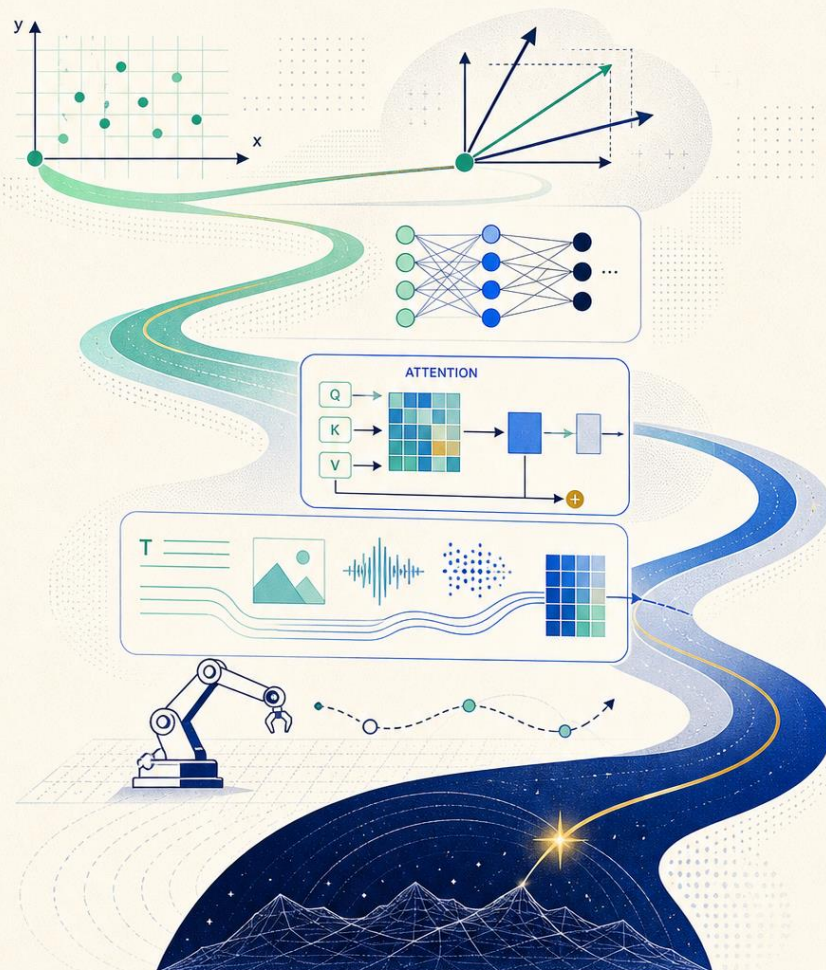


从零开始学AI

普通人与高中生的24个月互动教材

2026.08 Beta



NEWVAR

第四卷：前沿、系统与研究

版本、边界与使用方法

结论先行：这是一套面向普通人与高中生的系统教材，不是速成承诺。当前为 **2026.08 Beta**；只有技术评审、事实核查和领域专家签字全部完成后，才能进入 stable。

项目	内容
版本	2026.08 Beta
卷册	第四卷：前沿、系统与研究
本卷章节	24
全书结构	4 卷 · 24 单元 · 96 章
实践资产	96 个互动实验 · 24 个项目 · 288 道练习
正文规模	约 689,651 个汉字
核实日期	2026-08-06
版权	Copyright © 2026 NEWVAR LAB. All rights reserved.

能力终点

本书以独立开发、部署、论文阅读、实验复现、失败分析和研究问题提出为终点能力；不承诺仅靠一本书自动成为世界顶尖研究者。

全面覆盖的可审计边界

“全面”指对主要 AI 知识域、能力框架、工程链路和研究方法进行可审计覆盖，不等于收录互联网上每一页资料。

每章学习契约

- 先看结论、价值和完成后能做什么。
- 完成先修检查，用生活类比获得方向，再主动找出类比失效处。
- 一次只改变一个核心变量；网页提供键盘操作和静态替代说明。
- 按直觉、数学、最小实现、PyTorch、实验、失败、迁移的顺序推进。
- 工程线与研究线至少选择一条；三级练习先独立完成，再看答案。
- 所有结果必须标成实际测量、教学模拟或文献引用；不能混写。
- 遇到高风险决策、隐私数据或证据不足时停止自动化并交给给人。

隐私与材料规则

全书案例只使用虚构或公开数据。不要在练习、截图、Notebook 或网站笔记中输入姓名、账号、联系方式、密钥、医疗记录、成绩等个人信息。网页学习记录仅保存在当前设备，可匿名导入导出；不建账号、不做行为追踪。

目录

目录由 Word 字段生成。打开文件后若页码尚未刷新，请按 Ctrl+A，再按 F9。

版本、边界与使用方法..... 2

 能力终点..... 2

 全面覆盖的可审计边界..... 3

 每章学习契约..... 3

 隐私与材料规则..... 3

目录..... 4

本卷 24 章路线图..... 36

 月 19 · 图学习、推荐与时序智能..... 36

 月 20 · 基础模型和语言模型从零实现..... 36

 月 21 · 多模态、世界模型与 Agent..... 36

 月 22 · 数据、规模化训练、推理和 MLOps..... 37

 月 23 · 安全、隐私、可解释性与治理..... 37

 月 24 · AI for Science、论文复现与原创毕业项目..... 38

第 73 章 图表示、消息传递与 GNN..... 39

 1. 为什么重要，现在能做什么..... 39

 2. 先修检查与生活类比..... 40

 3. 互动实验：只改变一个变量..... 40

 4. 直觉到数学..... 41

 5. 最小实现..... 41

6. 五轮系统化理解	42
1. 问题边界	42
3. 机制与因果链	43
7. 评估与不确定性	44
9. 工程与复现	45
11. 迁移与研究	46
7. PyTorch 实现路线	47
8. 实验设计与结果记录	48
9. 常见失败与排查顺序	48
10. 迁移任务	49
11. 三级练习	49
12. 本章能力检查	49
13. 来源与核实	50
第 74 章 推荐、检索与排序系统	51
1. 为什么重要，现在能做什么	51
2. 先修检查与生活类比	52
3. 互动实验：只改变一个变量	52
4. 直觉到数学	53
5. 最小实现	53
6. 五轮系统化理解	54
1. 问题边界	54

3. 机制与因果链.....	55
7. 评估与不确定性	56
9. 工程与复现	57
11. 迁移与研究	58
7. PyTorch 实现路线	60
8. 实验设计与结果记录	60
9. 常见失败与排查顺序	60
10. 迁移任务.....	61
11. 三级练习.....	61
12. 本章能力检查.....	62
13. 来源与核实	62
第 75 章 时间序列、预测与回测.....	63
1. 为什么重要，现在能做什么.....	63
2. 先修检查与生活类比	64
3. 互动实验：只改变一个变量.....	64
4. 直觉到数学	65
5. 最小实现.....	65
6. 五轮系统化理解	66
1. 问题边界.....	66
3. 机制与因果链.....	67
7. 评估与不确定性	68

9. 工程与复现	69
11. 迁移与研究	70
7. PyTorch 实现路线	71
8. 实验设计与结果记录	72
9. 常见失败与排查顺序	72
10. 迁移任务	73
11. 三级练习	73
12. 本章能力检查	73
13. 来源与核实	74
第 76 章 时空建模、异常与决策智能	75
1. 为什么重要，现在能做什么	75
2. 先修检查与生活类比	76
3. 互动实验：只改变一个变量	76
4. 直觉到数学	77
5. 最小实现	77
6. 五轮系统化理解	78
1. 问题边界	78
3. 机制与因果链	79
7. 评估与不确定性	80
9. 工程与复现	81
11. 迁移与研究	82

7. PyTorch 实现路线	84
8. 实验设计与结果记录	84
9. 常见失败与排查顺序	84
10. 迁移任务	85
11. 三级练习	85
12. 本章能力检查	86
13. 来源与核实	86
月 19 阶段项目：构建一个不跟踪个人的内容发现原型	87
交付物	87
验收方法	87
第 77 章 语言模型词元器与嵌入	88
1. 为什么重要，现在能做什么	88
2. 先修检查与生活类比	89
3. 互动实验：只改变一个变量	89
4. 直觉到数学	90
5. 最小实现	90
6. 五轮系统化理解	91
1. 问题边界	91
3. 机制与因果链	92
7. 评估与不确定性	93
9. 工程与复现	94

11. 迁移与研究	95
7. PyTorch 实现路线	97
8. 实验设计与结果记录	97
9. 常见失败与排查顺序	97
10. 迁移任务	98
11. 三级练习	98
12. 本章能力检查	99
13. 来源与核实	99
第 78 章 从零实现因果 Transformer	101
1. 为什么重要，现在能做什么	101
2. 先修检查与生活类比	102
3. 互动实验：只改变一个变量	102
4. 直觉到数学	103
5. 最小实现	103
6. 五轮系统化理解	104
1. 问题边界	104
3. 机制与因果链	105
7. 评估与不确定性	106
9. 工程与复现	107
11. 迁移与研究	108
7. PyTorch 实现路线	110

8. 实验设计与结果记录	110
9. 常见失败与排查顺序	110
10. 迁移任务	111
11. 三级练习	111
12. 本章能力检查	112
13. 来源与核实	112
第 79 章 预训练目标、数据与缩放	114
1. 为什么重要，现在能做什么	114
2. 先修检查与生活类比	115
3. 互动实验：只改变一个变量	115
4. 直觉到数学	116
5. 最小实现	116
6. 五轮系统化理解	117
1. 问题边界	117
3. 机制与因果链	118
7. 评估与不确定性	119
9. 工程与复现	121
11. 迁移与研究	122
7. PyTorch 实现路线	123
8. 实验设计与结果记录	123
9. 常见失败与排查顺序	124

10. 迁移任务	124
11. 三级练习	125
12. 本章能力检查	125
13. 来源与核实	125
第 80 章 指令微调、偏好学习、RAG 与评估	127
1. 为什么重要，现在能做什么	127
2. 先修检查与生活类比	128
3. 互动实验：只改变一个变量	128
4. 直觉到数学	129
5. 最小实现	129
6. 五轮系统化理解	130
1. 问题边界	130
3. 机制与因果链	131
7. 评估与不确定性	132
9. 工程与复现	133
11. 迁移与研究	135
7. PyTorch 实现路线	136
8. 实验设计与结果记录	136
9. 常见失败与排查顺序	137
10. 迁移任务	137
11. 三级练习	137

12. 本章能力检查	138
13. 来源与核实	138
月 20 阶段项目：从零训练一个微型语言模型	140
交付物	140
验收方法	140
第 81 章 多模态表示、对齐与融合	141
1. 为什么重要，现在能做什么	141
2. 先修检查与生活类比	142
3. 互动实验：只改变一个变量	142
4. 直觉到数学	143
5. 最小实现	143
6. 五轮系统化理解	144
1. 问题边界	144
3. 机制与因果链	145
7. 评估与不确定性	146
9. 工程与复现	147
11. 迁移与研究	149
7. PyTorch 实现路线	150
8. 实验设计与结果记录	150
9. 常见失败与排查顺序	151
10. 迁移任务	151

11. 三级练习	151
12. 本章能力检查	152
13. 来源与核实	152
第 82 章 世界模型、视频与行动预测	154
1. 为什么重要，现在能做什么	154
2. 先修检查与生活类比	155
3. 互动实验：只改变一个变量	155
4. 直觉到数学	156
5. 最小实现	156
6. 五轮系统化理解	157
1. 问题边界	157
3. 机制与因果链	158
7. 评估与不确定性	159
9. 工程与复现	160
11. 迁移与研究	161
7. PyTorch 实现路线	162
8. 实验设计与结果记录	163
9. 常见失败与排查顺序	163
10. 迁移任务	164
11. 三级练习	164
12. 本章能力检查	164

13. 来源与核实	165
第 83 章 Agent 循环、工具、记忆与规划	166
1. 为什么重要，现在能做什么	166
2. 先修检查与生活类比	167
3. 互动实验：只改变一个变量	167
4. 直觉到数学	168
5. 最小实现	168
6. 五轮系统化理解	169
1. 问题边界	169
3. 机制与因果链	170
7. 评估与不确定性	171
9. 工程与复现	172
11. 迁移与研究	174
7. PyTorch 实现路线	175
8. 实验设计与结果记录	175
9. 常见失败与排查顺序	176
10. 迁移任务	176
11. 三级练习	176
12. 本章能力检查	177
13. 来源与核实	177
第 84 章 Agent 评估、安全与人在回路	179

1. 为什么重要，现在能做什么	179
2. 先修检查与生活类比	180
3. 互动实验：只改变一个变量	180
4. 直觉到数学	181
5. 最小实现	181
6. 五轮系统化理解	182
1. 问题边界	182
3. 机制与因果链	183
7. 评估与不确定性	184
9. 工程与复现	185
11. 迁移与研究	187
7. PyTorch 实现路线	188
8. 实验设计与结果记录	188
9. 常见失败与排查顺序	189
10. 迁移任务	189
11. 三级练习	189
12. 本章能力检查	190
13. 来源与核实	190
月 21 阶段项目：做一个只读、可审计的研究助理 Agent	192
交付物	192
验收方法	192

第 85 章 数据治理、去重、质量与许可.....193

1. 为什么重要，现在能做什么.....193

2. 先修检查与生活类比194

3. 互动实验：只改变一个变量.....194

4. 直觉到数学195

5. 最小实现.....195

6. 五轮系统化理解196

1. 问题边界.....196

3. 机制与因果链.....197

7. 评估与不确定性198

9. 工程与复现199

11. 迁移与研究201

7. PyTorch 实现路线202

8. 实验设计与结果记录202

9. 常见失败与排查顺序203

10. 迁移任务.....203

11. 三级练习.....203

12. 本章能力检查.....204

13. 来源与核实204

第 86 章 分布式训练、并行与检查点.....206

1. 为什么重要，现在能做什么.....206

2. 先修检查与生活类比	207
3. 互动实验：只改变一个变量.....	207
4. 直觉到数学	208
5. 最小实现.....	208
6. 五轮系统化理解	209
1. 问题边界.....	209
3. 机制与因果链.....	210
7. 评估与不确定性	211
9. 工程与复现	212
11. 迁移与研究	214
7. PyTorch 实现路线	215
8. 实验设计与结果记录	215
9. 常见失败与排查顺序	216
10. 迁移任务.....	216
11. 三级练习.....	216
12. 本章能力检查.....	217
13. 来源与核实	217
第 87 章 推理、量化、蒸馏与服务.....	219
1. 为什么重要，现在能做什么.....	219
2. 先修检查与生活类比	220
3. 互动实验：只改变一个变量.....	220

4. 直觉到数学	221
5. 最小实现.....	221
6. 五轮系统化理解	222
1. 问题边界.....	222
3. 机制与因果链.....	223
7. 评估与不确定性	224
9. 工程与复现	225
11. 迁移与研究	227
7. PyTorch 实现路线	228
8. 实验设计与结果记录	228
9. 常见失败与排查顺序	229
10. 迁移任务.....	229
11. 三级练习.....	229
12. 本章能力检查.....	230
13. 来源与核实	230
第 88 章 MLOps、监控、可靠性、成本与碳.....	232
1. 为什么重要，现在能做什么.....	232
2. 先修检查与生活类比	233
3. 互动实验：只改变一个变量.....	233
4. 直觉到数学	234
5. 最小实现.....	234

6. 五轮系统化理解	235
1. 问题边界	235
3. 机制与因果链	236
7. 评估与不确定性	238
9. 工程与复现	239
11. 迁移与研究	240
7. PyTorch 实现路线	241
8. 实验设计与结果记录	242
9. 常见失败与排查顺序	242
10. 迁移任务	243
11. 三级练习	243
12. 本章能力检查	243
13. 来源与核实	244
月 22 阶段项目：部署一个有监控和回滚的小模型服务	245
交付物	245
验收方法	245
第 89 章 安全、滥用与威胁建模	246
1. 为什么重要，现在能做什么	246
2. 先修检查与生活类比	247
3. 互动实验：只改变一个变量	247
4. 直觉到数学	248

5. 最小实现.....	248
6. 五轮系统化理解	249
1. 问题边界.....	249
3. 机制与因果链.....	250
7. 评估与不确定性	251
9. 工程与复现	252
11. 迁移与研究	254
7. PyTorch 实现路线	255
8. 实验设计与结果记录	255
9. 常见失败与排查顺序	256
10. 迁移任务.....	256
11. 三级练习.....	256
12. 本章能力检查.....	257
13. 来源与核实	257
第 90 章 隐私、差分隐私、联邦学习与模型安全.....	259
1. 为什么重要，现在能做什么.....	259
2. 先修检查与生活类比	260
3. 互动实验：只改变一个变量.....	260
4. 直觉到数学	261
5. 最小实现.....	261
6. 五轮系统化理解	262

1. 问题边界	262
3. 机制与因果链	263
7. 评估与不确定性	264
9. 工程与复现	265
11. 迁移与研究	267
7. PyTorch 实现路线	268
8. 实验设计与结果记录	268
9. 常见失败与排查顺序	269
10. 迁移任务	269
11. 三级练习	269
12. 本章能力检查	270
13. 来源与核实	270
第 91 章 可解释性、审计与模型文档	272
1. 为什么重要，现在能做什么	272
2. 先修检查与生活类比	273
3. 互动实验：只改变一个变量	273
4. 直觉到数学	274
5. 最小实现	274
6. 五轮系统化理解	275
1. 问题边界	275
3. 机制与因果链	276

7. 评估与不确定性	277
9. 工程与复现	278
11. 迁移与研究	280
7. PyTorch 实现路线	281
8. 实验设计与结果记录	281
9. 常见失败与排查顺序	282
10. 迁移任务.....	282
11. 三级练习.....	282
12. 本章能力检查.....	283
13. 来源与核实	283
第 92 章 治理、法律、标准与事故响应.....	285
1. 为什么重要，现在能做什么.....	285
2. 先修检查与生活类比	286
3. 互动实验：只改变一个变量.....	286
4. 直觉到数学	287
5. 最小实现.....	287
6. 五轮系统化理解	288
1. 问题边界.....	288
3. 机制与因果链.....	289
7. 评估与不确定性	291
9. 工程与复现	292

11. 迁移与研究	293
7. PyTorch 实现路线	294
8. 实验设计与结果记录	295
9. 常见失败与排查顺序	295
10. 迁移任务	296
11. 三级练习	296
12. 本章能力检查	296
13. 来源与核实	297
月 23 阶段项目：完成一次端到端 AI 系统红队与治理评审	299
交付物	299
验收方法	299
第 93 章 AI for Science：从预测到发现	300
1. 为什么重要，现在能做什么	300
2. 先修检查与生活类比	301
3. 互动实验：只改变一个变量	301
4. 直觉到数学	302
5. 最小实现	302
6. 五轮系统化理解	303
1. 问题边界	303
3. 机制与因果链	304
7. 评估与不确定性	305

9. 工程与复现	306
11. 迁移与研究	308
7. PyTorch 实现路线	309
8. 实验设计与结果记录	309
9. 常见失败与排查顺序	310
10. 迁移任务	310
11. 三级练习	310
12. 本章能力检查	311
13. 来源与核实	311
第 94 章 读论文、复现实验与统计核查	313
1. 为什么重要，现在能做什么	313
2. 先修检查与生活类比	314
3. 互动实验：只改变一个变量	314
4. 直觉到数学	315
5. 最小实现	315
6. 五轮系统化理解	316
1. 问题边界	316
3. 机制与因果链	317
7. 评估与不确定性	318
9. 工程与复现	319
11. 迁移与研究	321

7. PyTorch 实现路线	322
8. 实验设计与结果记录	322
9. 常见失败与排查顺序	323
10. 迁移任务	323
11. 三级练习	323
12. 本章能力检查	324
13. 来源与核实	324
第 95 章 研究问题、创新、消融与沟通	326
1. 为什么重要，现在能做什么	326
2. 先修检查与生活类比	327
3. 互动实验：只改变一个变量	327
4. 直觉到数学	328
5. 最小实现	328
6. 五轮系统化理解	329
1. 问题边界	329
3. 机制与因果链	330
7. 评估与不确定性	331
9. 工程与复现	332
11. 迁移与研究	334
7. PyTorch 实现路线	335
8. 实验设计与结果记录	335

9. 常见失败与排查顺序	335
10. 迁移任务	336
11. 三级练习	336
12. 本章能力检查	337
13. 来源与核实	337
第 96 章 毕业项目：部署、答辩与前沿追踪	338
1. 为什么重要，现在能做什么	338
2. 先修检查与生活类比	339
3. 互动实验：只改变一个变量	339
4. 直觉到数学	340
5. 最小实现	340
6. 五轮系统化理解	341
1. 问题边界	341
3. 机制与因果链	342
7. 评估与不确定性	343
9. 工程与复现	345
11. 迁移与研究	346
7. PyTorch 实现路线	347
8. 实验设计与结果记录	347
9. 常见失败与排查顺序	348
10. 迁移任务	348

11. 三级练习	349
12. 本章能力检查	349
13. 来源与核实	349
月 24 阶段项目：原创毕业项目与公开答辩包	351
交付物	351
验收方法	351
独立练习答案	352
第 73 章答案 · 图表示、消息传递与 GNN	352
基础题参考	352
应用题参考	352
挑战题参考	352
自评量规	353
第 74 章答案 · 推荐、检索与排序系统	353
基础题参考	353
应用题参考	353
挑战题参考	353
自评量规	354
第 75 章答案 · 时间序列、预测与回测	354
基础题参考	354
应用题参考	354
挑战题参考	354

自评量规.....	355
第 76 章答案 · 时空建模、异常与决策智能.....	355
基础题参考	355
应用题参考	355
挑战题参考	355
自评量规.....	356
第 77 章答案 · 语言模型词元器与嵌入.....	356
基础题参考	356
应用题参考	356
挑战题参考	356
自评量规.....	357
第 78 章答案 · 从零实现因果 Transformer.....	357
基础题参考	357
应用题参考	357
挑战题参考	357
自评量规.....	358
第 79 章答案 · 预训练目标、数据与缩放	358
基础题参考	358
应用题参考	358
挑战题参考	358
自评量规.....	359

第 80 章答案 · 指令微调、偏好学习、RAG 与评估.....359

 基础题参考359

 应用题参考359

 挑战题参考359

 自评量规.....360

第 81 章答案 · 多模态表示、对齐与融合360

 基础题参考360

 应用题参考360

 挑战题参考360

 自评量规.....361

第 82 章答案 · 世界模型、视频与行动预测.....361

 基础题参考361

 应用题参考361

 挑战题参考361

 自评量规.....362

第 83 章答案 · Agent 循环、工具、记忆与规划.....362

 基础题参考362

 应用题参考362

 挑战题参考362

 自评量规.....363

第 84 章答案 · Agent 评估、安全与人在回路363

基础题参考	363
应用题参考	363
挑战题参考	363
自评量规.....	364
第 85 章答案 · 数据治理、去重、质量与许可	364
基础题参考	364
应用题参考	364
挑战题参考	364
自评量规.....	365
第 86 章答案 · 分布式训练、并行与检查点.....	365
基础题参考	365
应用题参考	365
挑战题参考	365
自评量规.....	366
第 87 章答案 · 推理、量化、蒸馏与服务	366
基础题参考	366
应用题参考	366
挑战题参考	366
自评量规.....	367
第 88 章答案 · MLOps、监控、可靠性、成本与碳.....	367
基础题参考	367

应用题参考	367
挑战题参考	367
自评量规.....	368
第 89 章答案 · 安全、滥用与威胁建模.....	368
基础题参考	368
应用题参考	368
挑战题参考	368
自评量规.....	369
第 90 章答案 · 隐私、差分隐私、联邦学习与模型安全	369
基础题参考	369
应用题参考	369
挑战题参考	369
自评量规.....	370
第 91 章答案 · 可解释性、审计与模型文档.....	370
基础题参考	370
应用题参考	370
挑战题参考	370
自评量规.....	371
第 92 章答案 · 治理、法律、标准与事故响应	371
基础题参考	371
应用题参考	371

挑战题参考	372
自评量规.....	372
第 93 章答案 · AI for Science：从预测到发现	372
基础题参考	372
应用题参考	372
挑战题参考	373
自评量规.....	373
第 94 章答案 · 读论文、复现实验与统计核查	373
基础题参考	373
应用题参考	373
挑战题参考	374
自评量规.....	374
第 95 章答案 · 研究问题、创新、消融与沟通	374
基础题参考	374
应用题参考	374
挑战题参考	375
自评量规.....	375
第 96 章答案 · 毕业项目：部署、答辩与前沿追踪	375
基础题参考	375
应用题参考	375
挑战题参考	376

自评量规.....	376
本卷来源台账.....	377
[S004] Artificial Intelligence Risk Management Framework (AI RMF 1.0)	377
[S005] AI Index Report 2026	377
[S023] Fairlearn User Guide	378
[S024] Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations	378
[S030] Attention Is All You Need	379
[S046] CS224W: Machine Learning with Graphs	380
[S047] Semi-Supervised Classification with Graph Convolutional Networks	380
[S048] Matrix Factorization Techniques for Recommender Systems	381
[S049] Forecasting: Principles and Practice	381
[S050] CS336: Language Modeling from Scratch.....	382
[S051] SentencePiece	383
[S052] Language Models are Few-Shot Learners.....	383
[S053] Training Compute-Optimal Large Language Models	384
[S054] Training language models to follow instructions with human feedback	384
[S055] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.....	385
[S056] Learning Transferable Visual Models From Natural Language Supervision	386
[S057] World Models	386
[S058] ReAct: Synergizing Reasoning and Acting in Language Models	387
[S059] Toolformer	387

[S060] AgentBench.....	388
[S061] Fully Sharded Data Parallel	389
[S062] Megatron-LM.....	389
[S063] ZeRO.....	390
[S064] FlashAttention	390
[S065] Efficient Memory Management for Large Language Model Serving with PagedAttention.....	391
[S066] MLflow Documentation	391
[S067] Kubernetes Documentation	392
[S068] Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.....	393
[S069] Regulation (EU) 2024/1689 - Artificial Intelligence Act.....	393
[S070] OECD AI Principles	394
[S071] Model Cards for Model Reporting.....	394
[S072] Datasheets for Datasets.....	395
[S073] Why Should I Trust You?	395
[S074] A Unified Approach to Interpreting Model Predictions	396
[S075] The Algorithmic Foundations of Differential Privacy	397
[S076] Communication-Efficient Learning of Deep Networks from Decentralized Data	397
[S077] Membership Inference Attacks Against Machine Learning Models.....	398
[S078] Highly accurate protein structure prediction with AlphaFold	398
[S079] Learning skillful medium-range global weather forecasting.....	399

[S080] Artifact Review and Badging.....400

[S081] About Distill.....400

[S084] OWASP Top 10 for LLM Applications.....401

[S085] MITRE ATLAS401

[S086] Recommendation on the Ethics of Artificial Intelligence402

全书术语表403

版权、版本与勘误.....413

本卷 24 章路线图

月 19 · 图学习、推荐与时序智能

理解节点、边、属性、邻接、置换不变和消息传递。

1. [第 73 章 图表示、消息传递与 GNN — 理解节点、边、属性、邻接、置换不变和消息传递。](#)
2. [第 74 章 推荐、检索与排序系统 — 覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。](#)
3. [第 75 章 时间序列、预测与回测 — 理解趋势、季节、滞后、平稳性、预测区间和时间切分。](#)
4. [第 76 章 时空建模、异常与决策智能 — 结合位置、时间、网络和行动，避免把预测直接当决策。](#)

月 20 · 基础模型和语言模型从零实现

实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。

5. [第 77 章 语言模型词元器与嵌入 — 实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。](#)
6. [第 78 章 从零实现因果 Transformer — 实现归一化、多头因果注意力、前馈层、残差和输出头。](#)
7. [第 79 章 预训练目标、数据与缩放 — 理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。](#)
8. [第 80 章 指令微调、偏好学习、RAG 与评估 — 区分预训练、监督微调、偏好优化、检索增强和系统评估。](#)

月 21 · 多模态、世界模型与 Agent

理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。

9. [第 81 章 多模态表示、对齐与融合 — 理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。](#)
10. [第 82 章 世界模型、视频与行动预测 — 学习潜在动力学、未来预测、规划和模型误差累积。](#)

11. [第 83 章 Agent 循环、工具、记忆与规划 — 把观察、计划、调用、验证和停止条件做成显式状态机。](#)
12. [第 84 章 Agent 评估、安全与人在回路 — 测试任务成功、引用、权限、提示注入、工具故障和可恢复性。](#)

月 22 · 数据、规模化训练、推理和 MLOps

建立来源、许可、去重、污染、质量、退出和删除流程。

13. [第 85 章 数据治理、去重、质量与许可 — 建立来源、许可、去重、污染、质量、退出和删除流程。](#)
14. [第 86 章 分布式训练、并行与检查点 — 理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。](#)
15. [第 87 章 推理、量化、蒸馏与服务 — 理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。](#)
16. [第 88 章 MLOps、监控、可靠性、成本与碳 — 连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。](#)

月 23 · 安全、隐私、可解释性与治理

从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。

17. [第 89 章 安全、滥用与威胁建模 — 从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。](#)
18. [第 90 章 隐私、差分隐私、联邦学习与模型安全 — 理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。](#)
19. [第 91 章 可解释性、审计与模型文档 — 区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。](#)
20. [第 92 章 治理、法律、标准与事故响应 — 把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。](#)

月 24 · AI for Science、论文复现与原创毕业项目

理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。

21. [第 93 章 AI for Science：从预测到发现 — 理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。](#)
22. [第 94 章 读论文、复现实验与统计核查 — 从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。](#)
23. [第 95 章 研究问题、创新、消融与沟通 — 把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。](#)
24. [第 96 章 毕业项目：部署、答辩与前沿追踪 — 整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。](#)

第 73 章 图表示、消息传递与 GNN

Graphs and GNNs

项目	内容
学习单元	月 19 · 图学习、推荐与时序智能
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 72 章 机器人学习、仿真到现实与人身安全](#)

结论先行：理解节点、边、属性、邻接、置换不变和消息传递。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“图表示、消息传递与 GNN”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解节点、边、属性、邻接、置换不变和消息传递。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 robot-learning-sim2real。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：同学关系网中每个人从邻居收集信息，再更新自己的表示。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 73-1 图表示、消息传递与 GNN：单变量实验（教学模拟）

项目	内容
只改变量	消息传播跳数，范围 1-10 跳，步长 1
实验目标	只改变“消息传播跳数”，观察结果、代价和风险如何一起变化，理解理解节点、边、属性、邻接、置换不变和消息传递。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把消息传播跳数分别设为 1 跳、2 跳和 10 跳，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“消息传播跳数”，观察结果、代价和风险如何一起变化，理解理解节点、边、属性、邻接、置换不变和消息传递。

静态替代说明：静态替代：把消息传播跳数分别设为 1 跳、2 跳和 10 跳，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $h_v^{k+1} = \text{UPDATE}(h_v^k, \text{AGG}(\{h_u^k : u \in N(v)\}))$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 73-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
```

```
optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“图表示、消息传递与 GNN”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解节点、边、属性、邻接、置换不变和消息传递。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。同学关系网中每个人从邻居收集信息，再更新自己的表示。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消息传播跳数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $h_v^{k+1} = \text{UPDATE}(h_v^k, \text{AGG}(\{h_u^k: u \in N(v)\}))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么同学关系网中每个人从邻居收集信息，再更新自己的表示。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当消息传播跳数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“图表示、消息传递与 GNN”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解节点、边、属性、邻接、置换不变和消息传递。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。同学关系网中每个人从邻居收集信息，再更新自己的表示。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消息传播跳数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、

边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h_v^{k+1} = \text{UPDATE}(h_v^k, \text{AGG}(\{h_u^k: u \in N(v)\}))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么同学关系网中每个人从邻居收集信息，再更新自己的表示。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当消息传播跳数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“图表示、消息传递与 GNN”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解节点、边、属性、邻接、置换不变和消息传递。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。同学关系网中每个人从邻居收集信息，再更新自己的表示。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消息传播跳数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h_v^{k+1} = \text{UPDATE}(h_v^k, \text{AGG}(\{h_u^k: u \in N(v)\}))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么同学关系网中每个人从邻居收集信息，再更新自己的表示。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当消息传播跳数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“图表示、消息传递与 GNN”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解节点、边、属性、邻接、置换不变和消息传递。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。同学关系网中每个人从邻居收集信息，再更新自己的表示。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，

最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消息传播跳数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h_v^{k+1} = \text{UPDATE}(h_v^k, \text{AGG}(\{h_u^k: u \in N(v)\}))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么同学关系网中每个人从邻居收集信息，再更新自己的表示。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当消息传播跳数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“图表示、消息传递与 GNN”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解节点、边、属性、邻接、置换不变和消息传递。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。同学关系网中每个人从邻居收集信息，再更新自己的表示。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消息传播跳数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h_v^{k+1} = \text{UPDATE}(h_v^k, \text{AGG}(\{h_u^k: u \in N(v)\}))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么同学关系网中每个人从邻居收集信息，再更新自己的表示。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当消息传播跳数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存

`state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“消息传播跳数”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

25. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
26. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
27. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
28. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
29. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
30. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
31. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择不含个人信息的教学模拟天气记录作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开天文观测的缩小样例，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

32. **基础题：**用自己的话解释“图表示、消息传递与 GNN”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

33. **应用题：**在合成的植物叶片特征中设计一个只改变“消息传播跳数”的实验，列出基线、三档设置、指标和失败样例。

34. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/graphs-and-gnn.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。

- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S046] Stanford University, 《CS224W: Machine Learning with Graphs》, [访问 web.stanford.edu](https://web.stanford.edu))
- [S047] Thomas Kipf, Max Welling, 《Semi-Supervised Classification with Graph Convolutional Networks》, [访问 arxiv.org](https://arxiv.org))
- [S048] Yehuda Koren, Robert Bell, Chris Volinsky, 《Matrix Factorization Techniques for Recommender Systems》, [访问 doi.org](https://doi.org))
- [S049] Rob J Hyndman, George Athanasopoulos, 《Forecasting: Principles and Practice》, [访问 otexts.com](https://otexts.com))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 74 章 推荐、检索与排序系统

Recommender and ranking systems

项目	内容
学习单元	月 19 · 图学习、推荐与时序智能
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 73 章 图表示、消息传递与 GNN](#)

结论先行：覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“推荐、检索与排序系统”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `graphs-and-gnn`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 74-1 推荐、检索与排序系统：单变量实验（教学模拟）

项目	内容
只改变量	多样性权重，范围 0–1，步长 0.05
实验目标	只改变“多样性权重”，观察结果、代价和风险如何一起变化，理解覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把多样性权重分别设为 0、0.3 和 1，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“多样性权重”，观察结果、代价和风险如何一起变化，理解覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。

静态替代说明：静态替代：把多样性权重分别设为 0、0.3 和 1，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{score}(u, i) = p_u r_{u,i}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。

若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 74-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“推荐、检索与排序系统”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“多样性权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{score}(u,i)=p_u^T q_i$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、

空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当多样性权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“推荐、检索与排序系统”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“多样性权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、

边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{score}(u,i)=p_u^T q_i$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当多样性权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“推荐、检索与排序系统”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设

是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“多样性权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{score}(u,i)=p_u^T q_i$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当多样性权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“推荐、检索与排序系统”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“多样性权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{score}(u,i)=p_u^T q_i$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当多样性权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“推荐、检索与排序系统”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。。接着逐项标记观察到的事实、由公式推导的结果、

来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“多样性权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{score}(u,i)=p_u^T q_i$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当多样性权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“多样性权重”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

35. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

36. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

37. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

38. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

39. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

40. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

41. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开道路网络与虚构出行任务作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成音频波形与环境声标签，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

42. **基础题：** 用自己的话解释“推荐、检索与排序系统”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

43. **应用题：** 在虚构游戏角色的状态日志中设计一个只改变“多样性权重”的实验，列出基线、三档设置、指标和失败样例。

44. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/recommender-ranking.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S046] Stanford University, 《CS224W: Machine Learning with Graphs》，[访问 web.stanford.edu](https://web.stanford.edu))
- [S047] Thomas Kipf, Max Welling, 《Semi-Supervised Classification with Graph Convolutional Networks》，[访问 arxiv.org](https://arxiv.org))
- [S048] Yehuda Koren, Robert Bell, Chris Volinsky, 《Matrix Factorization Techniques for Recommender Systems》，[访问 doi.org](https://doi.org))
- [S049] Rob J Hyndman, George Athanasopoulos, 《Forecasting: Principles and Practice》，[访问 otexts.com](https://otexts.com))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 75 章 时间序列、预测与回测

Time series forecasting

项目	内容
学习单元	月 19 · 图学习、推荐与时序智能
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 74 章 推荐、检索与排序系统](#)

结论先行：理解趋势、季节、滞后、平稳性、预测区间和时间切分。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“时间序列、预测与回测”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解趋势、季节、滞后、平稳性、预测区间和时间切分。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 recommender-ranking。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 75-1 时间序列、预测与回测：单变量实验（教学模拟）

项目	内容
只改变量	回看窗口，范围 1-365 天，步长 1
实验目标	只改变“回看窗口”，观察结果、代价和风险如何一起变化，理解理解趋势、季节、滞后、平稳性、预测区间和时间切分。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把回看窗口分别设为 1 天、30 天和 365 天，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“回看窗口”，观察结果、代价和风险如何一起变化，理解理解趋势、季节、滞后、平稳性、预测区间和时间切分。

静态替代说明：静态替代：把回看窗口分别设为 1 天、30 天和 365 天，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $y_t = \text{trend}_t + \text{season}_t + \text{residual}_t$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 75-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
```

```
optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“时间序列、预测与回测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解趋势、季节、滞后、平稳性、预测区间和时间切分。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“回看窗口”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $y_t = \text{trend}_t + \text{season}_t + \text{residual}_t$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当回看窗口变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“时间序列、预测与回测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解趋势、季节、滞后、平稳性、预测区间和时间切分。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“回看窗口”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $y_t = \text{trend}_t + \text{season}_t + \text{residual}_t$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当回看窗口变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“时间序列、预测与回测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解趋势、季节、滞后、平稳性、预测区间和时间切分。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“回看窗口”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $y_t = \text{trend}_t + \text{season}_t + \text{residual}_t$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当回看窗口变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“时间序列、预测与回测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解趋势、季节、滞后、平稳性、预测区间和时间切分。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均

值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“回看窗口”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $y_t = \text{trend}_t + \text{season}_t + \text{residual}_t$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当回看窗口变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“时间序列、预测与回测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解趋势、季节、滞后、平稳性、预测区间和时间切分。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“回看窗口”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $y_t = \text{trend}_t + \text{season}_t + \text{residual}_t$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当回看窗口变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()`和

`optimizer.step()`完成更新。评估阶段使用 `model.eval()`与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“回看窗口”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

45. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
46. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
47. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
48. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
49. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
50. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
51. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成的植物叶片特征作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择校园能源使用的匿名汇总数据，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

52. **基础题：**用自己的话解释“时间序列、预测与回测”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

53. **应用题：**在公开天文观测的缩小样例中设计一个只改变“回看窗口”的实验，列出基线、三档设置、指标和失败样例。

54. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/time-series-forecasting.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。

- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S046] Stanford University, 《CS224W: Machine Learning with Graphs》, [访问 web.stanford.edu](https://web.stanford.edu))
- [S047] Thomas Kipf, Max Welling, 《Semi-Supervised Classification with Graph Convolutional Networks》, [访问 arxiv.org](https://arxiv.org))
- [S048] Yehuda Koren, Robert Bell, Chris Volinsky, 《Matrix Factorization Techniques for Recommender Systems》, [访问 doi.org](https://doi.org))
- [S049] Rob J Hyndman, George Athanasopoulos, 《Forecasting: Principles and Practice》, [访问 otexts.com](https://otexts.com))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 76 章 时空建模、异常与决策智能

Spatiotemporal intelligence

项目	内容
学习单元	月 19 · 图学习、推荐与时序智能
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 75 章 时间序列、预测与回测](#)

结论先行：结合位置、时间、网络和行动，避免把预测直接当决策。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“时空建模、异常与决策智能”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：结合位置、时间、网络和行动，避免把预测直接当决策。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 time-series-forecasting。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 76-1 时空建模、异常与决策智能：单变量实验（教学模拟）

项目	内容
只改变量	空间邻域，范围 1–100km，步长 1
实验目标	只改变“空间邻域”，观察结果、代价和风险如何一起变化，理解结合位置、时间、网络和行动，避免把预测直接当决策。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把空间邻域分别设为 1km、10km 和 100km，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“空间邻域”，观察结果、代价和风险如何一起变化，理解结合位置、时间、网络和行动，避免把预测直接当决策。

静态替代说明：静态替代：把空间邻域分别设为 1km、10km 和 100km，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{decision} = \text{forecast} + \text{constraints} + \text{utility} + \text{uncertainty}$ 。先不要急着推公式。

逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 76-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
```

```
optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“时空建模、异常与决策智能”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：结合位置、时间、网络和行动，避免把预测直接当决策。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“空间邻域”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $\text{decision} = \text{forecast} + \text{constraints} + \text{utility} + \text{uncertainty}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当空间邻域变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“时空建模、异常与决策智能”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：结合位置、时间、网络和行动，避免把预测直接当决策。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“空间邻域”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{decision} = \text{forecast} + \text{constraints} + \text{utility} + \text{uncertainty}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当空间邻域变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“时空建模、异常与决策智能”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：结合位置、时间、网络和行动，避免把预测直接当决策。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，

最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。

若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“空间邻域”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{decision} = \text{forecast} + \text{constraints} + \text{utility} + \text{uncertainty}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当空间邻域变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“时空建模、异常与决策智能”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：结合位置、时间、网络和行动，避免把预测直接当决策。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“空间邻域”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{decision} = \text{forecast} + \text{constraints} + \text{utility} + \text{uncertainty}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当空间邻域变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“时空建模、异常与决策智能”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：结合位置、时间、网络和行动，避免把预测直接当决策。。接着逐项标记观察到的事实、由公式推导的结果、

来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“空间邻域”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{decision} = \text{forecast} + \text{constraints} + \text{utility} + \text{uncertainty}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当空间邻域变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“空间邻域”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

55. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

56. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

57. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

58. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

59. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

60. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

61. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择虚构游戏角色的状态日志作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开古诗文本与版权已明确的材料，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

62. **基础题：** 用自己的话解释“时空建模、异常与决策智能”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

63. **应用题：** 在合成音频波形与环境声标签中设计一个只改变“空间邻域”的实验，列出基线、三档设置、指标和失败样例。

64. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/spatiotemporal-anomaly.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S046] Stanford University, 《CS224W: Machine Learning with Graphs》, [访问 web.stanford.edu](https://web.stanford.edu))
- [S047] Thomas Kipf, Max Welling, 《Semi-Supervised Classification with Graph Convolutional Networks》, [访问 arxiv.org](https://arxiv.org))
- [S048] Yehuda Koren, Robert Bell, Chris Volinsky, 《Matrix Factorization Techniques for Recommender Systems》, [访问 doi.org](https://doi.org))
- [S049] Rob J Hyndman, George Athanasopoulos, 《Forecasting: Principles and Practice》, [访问 otexts.com](https://otexts.com))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

月 19 阶段项目：构建一个不跟踪个人的内容发现原型

项目结论： 用公开项目或合成事件做图、推荐和时序分析，提供多样性、冷启动和退出机制。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

65. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
66. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
67. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
68. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 77 章 语言模型词元器与嵌入

LM tokenizers and embeddings

项目	内容
学习单元	月 20 · 基础模型和语言模型从零实现
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 76 章 时空建模、异常与决策智能](#)

结论先行：实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“语言模型词元器与嵌入”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条

件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 spatiotemporal-anomaly。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 77-1 语言模型词元器与嵌入：单变量实验（教学模拟）

项目	内容
只改变量	词表大小，范围 100–32000，步长 100
实验目标	只改变“词表大小”，观察结果、代价和风险如何一起变化，理解实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把词表大小分别设为 100、8000 和 32000，手工比较三组教学模拟输出。

项目	内容
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“词表大小”，观察结果、代价和风险如何一起变化，理解实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。

静态替代说明：静态替代：把词表大小分别设为 100、8000 和 32000，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`sequence_length ≈ tokens(text, vocabulary)`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 77-1 python 最小实现

```
import torch
torch.manual_seed(20260806)
```

```

x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“语言模型词元器与嵌入”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“词表大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{sequence_length} \approx \text{tokens}(\text{text}, \text{vocabulary})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当词表大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“语言模型词元器与嵌入”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“词表大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{sequence_length} \approx \text{tokens}(\text{text}, \text{vocabulary})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当词表大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“语言模型词元器与嵌入”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现子词切分、词表、位置 and 输入输出嵌入，分析不同语言的长度差异。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如

固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“词表大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{sequence_length} \approx \text{tokens}(\text{text}, \text{vocabulary})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当词表大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“语言模型词元器与嵌入”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能

力。积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“词表大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{sequence_length} \approx \text{tokens}(\text{text}, \text{vocabulary})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当词表大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“语言模型词元器与嵌入”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现子词切

分、词表、位置和输入输出嵌入，分析不同语言的长度差异。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“词表大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{sequence_length} \approx \text{tokens}(\text{text}, \text{vocabulary})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当词表大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“词表大小”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

69. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

70. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

71. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

72. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

73. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

74. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

75. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开天文观测的缩小样例作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择虚构机器人在二维网格中的轨迹，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

76. **基础题：** 用自己的话解释“语言模型词元器与嵌入”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

77. **应用题：** 在校园能源使用的匿名汇总数据中设计一个只改变“词表大小”的实验，列出基线、三档设置、指标和失败样例。

78. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/llm-tokenizer-embeddings.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S030] Ashish Vaswani et al., 《Attention Is All You Need》, [访问 arxiv.org](#))
- [S050] Stanford University, 《CS336: Language Modeling from Scratch》, [访问 cs336.stanford.edu](#))
- [S051] Taku Kudo, John Richardson, 《SentencePiece》, [访问 arxiv.org](#))
- [S052] Tom Brown et al., 《Language Models are Few-Shot Learners》, [访问 arxiv.org](#))
- [S053] Jordan Hoffmann et al., 《Training Compute-Optimal Large Language Models》, [访问 arxiv.org](#))
- [S054] Long Ouyang et al., 《Training language models to follow instructions with human feedback》, [访问 arxiv.org](#))

- [S055] Patrick Lewis et al., 《Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks》, [访问 arxiv.org](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 78 章 从零实现因果 Transformer

Transformer from scratch

项目	内容
学习单元	月 20 · 基础模型和语言模型从零实现
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 77 章 语言模型词元器与嵌入](#)

结论先行：实现归一化、多头因果注意力、前馈层、残差和输出头。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“从零实现因果 Transformer”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：实现归一化、多头因果注意力、前馈层、残差和输出头。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 llm-tokenizer-embeddings。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 78-1 从零实现因果 Transformer：单变量实验（教学模拟）

项目	内容
只改变量	上下文长度，范围 8–4096token，步长 8
实验目标	只改变“上下文长度”，观察结果、代价和风险如何一起变化，理解实现归一化、多头因果注意力、前馈层、残差和输出头。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把上下文长度分别设为 8token、256token 和 4096token，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“上下文长度”，观察结果、代价和风险如何一起变化，理解实现归一化、多头因果注意力、前馈层、残差和输出头。

静态替代说明：静态替代：把上下文长度分别设为 8token、256token 和 4096token，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $P(x) = \prod_t P(x_t | x_{<t})$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 78-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
```

```
optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“从零实现因果 Transformer”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现归一化、多头因果注意力、前馈层、残差和输出头。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“上下文长度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $P(x)=\prod_t P(x_t|x_{<t})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当上下文长度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“从零实现因果Transformer”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现归一化、多头因果注意力、前馈层、残差和输出头。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“上下文长度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、

边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x) = \prod_t P(x_t | x_{\{<t\}})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当上下文长度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“从零实现因果 Transformer”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现归一化、多头因果注意力、前馈层、残差和输出头。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低

可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。

若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“上下文长度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x) = \prod_t P(x_t | x_{\leq t})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当上下文长度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“从零实现因果 Transformer”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现归一化、多头因果注意力、前馈层、残差和输出头。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“上下文长度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x) = \prod_t P(x_t | x_{\{<t\}})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当上下文长度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“从零实现因果Transformer”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：实现归一化、多头因果注意力、前馈层、残差和输出头。。接着逐项标记观察到的事实、由公式推导的

结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“上下文长度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x) = \prod_t P(x_t | x_{<t})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当上下文长度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“上下文长度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

79. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

80. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

81. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

82. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

83. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

84. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

85. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成音频波形与环境声标签作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开科研论文的标题与摘要元数据，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

86. **基础题：** 用自己的话解释“从零实现因果 Transformer”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

87. **应用题：** 在公开古诗文本与版权已明确的材料中设计一个只改变“上下文长度”的实验，列出基线、三档设置、指标和失败样例。

88. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/transformer-from-scratch.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S030] Ashish Vaswani et al., 《Attention Is All You Need》, [访问 arxiv.org](#))
- [S050] Stanford University, 《CS336: Language Modeling from Scratch》, [访问 cs336.stanford.edu](#))
- [S051] Taku Kudo, John Richardson, 《SentencePiece》, [访问 arxiv.org](#))
- [S052] Tom Brown et al., 《Language Models are Few-Shot Learners》, [访问 arxiv.org](#))
- [S053] Jordan Hoffmann et al., 《Training Compute-Optimal Large Language Models》, [访问 arxiv.org](#))
- [S054] Long Ouyang et al., 《Training language models to follow instructions with human feedback》, [访问 arxiv.org](#))

- [S055] Patrick Lewis et al., 《Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks》, [访问 arxiv.org](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 79 章 预训练目标、数据与缩放

Pretraining, data and scaling

项目	内容
学习单元	月 20 · 基础模型和语言模型从零实现
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 78 章 从零实现因果 Transformer](#)

结论先行：理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“预训练目标、数据与缩放”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条

件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 transformer-from-scratch。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：增加练习量有效，但重复同一道题不会等价于增加新知识。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 79-1 预训练目标、数据与缩放：单变量实验（教学模拟）

项目	内容
只改变量	训练词元，范围 1000–1000000token，步长 1000
实验目标	只改变“训练词元”，观察结果、代价和风险如何一起变化，理解理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把训练词元分别设为 1000token、100000token 和 1000000token，手工比较三组教学模拟输出。

项目	内容
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“训练词元”，观察结果、代价和风险如何一起变化，理解理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。

静态替代说明：静态替代：把训练词元分别设为 1000token、100000token 和 1000000token，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`training_compute ≈ parameters × tokens × constant`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 79-1 python 最小实现

```
import torch
```

```

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化管理

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“预训练目标、数据与缩放”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。增加练习量有效，但重复同一道题不会等价于增加新知识。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“训练词元”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{training_compute} \approx \text{parameters} \times \text{tokens} \times \text{constant}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么增加练习量有效，但重复同一道题不会等价于增加新知识。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当训练词元变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“预训练目标、数据与缩放”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。增加练习量有效，但重复同一道题不会等价于增加新知识。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如

固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“训练词元”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{training_compute} \approx \text{parameters} \times \text{tokens} \times \text{constant}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么增加练习量有效，但重复同一道题不会等价于增加新知识。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当训练词元变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“预训练目标、数据与缩放”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能

力。增加练习量有效，但重复同一道题不会等价于增加新知识。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“训练词元”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{training_compute} \approx \text{parameters} \times \text{tokens} \times \text{constant}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么增加练习量有效，但重复同一道题不会等价于增加新知识。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当训练词元变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“预训练目标、数据与缩放”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。增加练习量有效，但重复同一道题不会等价于增加新知识。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“训练词元”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{training_compute} \approx \text{parameters} \times \text{tokens} \times \text{constant}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么增加练习量有效，但重复同一道题不会等价于增加新知识。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完

全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当训练词元变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“预训练目标、数据与缩放”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。增加练习量有效，但重复同一道题不会等价于增加新知识。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“训练词元”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{training_compute} \approx \text{parameters} \times \text{tokens} \times \text{constant}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么增加练习量有效，但重复同一道题不会等价于增加新知识。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当训练词元变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“训练词元”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，

记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

89. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

90. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

91. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

92. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

93. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

94. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

95. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择校园能源使用的匿名汇总数据作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成电商目录但不含真实用户行为，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

96. **基础题：** 用自己的话解释“预训练目标、数据与缩放”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。
97. **应用题：** 在虚构机器人在二维网格中的轨迹中设计一个只改变“训练词元”的实验，列出基线、三档设置、指标和失败样例。
98. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/pretraining-data-scaling.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S030] Ashish Vaswani et al., 《Attention Is All You Need》，[访问 arxiv.org](https://arxiv.org/abs/1706.03762)

- [S050] Stanford University, 《CS336: Language Modeling from Scratch》, [访问 cs336.stanford.edu](https://cs336.stanford.edu))
- [S051] Taku Kudo, John Richardson, 《SentencePiece》, [访问 arxiv.org](https://arxiv.org))
- [S052] Tom Brown et al., 《Language Models are Few-Shot Learners》, [访问 arxiv.org](https://arxiv.org))
- [S053] Jordan Hoffmann et al., 《Training Compute-Optimal Large Language Models》, [访问 arxiv.org](https://arxiv.org))
- [S054] Long Ouyang et al., 《Training language models to follow instructions with human feedback》, [访问 arxiv.org](https://arxiv.org))
- [S055] Patrick Lewis et al., 《Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks》, [访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 **draft** → **technical_review** → **fact_checked** → **stable** 前进。

第 80 章 指令微调、偏好学习、RAG 与评估

Post-training, RAG and evaluation

项目	内容
学习单元	月 20 · 基础模型和语言模型从零实现
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 79 章 预训练目标、数据与缩放](#)

结论先行： 区分预训练、监督微调、偏好优化、检索增强和系统评估。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“指令微调、偏好学习、RAG 与评估”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：区分预训练、监督微调、偏好优化、检索增强和系统评估。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 pretraining-data-scaling。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 80-1 指令微调、偏好学习、RAG 与评估：单变量实验（教学模拟）

项目	内容
只改变量	检索候选数，范围 1–30 条，步长 1
实验目标	只改变“检索候选数”，观察结果、代价和风险如何一起变化，理解区分预训练、监督微调、偏好优化、检索增强和系统评估。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把检索候选数分别设为 1 条、5 条和 30 条，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性

项目	内容
	能或现实世界因果效果。

实验目标：只改变“检索候选数”，观察结果、代价和风险如何一起变化，理解区分预训练、监督微调、偏好优化、检索增强和系统评估。

静态替代说明：静态替代：把检索候选数分别设为 1 条、5 条和 30 条，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`answer = model(prompt, retrieved_evidence, policy)`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 80-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
```

```

        torch.nn.Linear(4, 8),
        torch.nn.ReLU(),
        torch.nn.Linear(8, 1),
    )
    optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
    for _ in range(40):
        loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
    print({"loss": round(loss.item(), 4), "seed": 20260806})

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“指令微调、偏好学习、RAG 与评估”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分预训练、监督微调、偏好优化、检索增强和系统评估。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“检索候选数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、

边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `answer = model(prompt, retrieved_evidence, policy)`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当检索候选数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“指令微调、偏好学习、RAG 与评估”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分预训练、监督微调、偏好优化、检索增强和系统评估。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低

可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“检索候选数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `answer = model(prompt, retrieved_evidence, policy)`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当检索候选数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“指令微调、偏好学习、RAG 与评估”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分预训练、监督微调、偏好优化、检索增强和系统评估。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“检索候选数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `answer = model(prompt, retrieved_evidence, policy)`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当检索候选数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“指令微调、偏好学习、RAG与评估”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分预训练、监督微调、偏好优化、检索增强和系统评估。。接着逐项标记观察到的事实、由公式推导的结果、

来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“检索候选数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `answer = model(prompt, retrieved_evidence, policy)`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当检索候选数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“指令微调、偏好学习、RAG 与评估”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分预训练、监督微调、偏好优化、检索增强和系统评估。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“检索候选数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `answer = model(prompt, retrieved_evidence, policy)`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么基础模型像读过许多书的学生；指令

训练教作答方式，检索给现场资料，评估检查结果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当检索候选数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“检索候选数”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

99. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
100. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
101. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
102. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
103. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
104. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
105. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开古诗文本与版权已明确的材料作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择校园图书馆的公开借阅类别，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

106. **基础题：** 用自己的话解释“指令微调、偏好学习、RAG 与评估”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

107. **应用题：**在公开科研论文的标题与摘要元数据中设计一个只改变“检索候选数”的实验，列出基线、三档设置、指标和失败样例。

108. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/posttraining-rag-evaluation.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S030] Ashish Vaswani et al., 《Attention Is All You Need》，[访问 arxiv.org](https://arxiv.org/abs/1706.03762)
- [S050] Stanford University, 《CS336: Language Modeling from Scratch》，[访问 cs336.stanford.edu](https://cs336.stanford.edu)
- [S051] Taku Kudo, John Richardson, 《SentencePiece》，[访问 arxiv.org](https://arxiv.org/abs/1609.08147)
- [S052] Tom Brown et al., 《Language Models are Few-Shot Learners》，[访问 arxiv.org](https://arxiv.org/abs/2002.05207)

- [S053] Jordan Hoffmann et al., 《Training Compute-Optimal Large Language Models》, [访问 arxiv.org](#))
- [S054] Long Ouyang et al., 《Training language models to follow instructions with human feedback》, [访问 arxiv.org](#))
- [S055] Patrick Lewis et al., 《Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks》, [访问 arxiv.org](#))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 **draft** → **technical_review** → **fact_checked** → **stable** 前进。

月 20 阶段项目：从零训练一个微型语言模型

项目结论：在 CPU 小语料上实现词元器、因果注意力、训练、采样和评估；记录数据许可与算力边界。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

- 109. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
- 110. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
- 111. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
- 112. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 81 章 多模态表示、对齐与融合

Multimodal alignment and fusion

项目	内容
学习单元	月 21 · 多模态、世界模型与 Agent
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 80 章 指令微调、偏好学习、RAG 与评估](#)

结论先行：理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“多模态表示、对齐与融合”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 posttraining-rag-evaluation。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 81-1 多模态表示、对齐与融合：单变量实验（教学模拟）

项目	内容
只改变量	对齐温度，范围 0.01–1，步长 0.01
实验目标	只改变“对齐温度”，观察结果、代价和风险如何一起变化，理解理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把对齐温度分别设为 0.01、0.07 和 1，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“对齐温度”，观察结果、代价和风险如何一起变化，理解理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。

静态替代说明：静态替代：把对齐温度分别设为 0.01、0.07 和 1，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{loss} = -\log \frac{\exp(\text{sim}(i, t))}{\sum_j \exp(\text{sim}(i, t_j))}$ 。先不要急着推公式。逐个给符号

贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 81-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```

```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“多模态表示、对齐与融合”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“对齐温度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = -\log \exp(\text{sim}(i, t) / \tau) / \sum_j \exp(\text{sim}(i, t_j) / \tau)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当对齐温度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“多模态表示、对齐与融合”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“对齐温度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = -\log \exp(\text{sim}(i,t)/\tau) / \sum_j \exp(\text{sim}(i,t_j)/\tau)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当对齐温度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“多模态表示、对齐与融合”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均

值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“对齐温度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = -\log \exp(\text{sim}(i, t) / \tau) / \sum_j \exp(\text{sim}(i, t_j) / \tau)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当对齐温度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“多模态表示、对齐与融合”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能

力。给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“对齐温度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = -\log \exp(\text{sim}(i, t) / \tau) / \sum_j \exp(\text{sim}(i, t_j) / \tau)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当对齐温度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“多模态表示、对齐与融合”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“对齐温度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = -\log \exp(\text{sim}(i, t) / \tau) / \sum_j \exp(\text{sim}(i, t_j) / \tau)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。只是入口，并给出一个类比失效的地方。如果解释中出现“肯

定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当对齐温度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“对齐温度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 113. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 114. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 115. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 116. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 117. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 118. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 119. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择虚构机器人在二维网格中的轨迹作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择不含个人信息的教学模拟天气记录，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 120. **基础题：** 用自己的话解释“多模态表示、对齐与融合”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

121. **应用题：**在合成电商目录但不含真实用户行为中设计一个只改变“对齐温度”的实验，列出基线、三档设置、指标和失败样例。

122. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/multimodal-alignment.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S056] Alec Radford et al., 《Learning Transferable Visual Models From Natural Language Supervision》，[访问 arxiv.org](https://arxiv.org/abs/2109.01111))
- [S057] David Ha, Jürgen Schmidhuber, 《World Models》，[访问 arxiv.org](https://arxiv.org/abs/2109.01111))
- [S058] Shunyu Yao et al., 《ReAct: Synergizing Reasoning and Acting in Language Models》，[访问 arxiv.org](https://arxiv.org/abs/2210.03629))
- [S059] Timo Schick et al., 《Toolformer》，[访问 arxiv.org](https://arxiv.org/abs/2302.07549))

- [S060] Xiao Liu et al., 《AgentBench》, [访问 arxiv.org](#))
- [S084] OWASP Foundation, 《OWASP Top 10 for LLM Applications》, [访问 genai.owasp.org](#))
- [S085] MITRE, 《MITRE ATLAS》, [访问 atlas.mitre.org](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 82 章 世界模型、视频与行动预测

World models, video and action

项目	内容
学习单元	月 21 · 多模态、世界模型与 Agent
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 81 章 多模态表示、对齐与融合](#)

结论先行：学习潜在动力学、未来预测、规划和模型误差累积。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“世界模型、视频与行动预测”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：学习潜在动力学、未来预测、规划和模型误差累积。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 multimodal-alignment。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 82-1 世界模型、视频与行动预测：单变量实验（教学模拟）

项目	内容
只改变量	想象步数，范围 1–100 步，步长 1
实验目标	只改变“想象步数”，观察结果、代价和风险如何一起变化，理解学习潜在动力学、未来预测、规划和模型误差累积。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把想象步数分别设为 1 步、10 步和 100 步，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“想象步数”，观察结果、代价和风险如何一起变化，理解学习潜在动力学、未来预测、规划和模型误差累积。

静态替代说明：静态替代：把想象步数分别设为 1 步、10 步和 100 步，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $z_{t+1} = f(z_t, a_t) + \epsilon$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 82-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
    Gate("publish or write externally", "high", True),
]
for check in checks:
```

```
print(authorize(check))
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“世界模型、视频与行动预测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：学习潜在动力学、未来预测、规划和模型误差累积。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“想象步数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $z_{t+1}=f(z_t,a_t)+\varepsilon$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当想象步数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“世界模型、视频与行动预测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：学习潜在动力学、未来预测、规划和模型误差累积。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“想象步数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z_{t+1}=f(z_t, a_t)+\varepsilon$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当想象步数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“世界模型、视频与行动预测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：学习潜在动力学、未来预测、规划和模型误差累积。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“想象步数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z_{t+1}=f(z_t, a_t)+\epsilon$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当想象步数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“世界模型、视频与行动预测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：学习潜在动力学、未来预测、规划和模型误差累积。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，

最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。

若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“想象步数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z_{t+1}=f(z_t, a_t)+\epsilon$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当想象步数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“世界模型、视频与行动预测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：学习潜在动力学、未来预测、规划和模型误差累积。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“想象步数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z_{t+1}=f(z_t, a_t)+\epsilon$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当想象步数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存

`state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“想象步数”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

123. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
124. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
125. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
126. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
127. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
128. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
129. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开科研论文的标题与摘要元数据作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开道路网络与虚构出行任务，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

130. **基础题：**用自己的话解释“世界模型、视频与行动预测”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

131. **应用题：**在校园图书馆的公开借阅类别中设计一个只改变“想象步数”的实验，列出基线、三档设置、指标和失败样例。

132. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/world-models-video-action.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。

- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S056] Alec Radford et al., 《Learning Transferable Visual Models From Natural Language Supervision》，[访问 arxiv.org](#))
- [S057] David Ha, Jürgen Schmidhuber, 《World Models》，[访问 arxiv.org](#))
- [S058] Shunyu Yao et al., 《ReAct: Synergizing Reasoning and Acting in Language Models》，[访问 arxiv.org](#))
- [S059] Timo Schick et al., 《Toolformer》，[访问 arxiv.org](#))
- [S060] Xiao Liu et al., 《AgentBench》，[访问 arxiv.org](#))
- [S084] OWASP Foundation, 《OWASP Top 10 for LLM Applications》，[访问 genai.owasp.org](#))
- [S085] MITRE, 《MITRE ATLAS》，[访问 atlas.mitre.org](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 83 章 Agent 循环、工具、记忆与规划

Agent loops, tools and memory

项目	内容
学习单元	月 21 · 多模态、世界模型与 Agent
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 82 章 世界模型、视频与行动预测](#)

结论先行： 把观察、计划、调用、验证和停止条件做成显式状态机。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“Agent 循环、工具、记忆与规划”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：把观察、计划、调用、验证和停止条件做成显式状态机。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 world-models-video-action。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 83-1 Agent 循环、工具、记忆与规划：单变量实验（教学模拟）

项目	内容
只改变量	工具调用预算，范围 1-50 次，步长 1
实验目标	只改变“工具调用预算”，观察结果、代价和风险如何一起变化，理解把观察、计划、调用、验证和停止条件做成显式状态机。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把工具调用预算分别设为 1 次、8 次和 50 次，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“工具调用预算”，观察结果、代价和风险如何一起变化，理解把观察、计划、调用、验证和停止条件做成显式状态机。

静态替代说明：静态替代：把工具调用预算分别设为 1 次、8 次和 50 次，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：observe → plan → act → verify → stop。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 83-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```

```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“Agent 循环、工具、记忆与规划”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把观察、计划、调用、验证和停止条件做成显式状态机。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“工具调用预算”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 observe → plan → act → verify → stop，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当工具调用预算变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“Agent 循环、工具、记忆与规划”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把观察、计划、调用、验证和停止条件做成显式状态机。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“工具调用预算”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 observe → plan → act → verify → stop，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当工具调用预算变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“Agent 循环、工具、记忆与规划”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把观察、计划、调用、验证和停止条件做成显式状态机。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、

均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“工具调用预算”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 observe → plan → act → verify → stop，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当工具调用预算变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“Agent 循环、工具、记忆与规划”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把观察、计划、调用、验证和停止条件做成显式状态机。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。

项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“工具调用预算”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 observe → plan → act → verify → stop，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当工具调用预算变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“Agent 循环、工具、记忆与规划”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把观察、计划、调用、验证和停止条件做成显式状态机。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“工具调用预算”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 observe → plan → act → verify → stop，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。只是入口，并给出一个类比失效的地方。如果解释中出现

“肯定” “完全” “智能” “理解” 等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当工具调用预算变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“工具调用预算”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 133. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 134. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 135. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 136. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 137. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 138. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 139. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成电商目录但不含真实用户行为作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成的植物叶片特征，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 140. **基础题：** 用自己的话解释“Agent 循环、工具、记忆与规划”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

141. **应用题：** 在不含个人信息的教学模拟天气记录中设计一个只改变“工具调用预算”的实验，列出基线、三档设置、指标和失败样例。

142. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/agent-loop-tools-memory.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S056] Alec Radford et al., 《Learning Transferable Visual Models From Natural Language Supervision》，[访问 arxiv.org](https://arxiv.org/abs/2109.00859))
- [S057] David Ha, Jürgen Schmidhuber, 《World Models》，[访问 arxiv.org](https://arxiv.org/abs/2109.00859))
- [S058] Shunyu Yao et al., 《ReAct: Synergizing Reasoning and Acting in Language Models》，[访问 arxiv.org](https://arxiv.org/abs/2210.03629))
- [S059] Timo Schick et al., 《Toolformer》，[访问 arxiv.org](https://arxiv.org/abs/2302.07549))

- [S060] Xiao Liu et al., 《AgentBench》, [访问 arxiv.org](#))
- [S084] OWASP Foundation, 《OWASP Top 10 for LLM Applications》, [访问 genai.owasp.org](#))
- [S085] MITRE, 《MITRE ATLAS》, [访问 atlas.mitre.org](#))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 84 章 Agent 评估、安全与人在回路

Agent evaluation and security

项目	内容
学习单元	月 21 · 多模态、世界模型与 Agent
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 83 章 Agent 循环、工具、记忆与规划](#)

结论先行：测试任务成功、引用、权限、提示注入、工具故障和可恢复性。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“Agent 评估、安全与人在回路”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：测试任务成功、引用、权限、提示注入、工具故障和可恢复性。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 agent-loop-tools-memory。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 84-1 Agent 评估、安全与人在回路：单变量实验（教学模拟）

项目	内容
只改变量	自主级别，范围 0–5 级，步长 1
实验目标	只改变“自主级别”，观察结果、代价和风险如何一起变化，理解测试任务成功、引用、权限、提示注入、工具故障和可恢复性。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把自主级别分别设为 0 级、1 级和 5 级，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“自主级别”，观察结果、代价和风险如何一起变化，理解测试任务成功、引用、权限、提示注入、工具故障和可恢复性。

静态替代说明：静态替代：把自主级别分别设为 0 级、1 级和 5 级，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{risk} = \text{capability} \times \text{access} \times \text{autonomy} \times \text{consequence}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 84-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```

```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“Agent 评估、安全与人在回路”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试任务成功、引用、权限、提示注入、工具故障和可恢复性。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“自主级别”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{risk} = \text{capability} \times \text{access} \times \text{autonomy} \times \text{consequence}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当自主级别变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“Agent 评估、安全与人在回路”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试任务成功、引用、权限、提示注入、工具故障和可恢复性。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“自主级别”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{risk} = \text{capability} \times \text{access} \times \text{autonomy} \times \text{consequence}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当自主级别变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“Agent 评估、安全与人在回路”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试任务成功、引用、权限、提示注入、工具故障和可恢复性。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“自主级别”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{risk} = \text{capability} \times \text{access} \times \text{autonomy} \times \text{consequence}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当自主级别变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“Agent 评估、安全与人在回路”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试任务成功、引用、权限、提示注入、工具故障和可恢复性。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能

力。给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“自主级别”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{risk} = \text{capability} \times \text{access} \times \text{autonomy} \times \text{consequence}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当自主级别变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“Agent 评估、安全与人在回路”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试任务成功、引用、权限、提示注入、工具故障和可恢复性。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“自主级别”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{risk} = \text{capability} \times \text{access} \times \text{autonomy} \times \text{consequence}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。只是入口，并给出一个类比失效的地方。如果解释中出现

“肯定” “完全” “智能” “理解” 等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当自主级别变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“自主级别”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 143. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 144. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 145. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 146. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 147. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 148. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 149. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择校园图书馆的公开借阅类别作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择虚构游戏角色的状态日志，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 150. **基础题：** 用自己的话解释“Agent 评估、安全与人在回路”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

151. **应用题：** 在公开道路网络与虚构出行任务中设计一个只改变“自主级别”的实验，列出基线、三档设置、指标和失败样例。

152. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/agent-evaluation-security.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S056] Alec Radford et al., 《Learning Transferable Visual Models From Natural Language Supervision》，[访问 arxiv.org](https://arxiv.org/abs/2108.07258))
- [S057] David Ha, Jürgen Schmidhuber, 《World Models》，[访问 arxiv.org](https://arxiv.org/abs/2109.00859))
- [S058] Shunyu Yao et al., 《ReAct: Synergizing Reasoning and Acting in Language Models》，[访问 arxiv.org](https://arxiv.org/abs/2210.03629))
- [S059] Timo Schick et al., 《Toolformer》，[访问 arxiv.org](https://arxiv.org/abs/2302.01569))

- [S060] Xiao Liu et al., 《AgentBench》, [访问 arxiv.org](#))
- [S084] OWASP Foundation, 《OWASP Top 10 for LLM Applications》, [访问 genai.owasp.org](#))
- [S085] MITRE, 《MITRE ATLAS》, [访问 atlas.mitre.org](#))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

月 21 阶段项目：做一个只读、可审计的研究助理

Agent

项目结论：限制工具白名单，记录每次检索和引用，遇到发布、写入或高风险判断必须停下等待人工确认。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

- 153. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
- 154. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
- 155. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
- 156. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 85 章 数据治理、去重、质量与许可

Data curation and licensing

项目	内容
学习单元	月 22 · 数据、规模化训练、推理和 MLOps
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 84 章 Agent 评估、安全与人在回路](#)

结论先行： 建立来源、许可、去重、污染、质量、退出和删除流程。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“数据治理、去重、质量与许可”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：建立来源、许可、去重、污染、质量、退出和删除流程。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 agent-evaluation-security。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 85-1 数据治理、去重、质量与许可：单变量实验（教学模拟）

项目	内容
只改变量	相似阈值，范围 0–1，步长 0.05
实验目标	只改变“相似阈值”，观察结果、代价和风险如何一起变化，理解建立来源、许可、去重、污染、质量、退出和删除流程。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把相似阈值分别设为 0、0.8 和 1，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“相似阈值”，观察结果、代价和风险如何一起变化，理解建立来源、许可、去重、污染、质量、退出和删除流程。

静态替代说明：静态替代：把相似阈值分别设为 0、0.8 和 1，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{effective_data} = \text{diverse_valid_tokens} - \text{duplicates} - \text{contamination}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 85-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```



```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“数据治理、去重、质量与许可”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建立来源、许可、去重、污染、质量、退出和删除流程。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“相似阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{effective_data} = \text{diverse_valid_tokens} - \text{duplicates} - \text{contamination}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当相似阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“数据治理、去重、质量与许可”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建立来源、许可、去重、污染、质量、退出和删除流程。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“相似阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{effective_data} = \text{diverse_valid_tokens} - \text{duplicates} - \text{contamination}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当相似阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“数据治理、去重、质量与许可”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建立来源、许可、去重、污染、质量、退出和删除流程。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“相似阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{effective_data} = \text{diverse_valid_tokens} - \text{duplicates} - \text{contamination}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当相似阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“数据治理、去重、质量与许可”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建立来源、许可、去重、污染、质量、退出和删除流程。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。

图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“相似阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{effective_data} = \text{diverse_valid_tokens} - \text{duplicates} - \text{contamination}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当相似阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“数据治理、去重、质量与许可”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建立来源、许可、去重、污染、质量、退出和删除流程。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“相似阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{effective_data} = \text{diverse_valid_tokens} - \text{duplicates} - \text{contamination}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完

全” “智能” “理解” 等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当相似阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“相似阈值”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 157. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 158. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 159. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 160. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 161. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 162. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 163. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择不含个人信息的教学模拟天气记录作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开天文观测的缩小样例，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 164. **基础题：** 用自己的话解释“数据治理、去重、质量与许可”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

165. **应用题：**在合成的植物叶片特征中设计一个只改变“相似阈值”的实验，列出基线、三档设置、指标和失败样例。

166. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/data-curation-dedup-license.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S050] Stanford University, 《CS336: Language Modeling from Scratch》，[访问 cs336.stanford.edu](https://cs336.stanford.edu))
- [S061] PyTorch, 《Fully Sharded Data Parallel》，[访问 docs.pytorch.org](https://docs.pytorch.org))
- [S062] Mohammad Shoeybi et al., 《Megatron-LM》，[访问 arxiv.org](https://arxiv.org))
- [S063] Samyam Rajbhandari et al., 《ZeRO》，[访问 arxiv.org](https://arxiv.org))
- [S064] Tri Dao et al., 《FlashAttention》，[访问 arxiv.org](https://arxiv.org))

- [S065] Woosuk Kwon et al., 《Efficient Memory Management for Large Language Model Serving with PagedAttention》, [访问 arxiv.org](#))
- [S066] MLflow Project, 《MLflow Documentation》, [访问 mlflow.org](#))
- [S067] Kubernetes, 《Kubernetes Documentation》, [访问 kubernetes.io](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 86 章 分布式训练、并行与检查点

Distributed training

项目	内容
学习单元	月 22 · 数据、规模化训练、推理和 MLOps
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 85 章 数据治理、去重、质量与许可](#)

结论先行：理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“分布式训练、并行与检查点”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 data-curation-dedup-license。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 86-1 分布式训练、并行与检查点：单变量实验（教学模拟）

项目	内容
只改变量	并行工作者，范围 1-128 个，步长 1
实验目标	只改变“并行工作者”，观察结果、代价和风险如何一起变化，理解理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把并行工作者分别设为 1 个、8 个和 128 个，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“并行工作者”，观察结果、代价和风险如何一起变化，理解理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。

静态替代说明：静态替代：把并行工作者分别设为 1 个、8 个和 128 个，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{step_time} \approx \text{compute} + \text{communication} + \text{imbalance}$ 。先不要急着推公式。逐个给符号

贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 86-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```

```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“分布式训练、并行与检查点”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“并行工作者”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{step_time} \approx \text{compute} + \text{communication} + \text{imbalance}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当并行工作者变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“分布式训练、并行与检查点”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“并行工作者”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{step_time} \approx \text{compute} + \text{communication} + \text{imbalance}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当并行工作者变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“分布式训练、并行与检查点”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如

固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“并行工作者”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{step_time} \approx \text{compute} + \text{communication} + \text{imbalance}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当并行工作者变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“分布式训练、并行与检查点”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能

力。多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“并行工作者”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{step_time} \approx \text{compute} + \text{communication} + \text{imbalance}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当并行工作者变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“分布式训练、并行与检查点”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“并行工作者”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{step_time} \approx \text{compute} + \text{communication} + \text{imbalance}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多人搬家可以按箱子分工、按家具拆

件或按楼层流水线，但协调成本不同。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当并行工作者变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“并行工作者”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 167. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 168. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 169. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 170. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 171. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 172. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 173. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开道路网络与虚构出行任务作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成音频波形与环境声标签，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 174. **基础题：** 用自己的话解释“分布式训练、并行与检查点”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

175. **应用题：**在虚构游戏角色的状态日志中设计一个只改变“并行工作者”的实验，列出基线、三档设置、指标和失败样例。

176. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/distributed-training.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S050] Stanford University, 《CS336: Language Modeling from Scratch》，[访问 cs336.stanford.edu](https://cs336.stanford.edu))
- [S061] PyTorch, 《Fully Sharded Data Parallel》，[访问 docs.pytorch.org](https://docs.pytorch.org))
- [S062] Mohammad Shoeybi et al., 《Megatron-LM》，[访问 arxiv.org](https://arxiv.org))
- [S063] Samyam Rajbhandari et al., 《ZeRO》，[访问 arxiv.org](https://arxiv.org))
- [S064] Tri Dao et al., 《FlashAttention》，[访问 arxiv.org](https://arxiv.org))

- [S065] Woosuk Kwon et al., 《Efficient Memory Management for Large Language Model Serving with PagedAttention》, [访问 arxiv.org](#))
- [S066] MLflow Project, 《MLflow Documentation》, [访问 mlflow.org](#))
- [S067] Kubernetes, 《Kubernetes Documentation》, [访问 kubernetes.io](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 87 章 推理、量化、蒸馏与服务

Inference and serving

项目	内容
学习单元	月 22 · 数据、规模化训练、推理和 MLOps
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 86 章 分布式训练、并行与检查点](#)

结论先行：理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“推理、量化、蒸馏与服务”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 distributed-training。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。类比的 价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在 笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证 据。

3. 互动实验：只改变一个变量

实验 87-1 推理、量化、蒸馏与服务：单变量实验（教学模拟）

项目	内容
只改变量	量化位宽，范围 2–16bit，步长 1
实验目标	只改变“量化位宽”，观察结果、代价和风险如何一起变化，理解理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把量化位宽分别设为 2bit、8bit 和 16bit，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性

项目	内容
	能或现实世界因果效果。

实验目标：只改变“量化位宽”，观察结果、代价和风险如何一起变化，理解理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。

静态替代说明：静态替代：把量化位宽分别设为 2bit、8bit 和 16bit，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{throughput} \neq 1/\text{latency}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 87-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
```

```

    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“推理、量化、蒸馏与服务”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“量化位宽”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{throughput} \neq 1/\text{latency}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当量化位宽变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“推理、量化、蒸馏与服务”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，

最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“量化位宽”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{throughput} \neq 1/\text{latency}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当量化位宽变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“推理、量化、蒸馏与服务”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“量化位宽”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{throughput} \neq 1/\text{latency}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当量化位宽变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“推理、量化、蒸馏与服务”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。。接着逐项标记观察到的事实、由公式推导的结果、来

自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“量化位宽”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{throughput} \neq 1/\text{latency}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当量化位宽变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“推理、量化、蒸馏与服务”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“量化位宽”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{throughput} \neq 1/\text{latency}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么餐厅合并订单提高吞吐，但第一位顾

客可能等待更久；服务要平衡延迟和利用率。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当量化位宽变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“量化位宽”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 177. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 178. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 179. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 180. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 181. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 182. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 183. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成的植物叶片特征作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择校园能源使用的匿名汇总数据，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 184. **基础题：** 用自己的话解释“推理、量化、蒸馏与服务”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

185. **应用题：**在公开天文观测的缩小样例中设计一个只改变“量化位宽”的实验，列出基线、三档设置、指标和失败样例。

186. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/inference-compression-serving.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S050] Stanford University, 《CS336: Language Modeling from Scratch》，[访问 cs336.stanford.edu](https://cs336.stanford.edu))
- [S061] PyTorch, 《Fully Sharded Data Parallel》，[访问 docs.pytorch.org](https://docs.pytorch.org))
- [S062] Mohammad Shoeybi et al., 《Megatron-LM》，[访问 arxiv.org](https://arxiv.org))
- [S063] Samyam Rajbhandari et al., 《ZeRO》，[访问 arxiv.org](https://arxiv.org))
- [S064] Tri Dao et al., 《FlashAttention》，[访问 arxiv.org](https://arxiv.org))

- [S065] Woosuk Kwon et al., 《Efficient Memory Management for Large Language Model Serving with PagedAttention》, [访问 arxiv.org](#))
- [S066] MLflow Project, 《MLflow Documentation》, [访问 mlflow.org](#))
- [S067] Kubernetes, 《Kubernetes Documentation》, [访问 kubernetes.io](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 88 章 MLOps、监控、可靠性、成本与碳

MLOps and reliability

项目	内容
学习单元	月 22 · 数据、规模化训练、推理和 MLOps
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 87 章 推理、量化、蒸馏与服务](#)

结论先行： 连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。 **版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“MLOps、监控、可靠性、成本与碳”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署

条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 inference-compression-serving。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 88-1 MLOps、监控、可靠性、成本与碳：单变量实验（教学模拟）

项目	内容
只改变量	告警阈值，范围 0–100%，步长 5
实验目标	只改变“告警阈值”，观察结果、代价和风险如何一起变化，理解连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把告警阈值分别设为 0%、20%和 100%，手工

项目	内容
	比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“告警阈值”，观察结果、代价和风险如何一起变化，理解连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。

静态替代说明：静态替代：把告警阈值分别设为 0%、20%和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`total_cost = training + serving + people + failures`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 88-1 python 最小实现

```

from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“MLOps、监控、可靠性、成本与碳”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接

受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“告警阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{total_cost} = \text{training} + \text{serving} + \text{people} + \text{failures}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当告警阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“MLOps、监控、可靠性、成本与碳”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流

变化和备用方案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“告警阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{total_cost} = \text{training} + \text{serving} + \text{people} + \text{failures}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当告警阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“MLOps、监控、可靠性、成本与碳”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“告警阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{total_cost} = \text{training} + \text{serving} + \text{people} + \text{failures}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么上线模型像经营公交线路：不仅要发

车，还要看准点率、故障、客流变化和备用方案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当告警阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“MLOps、监控、可靠性、成本与碳”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“告警阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{total_cost} = \text{training} + \text{serving} + \text{people} + \text{failures}$ ，逐个说明符号的对象、形状、单

位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当告警阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“MLOps、监控、可靠性、成本与碳”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“告警阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{total_cost} = \text{training} + \text{serving} + \text{people} + \text{failures}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当告警阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“告警阈值”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

187. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
188. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
189. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
190. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
191. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
192. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
193. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择虚构游戏角色的状态日志作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开古诗文本与版权已明确的材料，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

194. **基础题：**用自己的话解释“MLOps、监控、可靠性、成本与碳”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

195. **应用题：**在合成音频波形与环境声标签中设计一个只改变“告警阈值”的实验，列出基线、三档设置、指标和失败样例。

196. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/mlops-monitoring-reliability.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。

- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S050] Stanford University, 《CS336: Language Modeling from Scratch》, [访问 cs336.stanford.edu](https://cs336.stanford.edu))
- [S061] PyTorch, 《Fully Sharded Data Parallel》, [访问 docs.pytorch.org](https://docs.pytorch.org))
- [S062] Mohammad Shoeybi et al., 《Megatron-LM》, [访问 arxiv.org](https://arxiv.org))
- [S063] Samyam Rajbhandari et al., 《ZeRO》, [访问 arxiv.org](https://arxiv.org))
- [S064] Tri Dao et al., 《FlashAttention》, [访问 arxiv.org](https://arxiv.org))
- [S065] Woosuk Kwon et al., 《Efficient Memory Management for Large Language Model Serving with PagedAttention》, [访问 arxiv.org](https://arxiv.org))
- [S066] MLflow Project, 《MLflow Documentation》, [访问 mlflow.org](https://mlflow.org))
- [S067] Kubernetes, 《Kubernetes Documentation》, [访问 kubernetes.io](https://kubernetes.io))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

月 22 阶段项目：部署一个有监控和回滚的小模型服务

项目结论： 建立数据版本、训练记录、离线评估、CPU 推理、延迟预算、漂移告警和回滚说明。
当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

- 197. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
- 198. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
- 199. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
- 200. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 89 章 安全、滥用与威胁建模

Safety and threat modeling

项目	内容
学习单元	月 23 · 安全、隐私、可解释性与治理
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 88 章 MLOps、监控、可靠性、成本与碳](#)

结论先行：从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“安全、滥用与威胁建模”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `mlops-monitoring-reliability`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。类比的价
值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔
记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 89-1 安全、滥用与威胁建模：单变量实验（教学模拟）

项目	内容
只改变量	控制措施强度，范围 0–100%，步长 5
实验目标	只改变“控制措施强度”，观察结果、代价和风险如何一起变化，理解从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把控制措施强度分别设为 0%、60%和 100%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“控制措施强度”，观察结果、代价和风险如何一起变化，理解从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。

静态替代说明：静态替代：把控制措施强度分别设为 0%、60%和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{residual_risk} = \text{inherent_risk} - \text{effective_controls}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 89-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"
```

```
checks = [
    Gate("read public metadata", "low", False),
    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“安全、滥用与威胁建模”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“控制措施强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{residual_risk} = \text{inherent_risk} - \text{effective_controls}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当控制措施强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“安全、滥用与威胁建模”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“控制措施强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{residual_risk} = \text{inherent_risk} - \text{effective_controls}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当控制措施强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“安全、滥用与威胁建模”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如

固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“控制措施强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{residual_risk} = \text{inherent_risk} - \text{effective_controls}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当控制措施强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“安全、滥用与威胁建模”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。

给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“控制措施强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{residual_risk} = \text{inherent_risk} - \text{effective_controls}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当控制措施强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“安全、滥用与威胁建模”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“控制措施强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{residual_risk} = \text{inherent_risk} - \text{effective_controls}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。只是入口，并给出一个类比失效的地方。如果解释中出现

“肯定” “完全” “智能” “理解” 等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当控制措施强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“控制措施强度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 201. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 202. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 203. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 204. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 205. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 206. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 207. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开天文观测的缩小样例作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择虚构机器人在二维网格中的轨迹，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 208. **基础题：** 用自己的话解释“安全、滥用与威胁建模”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

209. **应用题：**在校园能源使用的匿名汇总数据中设计一个只改变“控制措施强度”的实验，列出基线、三档设置、指标和失败样例。

210. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/safety-threat-modeling.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S004] NIST, 《Artificial Intelligence Risk Management Framework (AI RMF 1.0)》, [访问 nist.gov](https://nist.gov))
- [S023] Fairlearn contributors, 《Fairlearn User Guide》, [访问 fairlearn.org](https://fairlearn.org))
- [S024] NIST, 《Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations》, [访问 csrc.nist.gov](https://csrc.nist.gov))

- [S068] NIST, 《Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile》, [访问 nist.gov](https://nist.gov))
- [S069] European Union, 《Regulation (EU) 2024/1689 - Artificial Intelligence Act》, [访问 eur-lex.europa.eu](https://eur-lex.europa.eu))
- [S070] OECD, 《OECD AI Principles》, [访问 oecd.ai](https://oecd.ai))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))
- [S072] Timnit Gebru et al., 《Datasheets for Datasets》, [访问 arxiv.org](https://arxiv.org))
- [S073] Marco Tulio Ribeiro et al., 《Why Should I Trust You?》, [访问 arxiv.org](https://arxiv.org))
- [S074] Scott Lundberg, Su-In Lee, 《A Unified Approach to Interpreting Model Predictions》, [访问 arxiv.org](https://arxiv.org))
- [S075] Cynthia Dwork, Aaron Roth, 《The Algorithmic Foundations of Differential Privacy》, [访问 cis.upenn.edu](https://cis.upenn.edu))
- [S076] H. Brendan McMahan et al., 《Communication-Efficient Learning of Deep Networks from Decentralized Data》, [访问 arxiv.org](https://arxiv.org))
- [S077] Reza Shokri et al., 《Membership Inference Attacks Against Machine Learning Models》, [访问 arxiv.org](https://arxiv.org))
- [S084] OWASP Foundation, 《OWASP Top 10 for LLM Applications》, [访问 genai.owasp.org](https://genai.owasp.org))
- [S085] MITRE, 《MITRE ATLAS》, [访问 atlas.mitre.org](https://atlas.mitre.org))
- [S086] UNESCO, 《Recommendation on the Ethics of Artificial Intelligence》, [访问 unesco.org](https://unesco.org))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 90 章 隐私、差分隐私、联邦学习与模型安全

Privacy and model security

项目	内容
学习单元	月 23 · 安全、隐私、可解释性与治理
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 89 章 安全、滥用与威胁建模](#)

结论先行：理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“隐私、差分隐私、联邦学习与模型安全”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `safety-threat-modeling`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 90-1 隐私、差分隐私、联邦学习与模型安全：单变量实验（教学模拟）

项目	内容
只改变量	隐私参数，范围 0.1–10 ϵ ，步长 0.1
实验目标	只改变“隐私参数”，观察结果、代价和风险如何一起变化，理解理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把隐私参数分别设为 0.1 ϵ 、1 ϵ 和 10 ϵ ，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“隐私参数”，观察结果、代价和风险如何一起变化，理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。

静态替代说明：静态替代：把隐私参数分别设为 0.1ϵ 、 1ϵ 和 10ϵ ，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 90-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```

```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“隐私、差分隐私、联邦学习与模型安全”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐私参数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当隐私参数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“隐私、差分隐私、联邦学习与模型安全”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐私参数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当隐私参数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“隐私、差分隐私、联邦学习与模型安全”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐私参数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当隐私参数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“隐私、差分隐私、联邦学习与模型安全”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅

文字误当成稳定能力。只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐私参数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当隐私参数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“隐私、差分隐私、联邦学习与模型安全”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐私参数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么只把姓名删掉不一定匿名；多项特征

组合仍可能重新识别一个人。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当隐私参数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“隐私参数”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 211. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 212. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 213. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 214. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 215. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 216. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 217. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成音频波形与环境声标签作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开科研论文的标题与摘要元数据，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 218. **基础题：** 用自己的话解释“隐私、差分隐私、联邦学习与模型安全”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

219. **应用题：**在公开古诗文本与版权已明确的材料中设计一个只改变“隐私参数”的实验，列出基线、三档设置、指标和失败样例。

220. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/privacy-security.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S004] NIST, 《Artificial Intelligence Risk Management Framework (AI RMF 1.0)》, [访问 nist.gov](https://nist.gov))
- [S023] Fairlearn contributors, 《Fairlearn User Guide》, [访问 fairlearn.org](https://fairlearn.org))
- [S024] NIST, 《Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations》, [访问 csrc.nist.gov](https://csrc.nist.gov))

- [S068] NIST, 《Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile》, [访问 nist.gov](https://nist.gov))
- [S069] European Union, 《Regulation (EU) 2024/1689 - Artificial Intelligence Act》, [访问 eur-lex.europa.eu](https://eur-lex.europa.eu))
- [S070] OECD, 《OECD AI Principles》, [访问 oecd.ai](https://oecd.ai))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))
- [S072] Timnit Gebru et al., 《Datasheets for Datasets》, [访问 arxiv.org](https://arxiv.org))
- [S073] Marco Tulio Ribeiro et al., 《Why Should I Trust You?》, [访问 arxiv.org](https://arxiv.org))
- [S074] Scott Lundberg, Su-In Lee, 《A Unified Approach to Interpreting Model Predictions》, [访问 arxiv.org](https://arxiv.org))
- [S075] Cynthia Dwork, Aaron Roth, 《The Algorithmic Foundations of Differential Privacy》, [访问 cis.upenn.edu](https://cis.upenn.edu))
- [S076] H. Brendan McMahan et al., 《Communication-Efficient Learning of Deep Networks from Decentralized Data》, [访问 arxiv.org](https://arxiv.org))
- [S077] Reza Shokri et al., 《Membership Inference Attacks Against Machine Learning Models》, [访问 arxiv.org](https://arxiv.org))
- [S084] OWASP Foundation, 《OWASP Top 10 for LLM Applications》, [访问 genai.owasp.org](https://genai.owasp.org))
- [S085] MITRE, 《MITRE ATLAS》, [访问 atlas.mitre.org](https://atlas.mitre.org))
- [S086] UNESCO, 《Recommendation on the Ethics of Artificial Intelligence》, [访问 unesco.org](https://unesco.org))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 91 章 可解释性、审计与模型文档

Explainability and audit

项目	内容
学习单元	月 23 · 安全、隐私、可解释性与治理
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 90 章 隐私、差分隐私、联邦学习与模型安全](#)

结论先行：区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“可解释性、审计与模型文档”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `privacy-security`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 91-1 可解释性、审计与模型文档：单变量实验（教学模拟）

项目	内容
只改变量	解释层级，范围 1–10 层，步长 1
实验目标	只改变“解释层级”，观察结果、代价和风险如何一起变化，理解区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把解释层级分别设为 1 层、3 层和 10 层，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“解释层级”，观察结果、代价和风险如何一起变化，理解区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。

静态替代说明：静态替代：把解释层级分别设为 1 层、3 层和 10 层，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{explanation_quality} = \text{fidelity} + \text{stability} + \text{usefulness}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 91-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```

```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“可解释性、审计与模型文档”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“解释层级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{explanation_quality} = \text{fidelity} + \text{stability} + \text{usefulness}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当解释层级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“可解释性、审计与模型文档”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“解释层级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{explanation_quality} = \text{fidelity} + \text{stability} + \text{usefulness}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当解释层级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“可解释性、审计与模型文档”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如

固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“解释层级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{explanation_quality} = \text{fidelity} + \text{stability} + \text{usefulness}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当解释层级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“可解释性、审计与模型文档”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。

路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“解释层级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{explanation_quality} = \text{fidelity} + \text{stability} + \text{usefulness}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当解释层级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“可解释性、审计与模型文档”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“解释层级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{explanation_quality} = \text{fidelity} + \text{stability} + \text{usefulness}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。只是入口，并给出一个类比失效的地方。如果解释中出现“肯

定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当解释层级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“解释层级”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 221. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 222. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 223. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 224. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 225. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 226. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 227. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择校园能源使用的匿名汇总数据作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成电商目录但不含真实用户行为，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 228. **基础题：** 用自己的话解释“可解释性、审计与模型文档”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

229. **应用题：**在虚构机器人在二维网格中的轨迹中设计一个只改变“解释层级”的实验，列出基线、三档设置、指标和失败样例。

230. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/explainability-audit.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S004] NIST, 《Artificial Intelligence Risk Management Framework (AI RMF 1.0)》，[访问 nist.gov](https://nist.gov))
- [S023] Fairlearn contributors, 《Fairlearn User Guide》，[访问 fairlearn.org](https://fairlearn.org))
- [S024] NIST, 《Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations》，[访问 csrc.nist.gov](https://csrc.nist.gov))

- [S068] NIST, 《Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile》, [访问 nist.gov](https://nist.gov))
- [S069] European Union, 《Regulation (EU) 2024/1689 - Artificial Intelligence Act》, [访问 eur-lex.europa.eu](https://eur-lex.europa.eu))
- [S070] OECD, 《OECD AI Principles》, [访问 oecd.ai](https://oecd.ai))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))
- [S072] Timnit Gebru et al., 《Datasheets for Datasets》, [访问 arxiv.org](https://arxiv.org))
- [S073] Marco Tulio Ribeiro et al., 《Why Should I Trust You?》, [访问 arxiv.org](https://arxiv.org))
- [S074] Scott Lundberg, Su-In Lee, 《A Unified Approach to Interpreting Model Predictions》, [访问 arxiv.org](https://arxiv.org))
- [S075] Cynthia Dwork, Aaron Roth, 《The Algorithmic Foundations of Differential Privacy》, [访问 cis.upenn.edu](https://cis.upenn.edu))
- [S076] H. Brendan McMahan et al., 《Communication-Efficient Learning of Deep Networks from Decentralized Data》, [访问 arxiv.org](https://arxiv.org))
- [S077] Reza Shokri et al., 《Membership Inference Attacks Against Machine Learning Models》, [访问 arxiv.org](https://arxiv.org))
- [S084] OWASP Foundation, 《OWASP Top 10 for LLM Applications》, [访问 genai.owasp.org](https://genai.owasp.org))
- [S085] MITRE, 《MITRE ATLAS》, [访问 atlas.mitre.org](https://atlas.mitre.org))
- [S086] UNESCO, 《Recommendation on the Ethics of Artificial Intelligence》, [访问 unesco.org](https://unesco.org))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 92 章 治理、法律、标准与事故响应

Governance, law and incidents

项目	内容
学习单元	月 23 · 安全、隐私、可解释性与治理
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 91 章 可解释性、审计与模型文档](#)

结论先行：把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“治理、法律、标准与事故响应”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条

件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 explainability-audit。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。类比的 价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对 应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 92-1 治理、法律、标准与事故响应：单变量实验（教学模拟）

项目	内容
只改变量	风险等级，范围 1–4 级，步长 1
实验目标	只改变“风险等级”，观察结果、代价和风险如何一起变化，理解把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把风险等级分别设为 1 级、2 级和 4 级，手工比

项目	内容
	较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“风险等级”，观察结果、代价和风险如何一起变化，理解把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。

静态替代说明：静态替代：把风险等级分别设为 1 级、2 级和 4 级，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`governance = policy + evidence + ownership + enforcement`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 92-1 python 最小实现

```

from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化解理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“治理、法律、标准与事故响应”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，

最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“风险等级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{governance} = \text{policy} + \text{evidence} + \text{ownership} + \text{enforcement}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当风险等级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“治理、法律、标准与事故响应”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实

到责任。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“风险等级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{governance} = \text{policy} + \text{evidence} + \text{ownership} + \text{enforcement}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当风险等级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“治理、法律、标准与事故响应”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“风险等级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{governance} = \text{policy} + \text{evidence} + \text{ownership} + \text{enforcement}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么实验室安全制度规定谁能操作、怎样

记录、事故如何上报；AI 治理也需落实到责任。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当风险等级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“治理、法律、标准与事故响应”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“风险等级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $\text{governance} = \text{policy} + \text{evidence} + \text{ownership} + \text{enforcement}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当风险等级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“治理、法律、标准与事故响应”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“风险等级”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{governance} = \text{policy} + \text{evidence} + \text{ownership} + \text{enforcement}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当风险等级变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“风险等级”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 231. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 232. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 233. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 234. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 235. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 236. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 237. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开古诗文本与版权已明确的材料作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择校园图书馆的公开借阅类别，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

238. **基础题：**用自己的话解释“治理、法律、标准与事故响应”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

239. **应用题：**在公开科研论文的标题与摘要元数据中设计一个只改变“风险等级”的实验，列出基线、三档设置、指标和失败样例。

240. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/governance-law-incidents.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。

- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S004] NIST, 《Artificial Intelligence Risk Management Framework (AI RMF 1.0)》, [访问 nist.gov](https://nist.gov))
- [S023] Fairlearn contributors, 《Fairlearn User Guide》, [访问 fairlearn.org](https://fairlearn.org))
- [S024] NIST, 《Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations》, [访问 csrc.nist.gov](https://csrc.nist.gov))
- [S068] NIST, 《Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile》, [访问 nist.gov](https://nist.gov))
- [S069] European Union, 《Regulation (EU) 2024/1689 - Artificial Intelligence Act》, [访问 eur-lex.europa.eu](https://eur-lex.europa.eu))
- [S070] OECD, 《OECD AI Principles》, [访问 oecd.ai](https://oecd.ai))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))
- [S072] Timnit Gebru et al., 《Datasheets for Datasets》, [访问 arxiv.org](https://arxiv.org))
- [S073] Marco Tulio Ribeiro et al., 《Why Should I Trust You?》, [访问 arxiv.org](https://arxiv.org))
- [S074] Scott Lundberg, Su-In Lee, 《A Unified Approach to Interpreting Model Predictions》, [访问 arxiv.org](https://arxiv.org))
- [S075] Cynthia Dwork, Aaron Roth, 《The Algorithmic Foundations of Differential Privacy》, [访问 cis.upenn.edu](https://cis.upenn.edu))
- [S076] H. Brendan McMahan et al., 《Communication-Efficient Learning of Deep Networks from Decentralized Data》, [访问 arxiv.org](https://arxiv.org))

- [S077] Reza Shokri et al., 《Membership Inference Attacks Against Machine Learning Models》, [访问 arxiv.org](#))
- [S084] OWASP Foundation, 《OWASP Top 10 for LLM Applications》, [访问 genai.owasp.org](#))
- [S085] MITRE, 《MITRE ATLAS》, [访问 atlas.mitre.org](#))
- [S086] UNESCO, 《Recommendation on the Ethics of Artificial Intelligence》, [访问 unesco.org](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

月 23 阶段项目：完成一次端到端 AI 系统红队与治理 评审

项目结论：用虚构系统建立资产、威胁、风险、隐私、解释、法规映射、事件响应和残余风险清单。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

- 241. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
- 242. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
- 243. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
- 244. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 93 章 AI for Science：从预测到发现

AI for Science

项目	内容
学习单元	月 24 · AI for Science、论文复现与原创毕业项目
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 92 章 治理、法律、标准与事故响应](#)

结论先行：理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“AI for Science：从预测到发现”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `governance-law-incidents`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。类比的价
值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔
记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 93-1 AI for Science: 从预测到发现：单变量实验（教学模拟）

项目	内容
只改变量	模拟可信度，范围 0–100%，步长 5
实验目标	只改变“模拟可信度”，观察结果、代价和风险如何一起变化，理解理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把模拟可信度分别设为 0%、70%和 100%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“模拟可信度”，观察结果、代价和风险如何一起变化，理解理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。

静态替代说明：静态替代：把模拟可信度分别设为 0%、70%和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{scientific_value} = \text{prediction} + \text{mechanism} + \text{validation} + \text{reproducibility}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 93-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```

```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“AI for Science：从预测到发现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“模拟可信度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{scientific_value} = \text{prediction} + \text{mechanism} + \text{validation} + \text{reproducibility}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么 AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当模拟可信度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“AI for Science：从预测到发现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“模拟可信度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{scientific_value} = \text{prediction} + \text{mechanism} + \text{validation} + \text{reproducibility}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么 AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当模拟可信度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“AI for Science：从预测到发现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均

值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“模拟可信度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{scientific_value} = \text{prediction} + \text{mechanism} + \text{validation} + \text{reproducibility}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么 AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当模拟可信度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“AI for Science：从预测到发现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能

力。AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“模拟可信度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{scientific_value} = \text{prediction} + \text{mechanism} + \text{validation} + \text{reproducibility}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么 AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当模拟可信度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“AI for Science：从预测到发现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“模拟可信度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{scientific_value} = \text{prediction} + \text{mechanism} + \text{validation} + \text{reproducibility}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么 AI 可以帮助从巨大候选空间筛选实验，

但最终结论仍需要领域知识和实验验证。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当模拟可信度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“模拟可信度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 245. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 246. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 247. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 248. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 249. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 250. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 251. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择虚构机器人在二维网格中的轨迹作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择不含个人信息的教学模拟天气记录，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 252. **基础题：** 用自己的话解释“AI for Science：从预测到发现”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

253. **应用题：**在合成电商目录但不含真实用户行为中设计一个只改变“模拟可信度”的实验，列出基线、三档设置、指标和失败样例。

254. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/ai-for-science.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S005] Stanford HAI, 《AI Index Report 2026》, [访问 hai.stanford.edu](https://hai.stanford.edu))
- [S078] John Jumper et al., 《Highly accurate protein structure prediction with AlphaFold》, [访问 doi.org](https://doi.org/10.26434/chemrxiv-2021-05-01))
- [S079] Remi Lam et al., 《Learning skillful medium-range global weather forecasting》, [访问 doi.org](https://doi.org/10.26434/chemrxiv-2021-05-01))
- [S080] ACM, 《Artifact Review and Badging》, [访问 acm.org](https://artifacts.acm.org))
- [S081] Distill, 《About Distill》, [访问 distill.pub](https://distill.pub))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 94 章 读论文、复现实验与统计核查

Reading papers and replication

项目	内容
学习单元	月 24 · AI for Science、论文复现与原创毕业项目
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 93 章 AI for Science：从预测到发现](#)

结论先行：从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“读论文、复现实验与统计核查”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 ai-for-science。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 94-1 读论文、复现实验与统计核查：单变量实验（教学模拟）

项目	内容
只改变量	独立复现次数，范围 1-30 次，步长 1
实验目标	只改变“独立复现次数”，观察结果、代价和风险如何一起变化，理解从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把独立复现次数分别设为 1 次、5 次和 30 次，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“独立复现次数”，观察结果、代价和风险如何一起变化，理解从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。

静态替代说明：静态替代：把独立复现次数分别设为 1 次、5 次和 30 次，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{replication_gap} = \text{reported_result} - \text{reproduced_result}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 94-1 python 最小实现

```
from dataclasses import dataclass

@dataclass
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
```



```

    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“读论文、复现实验与统计核查”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“独立复现次数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{replication_gap} = \text{reported_result} - \text{reproduced_result}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当独立复现次数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“读论文、复现实验与统计核查”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“独立复现次数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{replication_gap} = \text{reported_result} - \text{reproduced_result}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当独立复现次数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“读论文、复现实验与统计核查”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、

均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“独立复现次数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{replication_gap} = \text{reported_result} - \text{reproduced_result}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当独立复现次数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“读论文、复现实验与统计核查”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当

成稳定能力。复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“独立复现次数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{replication_gap} = \text{reported_result} - \text{reproduced_result}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当独立复现次数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“读论文、复现实验与统计核查”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“独立复现次数”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{replication_gap} = \text{reported_result} - \text{reproduced_result}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完

全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当独立复现次数变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“独立复现次数”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 255. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 256. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 257. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 258. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 259. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 260. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 261. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开科研论文的标题与摘要元数据作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开道路网络与虚构出行任务，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 262. **基础题：** 用自己的话解释“读论文、复现实验与统计核查”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

263. **应用题：**在校园图书馆的公开借阅类别中设计一个只改变“独立复现次数”的实验，列出基线、三档设置、指标和失败样例。

264. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/reading-and-replication.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S005] Stanford HAI, 《AI Index Report 2026》, [访问 hai.stanford.edu](https://hai.stanford.edu))
- [S078] John Jumper et al., 《Highly accurate protein structure prediction with AlphaFold》, [访问 doi.org](https://doi.org/10.26434/chemrxiv-2021-03-01))
- [S079] Remi Lam et al., 《Learning skillful medium-range global weather forecasting》, [访问 doi.org](https://doi.org/10.26434/chemrxiv-2021-03-01))
- [S080] ACM, 《Artifact Review and Badging》, [访问 acm.org](https://artifacts.acm.org))

- [S081] Distill, 《About Distill》, [访问 distill.pub](https://distill.pub))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 95 章 研究问题、创新、消融与沟通

Research questions and ablations

项目	内容
学习单元	月 24 · AI for Science、论文复现与原创毕业项目
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 94 章 读论文、复现实验与统计核查](#)

结论先行：把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“研究问题、创新、消融与沟通”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条

件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 reading-and-replication。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 95-1 研究问题、创新、消融与沟通：单变量实验（教学模拟）

项目	内容
只改变量	消融数量，范围 1-20 项，步长 1
实验目标	只改变“消融数量”，观察结果、代价和风险如何一起变化，理解把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把消融数量分别设为 1 项、4 项和 20 项，手工比较三组教学模拟输出。

项目	内容
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“消融数量”，观察结果、代价和风险如何一起变化，理解把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。

静态替代说明：静态替代：把消融数量分别设为 1 项、4 项和 20 项，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`contribution = new_question | new_method | new_evidence | new_system`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 95-1 python 最小实现

```
from dataclasses import dataclass
@dataclass
```

```

class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“研究问题、创新、消融与沟通”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消融数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `contribution = new_question | new_method | new_evidence | new_system`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当消融数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“研究问题、创新、消融与沟通”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、

均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消融数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{contribution} = \text{new_question} \mid \text{new_method} \mid \text{new_evidence} \mid \text{new_system}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当消融数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“研究问题、创新、消融与沟通”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消融数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{contribution} = \text{new_question} \mid \text{new_method} \mid \text{new_evidence} \mid \text{new_system}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当消融数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“研究问题、创新、消融与沟通”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误

当成稳定能力。拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消融数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `contribution = new_question | new_method | new_evidence | new_system`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当消融数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“研究问题、创新、消融与沟通”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“消融数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `contribution = new_question | new_method | new_evidence | new_system`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完

全” “智能” “理解” 等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当消融数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“消融数量”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

265. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

- 266. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 267. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 268. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 269. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 270. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 271. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成电商目录但不含真实用户行为作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成的植物叶片特征，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 272. **基础题：** 用自己的话解释“研究问题、创新、消融与沟通”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。
- 273. **应用题：** 在不含个人信息的教学模拟天气记录中设计一个只改变“消融数量”的实验，列出基线、三档设置、指标和失败样例。
- 274. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/research-questions-ablations.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S005] Stanford HAI, 《AI Index Report 2026》, [访问 hai.stanford.edu](https://hai.stanford.edu))
- [S078] John Jumper et al., 《Highly accurate protein structure prediction with AlphaFold》, [访问 doi.org](https://doi.org))
- [S079] Remi Lam et al., 《Learning skillful medium-range global weather forecasting》, [访问 doi.org](https://doi.org))
- [S080] ACM, 《Artifact Review and Badging》, [访问 acm.org](https://acm.org))
- [S081] Distill, 《About Distill》, [访问 distill.pub](https://distill.pub))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 96 章 毕业项目：部署、答辩与前沿追踪

Capstone and frontiers

项目	内容
学习单元	月 24 · AI for Science、论文复现与原创毕业项目
建议时间	10 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 95 章 研究问题、创新、消融与沟通](#)

结论先行：整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“毕业项目：部署、答辩与前沿追踪”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署

条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 research-questions-ablations。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 96-1 毕业项目：部署、答辩与前沿追踪：单变量实验（教学模拟）

项目	内容
只改变量	项目范围，范围 1–12 周，步长 1
实验目标	只改变“项目范围”，观察结果、代价和风险如何一起变化，理解整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把项目范围分别设为 1 周、4 周和 12 周，手工比较三组教学模拟输出。

项目	内容
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“项目范围”，观察结果、代价和风险如何一起变化，理解整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。

静态替代说明：静态替代：把项目范围分别设为 1 周、4 周和 12 周，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`frontier_skill = foundations × practice × critique × community × time`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 96-1 python 最小实现

```
from dataclasses import dataclass
@dataclass
```

```
class Gate:
    action: str
    risk: str
    requires_human_approval: bool

def authorize(gate: Gate) -> str:
    if gate.requires_human_approval:
        return f"STOP: wait for a human before {gate.action}"
    return f"ALLOW: {gate.action}"

checks = [
    Gate("read public metadata", "low", False),
    Gate("publish or write externally", "high", True),
]
for check in checks:
    print(authorize(check))
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“毕业项目：部署、答辩与前沿追踪”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“项目范围”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{frontier_skill} = \text{foundations} \times \text{practice} \times \text{critique} \times \text{community} \times \text{time}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当项目范围变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“毕业项目：部署、答辩与前沿追踪”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“项目范围”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{frontier_skill} = \text{foundations} \times \text{practice} \times \text{critique} \times \text{community} \times \text{time}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当项目范围变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“毕业项目：部署、答辩与前沿追踪”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。。接着逐项标记观察到的事实、由公

式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“项目范围”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{frontier_skill} = \text{foundations} \times \text{practice} \times \text{critique} \times \text{community} \times \text{time}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当项目范围变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“毕业项目：部署、答辩与前沿追踪”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“项目范围”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{frontier_skill} = \text{foundations} \times \text{practice} \times \text{critique} \times \text{community} \times \text{time}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。只是入口，并给出一个类比失效的地方。如果解释中出现“肯

定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当项目范围变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“毕业项目：部署、答辩与前沿追踪”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“项目范围”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{frontier_skill} = \text{foundations} \times \text{practice} \times \text{critique} \times \text{community} \times \text{time}$ ，逐个说明

符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当项目范围变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“项目范围”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 275. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 276. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 277. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 278. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 279. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 280. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 281. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择校园图书馆的公开借阅类别作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择虚构游戏角色的状态日志，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

282. **基础题：** 用自己的话解释“毕业项目：部署、答辩与前沿追踪”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。
283. **应用题：** 在公开道路网络与虚构出行任务中设计一个只改变“项目范围”的实验，列出基线、三档设置、指标和失败样例。
284. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/capstone-and-frontiers.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S005] Stanford HAI，《AI Index Report 2026》，[访问 hai.stanford.edu](https://hai.stanford.edu)

- [S078] John Jumper et al., 《Highly accurate protein structure prediction with AlphaFold》, [访问 doi.org](#))
- [S079] Remi Lam et al., 《Learning skillful medium-range global weather forecasting》, [访问 doi.org](#))
- [S080] ACM, 《Artifact Review and Badging》, [访问 acm.org](#))
- [S081] Distill, 《About Distill》, [访问 distill.pub](#))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

月 24 阶段项目：原创毕业项目与公开答辩包

项目结论：选择低风险问题，完成文献图谱、基线、复现、消融、失败分析、部署、模型卡、演示和下一步研究问题。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

- 285. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
- 286. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
- 287. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
- 288. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

独立练习答案

先完成题目、保留自己的推理和错误，再查看参考答案。答案给出最低证据结构，不是唯一正确表述。

第 73 章答案 · 图表示、消息传递与 GNN

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解节点、边、属性、邻接、置换不变和消息传递。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“同学关系网中每个人从邻居收集信息，再更新自己的表示。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成的植物叶片特征，先用固定规则或简单统计量作为基线。把“消息传播跳数”设置为 1 跳、2 跳和 10 跳，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 74 章答案 · 推荐、检索与排序系统

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，覆盖召回、排序、矩阵分解、反馈回路、多样性和长期价值。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“图书馆先从百万本书召回候选，再按相关性排序；只追点击会放大狭窄偏好。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择虚构游戏角色的状态日志，先用固定规则或简单统计量作为基线。把“多样性权重”设置为 0、0.3 和 1，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 75 章答案 · 时间序列、预测与回测

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解趋势、季节、滞后、平稳性、预测区间和时间切分。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“预测明天气温不能随机打乱日期；未来信息绝不能泄漏进过去。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开天文观测的缩小样例，先用固定规则或简单统计量作为基线。把“回看窗口”设置为 1 天、30 天和 365 天，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 76 章答案 · 时空建模、异常与决策智能

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，结合位置、时间、网络和行动，避免把预测直接当决策。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“公交拥堵既沿道路传播又随时段变化；空间邻接和时间滞后要一起看。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成音频波形与环境声标签，先用固定规则或简单统计量作为基线。把“空间邻域”设置为 1km、10km 和 100km，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 77 章答案 · 语言模型词元器与嵌入

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，实现子词切分、词表、位置和输入输出嵌入，分析不同语言的长度差异。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“积木尺寸会影响一段话需要多少块，也影响稀有词如何共享部件。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择校园能源使用的匿名汇总数据，先用固定规则或简单统计量作为基线。把“词表大小”设置为 100、8000 和 32000，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 78 章答案 · 从零实现因果 Transformer

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，实现归一化、多头因果注意力、前馈层、残差和输出头。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“每个词只能看左侧已读内容，因果遮罩像考试时遮住未来答案。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开古诗文本与版权已明确的材料，先用固定规则或简单统计量作为基线。把“上下文长度”设置为 8token、256token 和 4096token，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 79 章答案 · 预训练目标、数据与缩放

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解下一个词预测、数据混合、去重、训练计算和缩放规律的适用边界。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“增加练习量有效，但重复同一道题不会等价于增加新知识。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择虚构机器人在二维网格中的轨迹，先用固定规则或简单统计量作为基线。把“训练词元”设置为 1000token、100000token 和 1000000token，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 80 章答案 · 指令微调、偏好学习、RAG 与评估

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，区分预训练、监督微调、偏好优化、检索增强和系统评估。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“基础模型像读过许多书的学生；指令训练教作答方式，检索给现场资料，评估检查结果。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开科研论文的标题与摘要元数据，先用固定规则或简单统计量作为基线。把“检索候选数”设置为 1 条、5 条和 30 条，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 81 章答案 · 多模态表示、对齐与融合

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解图像、文本、音频的编码、对比对齐、交叉注意力和模态缺失。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“给同一事件配照片、说明和声音，模型学习它们在共享空间的对应关系。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成电商目录但不含真实用户行为，先用固定规则或简单统计量作为基线。把“对齐温度”设置为 0.01、0.07 和 1，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 82 章答案 · 世界模型、视频与行动预测

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，学习潜在动力学、未来预测、规划和模型误差累积。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“脑中先模拟篮球落点再行动；模拟错得越远，长期计划越危险。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择校园图书馆的公开借阅类别，先用固定规则或简单统计量作为基线。把“想象步数”设置为 1 步、10 步和 100 步，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 83 章答案 · Agent 循环、工具、记忆与规划

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，把观察、计划、调用、验证和停止条件做成显式状态机。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“项目助理先读任务，再查资料、写草稿、核对；每个外部动作都有权限边界。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择不含个人信息的教学模拟天气记录，先用固定规则或简单统计量作为基线。把“工具调用预算”设置为 1 次、8 次和 50 次，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 84 章答案 · Agent 评估、安全与人在回路

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，测试任务成功、引用、权限、提示注入、工具故障和可恢复性。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“给实习生钥匙前要限定房间、记录使用、设置审批，而不是只看他是否聪明。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开道路网络与虚构出行任务，先用固定规则或简单统计量作为基线。把“自主级别”设置为 0 级、1 级和 5 级，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 85 章答案 · 数据治理、去重、质量与许可

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，建立来源、许可、去重、污染、质量、退出和删除流程。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“图书馆不能只堆书，还要编目、查重、确认授权并处理撤架。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成的植物叶片特征，先用固定规则或简单统计量作为基线。把“相似阈值”设置为 0、0.8 和 1，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 86 章答案 · 分布式训练、并行与检查点

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解数据并行、张量并行、流水线并行、通信、显存和故障恢复。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“多人搬家可以按箱子分工、按家具拆件或按楼层流水线，但协调成本不同。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择虚构游戏角色的状态日志，先用固定规则或简单统计量作为基线。把“并行工作者”设置为 1 个、8 个和 128 个，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 87 章答案·推理、量化、蒸馏与服务

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解批处理、缓存、量化、剪枝、蒸馏、吞吐、延迟和质量折衷；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“餐厅合并订单提高吞吐，但第一位顾客可能等待更久；服务要平衡延迟和利用率。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开天文观测的缩小样例，先用固定规则或简单统计量作为基线。把“量化位宽”设置为 2bit、8bit 和 16bit，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 88 章答案 · MLOps、监控、可靠性、成本与碳

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，连接持续集成、模型注册、漂移、服务目标、事故响应、成本和资源影响。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“上线模型像经营公交线路：不仅要发车，还要看准点率、故障、客流变化和备用方案。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成音频波形与环境声标签，先用固定规则或简单统计量作为基线。把“告警阈值”设置为 0%、20%和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上

一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 89 章答案 · 安全、滥用与威胁建模

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，从资产、攻击者、能力、路径、影响和缓解构建分层风险模型。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“给房子做安全设计要先知道保护什么、谁可能进入、从哪里进入以及失守后果。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择校园能源使用的匿名汇总数据，先用固定规则或简单统计量作为基线。把“控制措施强度”设置为 0%、60%和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上

一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 90 章答案·隐私、差分隐私、联邦学习与模型安全

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解数据最小化、访问控制、差分隐私、联邦学习、反演与成员推断。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“只把姓名删掉不一定匿名；多项特征组合仍可能重新识别一个人。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开古诗文本与版权已明确的材料，先用固定规则或简单统计量作为基线。把“隐私参数”设置为 0.1ϵ 、 1ϵ 和 10ϵ ，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上

一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 91 章答案 · 可解释性、审计与模型文档

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，区分全局/局部、内在/事后解释，验证忠实度、稳定性与用途。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“路线图可以解释导航建议，但若地图本身错了，清晰解释仍不等于正确。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择虚构机器人在二维网格中的轨迹，先用固定规则或简单统计量作为基线。把“解释层级”设置为 1 层、3 层和 10 层，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上

一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 92 章答案 · 治理、法律、标准与事故响应

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，把角色责任、风险分级、记录、监测、申诉和事件响应映射到具体法域。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“实验室安全制度规定谁能操作、怎样记录、事故如何上报；AI 治理也需落实到责任。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开科研论文的标题与摘要元数据，先用固定规则或简单统计量作为基线。把“风险等级”设置为 1 级、2 级和 4 级，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 93 章答案 · AI for Science：从预测到发现

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解结构预测、模拟加速、逆向设计、科学机器学习和实验闭环。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“AI 可以帮助从巨大候选空间筛选实验，但最终结论仍需要领域知识和实验验证。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成电商目录但不含真实用户行为，先用固定规则或简单统计量作为基线。把“模拟可信度”设置为 0%、70%和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 94 章答案 · 读论文、复现实验与统计核查

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，从问题、相关工作、方法、实验、限制阅读，先复现基线再追求创新。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“复刻一道菜先确认原料、步骤和温度，不能只看成品照片猜做法。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择校园图书馆的公开借阅类别，先用固定规则或简单统计量作为基线。把“独立复现次数”设置为 1 次、5 次和 30 次，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 95 章答案 · 研究问题、创新、消融与沟通

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，把模糊兴趣缩成可证伪问题，用基线、对照、消融和误差分析支撑贡献。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“拆掉自行车的一个部件观察变化，才能判断它是否真正带来效果。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择不含个人信息的教学模拟天气记录，先用固定规则或简单统计量作为基线。把“消融数量”设置为 1 项、4 项和 20 项，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 96 章答案 · 毕业项目：部署、答辩与前沿追踪

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，整合问题、数据、模型、系统、安全、评估和研究沟通，建立持续更新机制。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“毕业不是地图终点，而是你能独立规划下一段路线并解释为何这样走。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开道路网络与虚构出行任务，先用固定规则或简单统计量作为基线。把“项目范围”设置为 1 周、4 周和 12 周，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

本卷来源台账

本卷共引用 44 条来源。正文为原创解释；不复制受版权保护的图表、习题答案或大段原文。访问日期均为当前 Beta 版核实记录。

[S004] Artificial Intelligence Risk Management Framework (AI RMF 1.0)

项目	内容
作者/机构	NIST
来源等级	标准机构一手资料
发布日期	2023
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 nist.gov

[S005] AI Index Report 2026

项目	内容
作者/机构	Stanford HAI
来源等级	研究机构年度报告

项目	内容
发布日期	2026
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 hai.stanford.edu

[S023] Fairlearn User Guide

项目	内容
作者/机构	Fairlearn contributors
来源等级	开源项目官方文档
发布日期	持续更新
访问日期	2026-08-06
许可	MIT
DOI/URL	访问 fairlearn.org

[S024] Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

项目	内容
----	----

项目	内容
作者/机构	NIST
来源等级	标准机构一手资料
发布日期	2025
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 csrc.nist.gov

[S030] Attention Is All You Need

项目	内容
作者/机构	Ashish Vaswani et al.
来源等级	同行评审论文预印本
发布日期	2017
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名

项目	内容
DOI/URL	访问 arxiv.org

[S046] CS224W: Machine Learning with Graphs

项目	内容
作者/机构	Stanford University
来源等级	大学课程一手资料
发布日期	持续更新
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 web.stanford.edu

[S047] Semi-Supervised Classification with Graph Convolutional Networks

项目	内容
作者/机构	Thomas Kipf, Max Welling
来源等级	同行评审论文预印本
发布日期	2016

项目	内容
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S048] Matrix Factorization Techniques for Recommender Systems

项目	内容
作者/机构	Yehuda Koren, Robert Bell, Chris Volinsky
来源等级	同行评审综述
发布日期	2009
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 doi.org

[S049] Forecasting: Principles and Practice

项目	内容
----	----

项目	内容
作者/机构	Rob J Hyndman, George Athanasopoulos
来源等级	开放教材
发布日期	2021
访问日期	2026-08-06
许可	CC BY-NC-ND 4.0
DOI/URL	访问 otexts.com

[S050] CS336: Language Modeling from Scratch

项目	内容
作者/机构	Stanford University
来源等级	大学课程一手资料
发布日期	2026
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 cs336.stanford.edu

[S051] SentencePiece

项目	内容
作者/机构	Taku Kudo, John Richardson
来源等级	同行评审论文预印本
发布日期	2018
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S052] Language Models are Few-Shot Learners

项目	内容
作者/机构	Tom Brown et al.
来源等级	同行评审论文预印本
发布日期	2020
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名

项目	内容
	名
DOI/URL	访问 arxiv.org

[S053] Training Compute-Optimal Large Language Models

项目	内容
作者/机构	Jordan Hoffmann et al.
来源等级	同行评审论文预印本
发布日期	2022
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S054] Training language models to follow instructions with human feedback

项目	内容
作者/机构	Long Ouyang et al.
来源等级	同行评审论文预印本

项目	内容
发布日期	2022
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S055] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

项目	内容
作者/机构	Patrick Lewis et al.
来源等级	同行评审论文预印本
发布日期	2020
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S056] Learning Transferable Visual Models From Natural Language Supervision

项目	内容
作者/机构	Alec Radford et al.
来源等级	同行评审论文预印本
发布日期	2021
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S057] World Models

项目	内容
作者/机构	David Ha, Jürgen Schmidhuber
来源等级	同行评审论文预印本
发布日期	2018
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名

项目	内容
	名
DOI/URL	访问 arxiv.org

[S058] ReAct: Synergizing Reasoning and Acting in Language Models

项目	内容
作者/机构	Shunyu Yao et al.
来源等级	同行评审论文预印本
发布日期	2022
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S059] Toolformer

项目	内容
作者/机构	Timo Schick et al.
来源等级	同行评审论文预印本

项目	内容
发布日期	2023
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S060] AgentBench

项目	内容
作者/机构	Xiao Liu et al.
来源等级	研究论文预印本
发布日期	2023
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S061] Fully Sharded Data Parallel

项目	内容
作者/机构	PyTorch
来源等级	官方文档
发布日期	持续更新
访问日期	2026-08-06
许可	BSD-3-Clause
DOI/URL	访问 docs.pytorch.org

[S062] Megatron-LM

项目	内容
作者/机构	Mohammad Shoeybi et al.
来源等级	研究论文预印本
发布日期	2019
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名

项目	内容
DOI/URL	访问 arxiv.org

[S063] ZeRO

项目	内容
作者/机构	Samyam Rajbhandari et al.
来源等级	同行评审论文预印本
发布日期	2019
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S064] FlashAttention

项目	内容
作者/机构	Tri Dao et al.
来源等级	同行评审论文预印本
发布日期	2022

项目	内容
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S065] Efficient Memory Management for Large Language Model Serving with PagedAttention

项目	内容
作者/机构	Woosuk Kwon et al.
来源等级	同行评审论文预印本
发布日期	2023
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S066] MLflow Documentation

项目	内容
----	----

项目	内容
作者/机构	MLflow Project
来源等级	官方文档
发布日期	持续更新
访问日期	2026-08-06
许可	Apache-2.0
DOI/URL	访问 mlflow.org

[S067] Kubernetes Documentation

项目	内容
作者/机构	Kubernetes
来源等级	官方文档
发布日期	持续更新
访问日期	2026-08-06
许可	CC BY 4.0
DOI/URL	访问 kubernetes.io

[S068] Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

项目	内容
作者/机构	NIST
来源等级	标准机构一手资料
发布日期	2024
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 nist.gov

[S069] Regulation (EU) 2024/1689 - Artificial Intelligence Act

项目	内容
作者/机构	European Union
来源等级	法规原文
发布日期	2024
访问日期	2026-08-06
许可	EU legal notice applies

项目	内容
DOI/URL	访问 eur-lex.europa.eu

[S070] OECD AI Principles

项目	内容
作者/机构	OECD
来源等级	政府间组织一手框架
发布日期	2019/2024 更新
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 oecd.ai

[S071] Model Cards for Model Reporting

项目	内容
作者/机构	Margaret Mitchell et al.
来源等级	同行评审论文预印本
发布日期	2018

项目	内容
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S072] Datasheets for Datasets

项目	内容
作者/机构	Timnit Gebru et al.
来源等级	同行评审论文预印本
发布日期	2018
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S073] Why Should I Trust You?

项目	内容
----	----

项目	内容
作者/机构	Marco Tulio Ribeiro et al.
来源等级	同行评审论文预印本
发布日期	2016
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S074] A Unified Approach to Interpreting Model Predictions

项目	内容
作者/机构	Scott Lundberg, Su-In Lee
来源等级	同行评审论文预印本
发布日期	2017
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名

项目	内容
DOI/URL	访问 arxiv.org

[S075] The Algorithmic Foundations of Differential Privacy

项目	内容
作者/机构	Cynthia Dwork, Aaron Roth
来源等级	开放专著
发布日期	2014
访问日期	2026-08-06
许可	作者公开版本
DOI/URL	访问 cis.upenn.edu

[S076] Communication-Efficient Learning of Deep Networks from Decentralized Data

项目	内容
作者/机构	H. Brendan McMahan et al.
来源等级	同行评审论文预印本
发布日期	2016

项目	内容
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S077] Membership Inference Attacks Against Machine Learning Models

项目	内容
作者/机构	Reza Shokri et al.
来源等级	同行评审论文预印本
发布日期	2016
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S078] Highly accurate protein structure prediction with AlphaFold

项目	内容
----	----

项目	内容
作者/机构	John Jumper et al.
来源等级	同行评审论文
发布日期	2021
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 doi.org

[S079] Learning skillful medium-range global weather forecasting

项目	内容
作者/机构	Remi Lam et al.
来源等级	同行评审论文
发布日期	2023
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名

项目	内容
DOI/URL	访问 doi.org

[S080] Artifact Review and Badging

项目	内容
作者/机构	ACM
来源等级	专业学会复现规范
发布日期	持续更新
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 acm.org

[S081] About Distill

项目	内容
作者/机构	Distill
来源等级	互动科学传播参考
发布日期	持续更新

项目	内容
访问日期	2026-08-06
许可	以各文章许可为准
DOI/URL	访问 distill.pub

[S084] OWASP Top 10 for LLM Applications

项目	内容
作者/机构	OWASP Foundation
来源等级	行业安全项目一手资料
发布日期	持续更新
访问日期	2026-08-06
许可	CC BY-SA 4.0
DOI/URL	访问 genai.owasp.org

[S085] MITRE ATLAS

项目	内容
作者/机构	MITRE

项目	内容
来源等级	威胁知识库一手资料
发布日期	持续更新
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 atlas.mitre.org

[S086] Recommendation on the Ethics of Artificial Intelligence

项目	内容
作者/机构	UNESCO
来源等级	国际组织规范性文件
发布日期	2021
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 unesco.org

全书术语表

共 81 个核心术语。定义用于建立共同语言，不能替代章节中的数学、实验和边界分析。

术语	英文	小白解释
人工智能	Artificial Intelligence	让机器执行通常需要感知、推理、学习、规划或生成等能力的研究与工程领域。
算法	Algorithm	把输入转换为输出的一组明确步骤；算法本身不等于训练后的模型。
模型	Model	从数据或规则中得到、用于预测或决策的数学结构及其参数。
参数	Parameter	训练过程中被优化的数值，例如神经网络权重。
超参数	Hyperparameter	由开发者或搜索过程设定、不是直接从单次梯度更新学得s的配置。
数据集	Dataset	带有明确来源、字段、许可和用途的一组样本。

术语	英文	小白解释
特征	Feature	模型使用的输入表示。
标签	Label	监督学习中希望模型预测的目标值。
训练	Training	用目标函数和优化方法调整模型参数的过程。
推理	Inference	用已经得到的模型对新输入计算输出。
损失函数	Loss function	把预测与目标之间的差异变成可优化数值的函数。
目标函数	Objective	训练或决策希望最大化或最小化的量，可能包含多个损失和约束。
梯度	Gradient	函数对多个变量的一阶变化率组成的向量。
泛化	Generalization	模型在未参与训练、但与目标场景相关的数据上保持能力。
过拟合	Overfitting	模型过度适应训练数据细节，导致新数据表现下降。

术语	英文	小白解释
欠拟合	Underfitting	模型或训练不足以捕捉任务中的重要结构。
正则化	Regularization	通过惩罚、噪声或结构限制降低过拟合风险的方法。
基线	Baseline	用于判断新方法是否真正改进的简单或既有方法。
消融实验	Ablation	移除或替换系统一部分，观察它对结果的贡献。
随机种子	Random seed	初始化伪随机数生成器的值； 固定种子不是完整可复现性的替代品。
张量	Tensor	带有一个或多个轴的数值数组。
嵌入	Embedding	把离散对象映射到连续向量空间的表示。
注意力	Attention	根据查询与键的匹配程度加权汇总值的机制。
Transformer	—	以注意力、前馈层、残差和归一化为核心的序列模型架构。

术语	英文	小白解释
词元	Token	文本经过词元器切分后的基本单位，可是字符、子词或其他片段。
上下文窗口	Context window	模型一次计算中能够直接处理的词元范围。
基础模型	Foundation model	在广泛数据上训练、可适配多种下游任务的大规模模型。
语言模型	Language model	为词元序列分配概率或预测后续词元的模型。
检索增强生成	RAG	在生成前检索外部资料，并把证据提供给生成模型的系统模式。
Agent	—	围绕目标执行观察、计划、工具调用、验证和停止循环的系统。
提示注入	Prompt injection	不可信输入试图改变模型或 Agent 原定指令和权限的攻击。
强化学习	Reinforcement learning	智能体通过与环境交互、基于

术语	英文	小白解释
		奖励学习策略的方法。
策略	Policy	在给定状态下选择动作的规则或概率分布。
价值函数	Value function	从状态或状态动作对出发的预期累计回报。
马尔可夫决策过程	MDP	用状态、动作、转移、奖励和折扣描述序列决策的形式框架。
卷积	Convolution	用共享局部核在输入位置上滑动计算的运算。
计算机视觉	Computer vision	让机器从图像或视频中提取、生成和使用信息的领域。
自然语言处理	NLP	让计算机处理、理解、检索或生成人类语言的领域。
自动语音识别	ASR	把语音波形转换为文字或其他符号序列。
扩散模型	Diffusion model	通过学习逐步去噪过程生成样本的一类模型。

术语	英文	小白解释
生成对抗网络	GAN	通过生成器与判别器对抗训练学习生成分布的架构。
图神经网络	GNN	在图结构上通过邻域聚合或消息传递学习表示的模型。
校准	Calibration	预测概率与事件长期发生频率之间的一致程度。
分布偏移	Distribution shift	训练、评估与部署数据分布不一致的现象。
分布外	OOD	与训练或预期部署分布明显不同的输入。
因果推断	Causal inference	在明确假设下估计干预对结果影响的方法。
公平性	Fairness	围绕不同群体或个体的机会、错误、待遇和影响制定并审查的规范目标。
鲁棒性	Robustness	系统面对扰动、噪声、故障或环境变化仍保持可接受行为的能力。
可解释性	Explainability	为模型行为提供可理解说明的

术语	英文	小白解释
		方法和性质；解释清楚不等于预测正确。
差分隐私	Differential privacy	通过限制单条记录对输出分布的影响提供可量化隐私保证的框架。
联邦学习	Federated learning	在数据留在多个设备或组织的情况下协同训练的范式。
模型卡	Model card	记录模型用途、评估、限制和风险的结构化文档。
数据表	Datasheet	记录数据来源、构建、用途、限制和维护方式的结构化文档。
MLOps	—	管理机器学习数据、训练、部署、监控、版本和事故响应的工程实践。
延迟	Latency	一次请求从进入到完成所需的时间。
吞吐	Throughput	单位时间内系统能处理的请求或词元数量。

术语	英文	小白解释
量化	Quantization	用较低精度数值表示权重或激活以降低资源成本的方法。
蒸馏	Distillation	让较小模型学习较大模型输出或内部结构的方法。
并行训练	Distributed training	把训练计算、数据或参数分布到多个设备或进程。
检查点	Checkpoint	保存训练状态以便恢复、评估或部署的版本化工件。
可复现	Reproducible	他人依据充分工件和说明能够得到相容结果的性质。
同行评审	Peer review	由相关领域同行对方法、证据和表述进行独立审查。
置信区间	Confidence interval	按指定重复抽样程序构造、具有长期覆盖率解释的区间。
p 值	p-value	在零假设及检验假设成立时，观察到至少同样极端统计量的概率。
效应量	Effect size	差异或关联实际大小的度量。

术语	英文	小白解释
熵	Entropy	概率分布不确定性的度量。
交叉熵	Cross entropy	用一个分布编码另一个分布时的平均信息量，也是常见分类损失。
KL 散度	KL divergence	衡量一个概率分布相对另一个分布差异的非对称量。
贝叶斯更新	Bayesian update	用似然把先验信念更新为后验信念。
知识图谱	Knowledge graph	用实体、关系和属性构成的结构化知识网络。
本体	Ontology	对领域概念、关系和约束的显式规范。
搜索启发式	Heuristic	估计距离或代价、用于决定优先探索方向的函数。
约束满足	CSP	为变量选择值，使所有约束同时满足的问题。
SLAM	—	机器人同时定位并建立环境地图的问题。

术语	英文	小白解释
控制	Control	根据目标与反馈选择输入，使动态系统按期望演化。
仿真到现实	Sim-to-real	把在仿真中得到的方法迁移到真实系统的过程及差距问题。
世界模型	World model	学习环境状态和动作如何随时间演化的内部模型。
多模态	Multimodal	联合处理文本、图像、音频、视频或动作等多种信息形式。
AI for Science	—	使用 AI 辅助科学建模、模拟、实验设计、分析和发现的方向。
人机协同	Human-in-the-loop	在系统关键环节保留明确人工检查、判断或控制权的设计。
风险	Risk	事件发生可能性与后果的组合，必须结合具体情境评估。

版权、版本与勘误

2026.08 Beta · 第四卷：前沿、系统与研究 · 评审状态 draft。

Copyright © 2026 NEWVAR LAB. All rights reserved.

全部原创正文、实验定义和代码保留全部版权。第三方名称、论文与链接仅作明确署名的教学引用，其权利归原权利人。

本书仍处于 Beta：可能存在表达、事实、代码或排版问题。请通过 NEWVAR 版本页查看核实日期、评审状态与更正记录。

任何时效性结论都应在使用前回到一手来源重新核实；医疗、法律、财务、教育机会和安全关键场景必须由合格专业人员判断。

学习网站：[访问 newvarlab.com](https://newvarlab.com)）。