用了哪些 AI 工具、AI 在哪些环节发挥了作用、研究者采取了哪些验证措施。这种披露不会减损研究的价值，反而展示了研究者的方法论自觉。

需要强调的是，披露 AI 的使用并不能免除研究者对最终内容的责任。无论 AI 在生成过程中扮演了多大的角色，发表的内容归研究者所有，研究者对其准确性和诚信性负有完全的责任。AI 是工具，而使用工具的方式——以及对工具输出的把关——是研究者的选择和责任。

## 5.5 四、从工具到能力：信息素养的新维度

### 5.5.1 4.1 信息压缩作为研究者的核心能力

本章的讨论可能给读者一个印象：AI 正在接管文献整理的工作。但我想在结束时强调一个相反的观点：信息压缩是研究者的核心能力，AI 的参与不应削弱这一能力的培养。

信息压缩能力包括几个层面。第一个层面是快速识别信息价值的能力——看一篇论文的摘要，就能判断它与研究问题的相关程度。第二个层面是提取核心论点的能力——阅读一篇论文，能够准确把握其主要贡献和局限。第三个层面是综合多个来源的能力——将多篇论文的发现整合为连贯的知识图景。第四个层面是识别知识空白的能力——看到现有文献的整体轮廓，能够发现尚未被研究的问题。

这些能力需要通过实践来培养。阅读大量论文、写作多篇综述、经历导师的反馈和修改——这个过程不能被跳过或压缩。AI 可以帮助提高效率，但不能替代学习过程本身。

### 5.5.2 4.2 人机协作的最优配置

对于研究者而言，关键问题是：如何配置人类和 AI 在文献整理工作中的角色，以实现最优的结果？

我建议的配置原则是：AI 负责规模化处理和格式化任务，人类负责判断和验证任务。

在文献筛选中，AI 可以处理数百篇文献的初步分类，人类负责边界情况的判断和分类准确性的验证。在主题聚类中，AI 可以提供初步的聚类方案，人类负责评估聚类的合理性并进行调整。在信息提取中，AI 可以从论文中提取结构化信息，人类负责检查提取的准确性。在综述写作中，AI 可以协助生成草稿和组织框架，人类负责内容的验证、判断和最终表述。

这种配置的核心逻辑是：利用 AI 的速度优势处理那些规则明确、可批量执行的任务，同时保留人类对关键节点的控制权。人类的时间和认知资源应当集中在那些真正需要专业判断的环节。

### 5.5.3 4.3 未来研究者的信息素养

AI 工具的普及正在改变研究者所需的信息素养。传统的信息素养强调检索技能——如何使用数据库、如何构建检索策略、如何管理文献。这些技能仍然重要，但不再足够。

未来的研究者需要掌握新的信息素养。第一是 AI 素养——理解 AI 工具的能力和局限，知道如何有效地使用它们，也知道在什么情况下不应该使用它们。第二是验证素养——能够系统地检查 AI 输出的准确性，建立可追溯的证据链，识别和纠正错误。第三是工作流设计素养——能够设计人机协作的工作流程，优化任务在人类和 AI 之间的分配。第四是元认知素养——保持对自己认知过程的反思，避免因过度依赖 AI 而导致的能力退化。

这些素养不会自动获得，需要有意识地培养。研究者应当在实践中不断反思自己使用 AI 的方式，评估效果，调整策略。学术机构和导师也应当为研究生提供相关的培训和指导。

## 5.6 结语

本章讨论了 AI 辅助文献整理的方法和原则。核心论点可以总结为：AI 能够显著提高信息压缩的效率，但这种效率提升必须以可靠性为前提。研究者应当把 AI 的输出视为“索引”而非“结论”，是需要验证的起点而非可以信任的终点。

在文献筛选中，两阶段模型——AI 辅助粗筛加人工复核精筛——能够在效率和准确性之间取得平衡。在使用长上下文模型时，分层处理策略能够避免信息过载导致的质量下降。在综述生成中，严格的引用验证和证据链管理是确保可信度的必要措施。

贯穿这些讨论的，是对研究者主体性的强调。AI 改变了文献整理的方式，但没有改变文献整理的目的——那就是建立研究者对一个领域的深入理解。这种理解不能被外包给机器，只能通过研究者自己的阅读、思考和综合来建立。AI 是这个过程中的有力工具，但工具的价值最终取决于使用工具的人。

## 6 第 5 章 结构化表达与写作

AI 能快速生成提纲、段落与结构，这在写作初期非常有用。你可以用它来整理研究框架、提出章节结构、梳理理论证顺序。但要记住：结构不是思想，结构只是容器，关键是你要把真正的论证放进去。

“学术语言”与“AI 语言”之间的差别常常被忽视。AI 写出来的文字通常流畅、完整，但容易缺乏学术写作所需的精确性与责任感。它倾向于“把话说满”，而研究写作更强调“留出不确定性”。因此，AI 生成文本需要二次加工：删去过度自信的表达，补上证据与限定条件。

在翻译与润色上，AI 的优势是速度，但风险是术语与语境的偏差。最好的策略是：让 AI 做“第一遍粗修”，再由研究者做“关键点审校”。尤其是方法与结果部分，不能完全依赖 AI 完成。

## 7 第 6 章 代码与计算能力

AI 对代码的帮助不在于“自动生成”，而在于“降低门槛”。它可以帮助你把模糊想法转成可运行的脚本，减少重复劳动，并在遇到错误时快速定位问题。但这不意味着你可以不理解代码。任何研究级分析都要求你能解释每一步处理逻辑。

在数据清洗、格式转换、批量处理等任务上，AI 的效率优势非常明显。它能快速给出脚本框架和可复用函数，适合用作“草稿生成器”。你应当把它当作代码搭子：让它产出、让你审核、共同迭代。

在统计分析与可视化上，AI 可以协助你构建模型、输出图表，但不能替代对模型假设的理解。模型选择、变量定义、结果解释都属于研究者的责任。AI 提供的是加速，而不是合法性。

## 8 第 7 章 知识管理与协作

AI 的最大价值之一是“对话式启发”，但对话如果不被整理，就会迅速消失。研究需要记录、归档与版本管理。你需要把 AI 对话转成可检索的研究材料，而不是散落的聊天记录。

一个可用的流程是：对话 → 摘要 → 标签 → 归档。每次关键对话都应输出结构化总结，并注明时间、模型、提示词与用途。这些记录构成研究的“隐形材料”，决定了你能否复现自己的思路。

在团队协作中，AI 既能提升效率，也会模糊责任边界。谁写的、谁验证的、谁承担错误？这些问题必须在协作中明确。开源和透明不只是技术选择，也是团队协作的伦理底线。

## 9 第8章 选题与研究问题

选题阶段是 AI 最容易“帮倒忙”的环节。它能快速生成许多漂亮的问题，但这些问题往往缺乏可检验性、现实可行性或理论价值。研究者必须先明确研究问题的“可研究性”，再使用 AI 辅助拆解。

一个有效的方法是先由人写出核心问题，再让 AI 做“多角度压力测试”：是否存在可用数据？是否已有大量研究？是否存在明确的识别路径？AI 擅长从多个角度给出提醒，但这些提醒必须由研究者来做最终判断。

最常见的风险是“问题被做空”：AI 倾向于把复杂问题简化为泛化叙述，导致研究失去锋利度。研究者要主动保持问题的边界感，避免被 AI“优化成毫无棱角的宏大命题”。

## 10 第 9 章 文献与理论构建

AI 可以帮你搭建“文献地图”，但搭建“理论链条”仍然需要研究者的判断力。文献地图强调覆盖面，理论链条强调逻辑性。二者不是同一件事。

在文献整理阶段，AI 适合做主题聚合与脉络描述：哪些研究关注同一问题、哪些方法常被使用、哪些争议存在。但理论构建需要你明确变量关系与因果机制，AI 只能辅助解释，不能替你承担理论责任。

所谓“AI 综述陷阱”是指：输出看起来完整，但引用无法核对，或者概念混杂不清。避免陷阱的办法是把 AI 综述当成“初稿目录”，逐条回到原文校验，并用“证据链”替换“叙述链”。

# 11 第 10 章 数据获取与构造

数据获取是 AI 加速最明显的环节之一。无论是公开数据、爬虫还是 API 接口，AI 都能帮助你快速写出脚本并完成批量抓取。但“获取速度”不等于“数据质量”，任何数据都需要清晰的采集记录与合法性说明。

AI 生成数据是更具争议的方向。它可用于模拟、训练或思维实验，但不能轻易替代真实样本。研究者必须区分“生成数据用于方法测试”和“生成数据用于经验结论”这两种完全不同的用途。

“硅样本”的方法论争议在于：AI 是否能代表人类行为？如果 AI 是研究对象，它本身就不是“人”。使用硅样本必须明确其理论定位，否则会把“模型输出”误当成“社会事实”。

## 12 第 11 章 数据清洗与分析

数据清洗是科研中最耗时也最容易被忽略的环节。AI 的优势在于快速生成清洗脚本和规则，尤其适用于文本数据的结构化处理。但清洗规则背后往往包含方法论选择，这些选择必须透明化、可解释、可复现。

在统计分析阶段，AI 可以帮助搭建模型、生成图表、解释结果，但研究者必须对模型假设负责。AI 擅长“讲故事”，但不擅长“守住边界”。如果你让 AI 解释结果，它可能会给出看似合理但并不严谨的因果叙述。

因此，AI 的角色应该是“分析助手”，而不是“结论生成器”。任何结论性表述必须回到数据与方法本身，这是学术研究的底线。

## 13 第 12 章 写作、投稿与传播

AI 最受欢迎的应用场景是写作。你可以用它生成初稿、润色语言、调整结构，但这并不意味着它能替你“完成论文”。论文的核心价值来自研究设计与证据，而不是表达方式。

在投稿流程中，AI 可以帮助你整理摘要、润色投稿信、模拟审稿人视角。但“反 AI 检测”不应该成为研究者的目标。真正合理的策略是：确保内容真实、证据充分、逻辑清晰。所谓“AI 味”，往往来自缺乏具体细节与可验证信息，而不是语言风格本身。

在答辩与传播阶段，AI 适合帮助你做摘要、演讲稿与受众调整，但关键观点必须是你自己的。AI 可以把话说漂亮，但不能替你承担学术立场。

## 14 第 13 章 RA Level：工具层

在工具层，AI 等同于一个高效率的研究助理。它能完成大量重复性工作：整理数据、生成初稿、清洗文本、批量翻译。这个层级的核心价值是节省时间和成本。

但工具层的风险在于“过度依赖”。当 AI 替代了研究者的基础劳动，研究者也可能丧失对材料的直觉理解。你必须确保自己仍然掌握数据细节和研究逻辑，否则 AI 的效率会把你带向“看似完成但无法解释”的境地。

最重要的问题是：谁在控制流程？如果 AI 的输出驱动了你的研究方向，那么你已经从“使用工具”变成“被工具牵引”。工具层的原则是：AI 做脏活累活，人保留判断权。

## 15 第 14 章 Supervisor Level：认知协作层

在认知协作层，AI 不只是执行者，而是“对话式导师”。它能帮助你提出假设、模拟审稿人、补足背景知识，甚至推动你跨学科思考。这一层是 AI 最有创造性地应用。

但同样，风险也在于“依赖与懒惰”。当你习惯用 AI 生成思路，可能会丧失独立构建问题的能力。AI 会不断给出“合理建议”，但合理并不等于有价值。研究者必须保持自己的问题意识与判断标准。

认知协作层的关键是：让 AI 成为“思维刺激器”，而不是“思维替代品”。你可以借助 AI 扩展视野，但不能放弃自己的判断。

## 16 第 15 章 Domain Expert Level：推理与建模层

在推理与建模层，AI 能够模拟“领域专家”的思维方式，帮助你理解模型、构建算法或补齐技术短板。对跨学科研究者而言，这一层极具价值，因为它能缩短学习曲线。

但必须明确：推理不等于理论。AI 可以生成“看似合理”的模型，却不一定满足学科内部的规范与逻辑要求。理论构建需要领域知识、文献脉络与方法论约束，而这些不是 AI 能自动承担的。

这一层的正确使用方式是：让 AI 提供“草稿与解释”，由研究者完成“理论合法性与方法论判断”。领域知识的不可替代性，恰恰是学术研究者的核心价值。

## 17 第 16 章 Agent Level: AI 作为行动者

当 AI 开始模拟人类、参与互动或自动决策时，它不再只是工具，而成为“行动者”。这对社会科学提出新的本体论问题：研究对象是否还仅限于“人”？当 AI 参与社会系统，它是否也构成社会事实的一部分？

在这一层，AI 被用来构建虚拟社会、模拟行为、生成硅样本。这些方法可以帮助研究者探索复杂系统，但也必须承认其局限性：模型输出并不等于真实社会。虚拟样本的价值在于“实验性启发”，而不是“经验结论”。

Agent 层提醒我们：AI 不仅改变研究工具，也改变研究对象。研究者必须明确自己是在研究“人”、还是研究“人—AI 互动”，或是研究“AI 本体”。这些区分决定了理论框架与解释边界。

## 18 第 17 章 Governance Level：制度与治理层

在治理层，问题已经超越“如何使用 AI”，而是“AI 如何改变学术制度”。当 AI 可以大规模生成文本，学术评价体系会发生什么变化？当研究流程被自动化，学术劳动的价值如何被重新定义？

这一层的核心是结构性影响：论文产出可能变得更快，但“学术质量”可能被稀释；评审机制可能更依赖表面表达，忽视真实研究价值。AI 带来的不是简单的效率提升，而是制度逻辑的重构。

研究者需要在治理层承担新的责任：明确 AI 使用边界、推动透明化、参与学术伦理讨论。真正的挑战不是“AI 能不能用”，而是“学术共同体如何定义可信”。

## 19 终章 | AI 时代研究者的新能力结构

AI 让研究更快，这是事实。但更快不等于更深。真正的问题是：在 AI 时代，研究者的不可替代能力是什么？答案不是“会用工具”，而是“能够提出问题、建立证据链、守住方法论边界”。

未来的研究者需要新的能力结构：一是系统化的问题意识；二是对证据的严格要求；三是对 AI 输出的批判性判断。AI 可以帮助你节省时间，但它不会替你承担学术责任。

开源在这个时代变得更重要。它不仅是分享资源，更是对学术可信度的回应。面对 AI 可能带来的“空洞高产”，开源与可复现是一种必要的抵抗：把研究拉回事实、证据与共同体的监督之中。

## 20 附录（可选）

本书后续版本将提供以下配套材料，全部开源更新：

- AI 科研工具箱清单（按任务分类）
- 提示词模板与工作流脚本（可复用）
- 案例复现材料与数据链接（含版本号）
- AI 使用伦理与声明模板（投稿与答辩场景）

附录的目的不是堆工具，而是让方法论落地。读者可以根据自己的研究领域替换工具，但不应跳过验证与记录步骤。

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. <https://doi.org/10.1093/comjnl/27.2.97>.