

科研素养训练II

第4次课：复现结果比较与偏差解释

宋庆禹
2026.10.12

分组汇报安排



廈門大學
XIAMEN UNIVERSITY

10.19

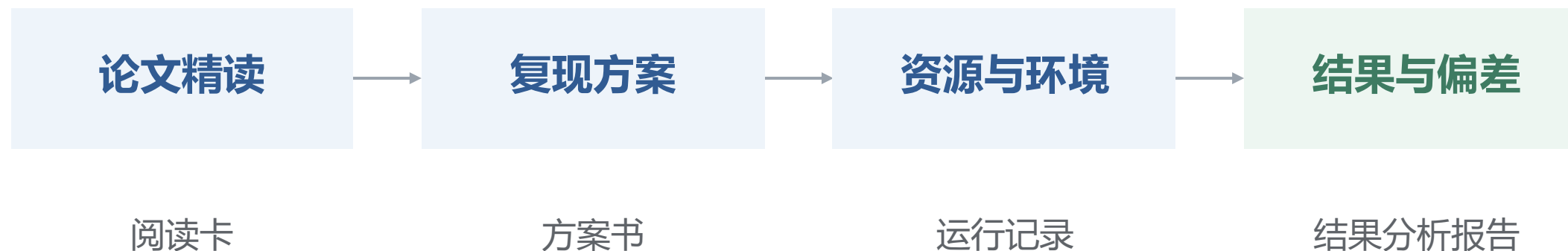
论文选题汇报

10.26

复现结果汇报



四次训练如何形成证据链



最终判断应能回查：依据 → 运行 → 条件 → 原文目标



从一句结果陈述开始

“我跑出了75.83%，所以已经复现成功。”

与谁比较？

条件是否对应？

满足什么判据？

问题不在于75.83%高不高，而在于缺少来源、比较对象和判断标准。

数值与目标、条件、证据的关系，才是判断依据



课程案例中的两个分析问题

20 Newsgroups四类别文本分类

新闻组文本 → TF-IDF特征 → RidgeClassifier → 主题类别

无神论 / 图形学 / 太空 / 宗教杂谈；训练2034条，测试1353条

A 实现一致性

官方脚本与我的教学封装
在相同条件下预测是否一致？

B 处理条件比较

保留与移除元信息
准确率及错误分布如何变化？

这是官方公开实验示例，不是某篇论文全部实验的复现



冒烟结果与正式结果的作用

配置	冒烟测试	完整实验
训练样本数	320	2034
测试样本数	120	1353
准确率	58.33%	75.83%
用途	流程检查	结果分析

提高17.50个百分点，不能直接简单归因，例如“修复算法”



结果数值必须带有来源

原结论

“原文准确率为75.83%。”

修订之后的

“教学封装在本地完整运行、移除指定元信息条件下，准确率为75.83%。”

代码来自官方，不代表本地数值就是作者历史报告值。

原文报告、本地官方代码输出、本地实验结果，应分别标注



从运行记录找到一项结果

运行编号: full-without_metadata20260922115534

训练样本: 2034
测试样本: 1353
正确预测: 1026
准确率: 0.7583148559

样本身份	真实类别	预测类别
sci.space/62410	太空	太空
...	...	...

有效配置 → 1353条预测 → 1026条正确 → 75.83%
运行编号与样本身份, 让数字可以核对

从汇总数字返回对应运行, 再定位到实际预测记录



复现结果比较：1 先写清这次比较要回答什么

目标A：实现一致性比较

1. 相同条件、流程
2. 官方脚本与本次运行的预测结果

比较：是否逐条一致？

目标B：处理条件比较

1. 固定任务与指定参数
2. 保留与移除元信息时

比较：准确率及错误分布如何变化？

注意：目标不同，所需的比较条件和证据也不同，取决于论文所属领域、不同流程等

复用已填写的比较条件矩阵

核对项	保留元信息	移除指定元信息
类别、样本与划分	同一四类别；2034训练 / 1353测试	相同ID集合与真实标签
处理规则	remove = []	headers、footers、quotes
向量器参数	sublinear_tf=True; max_df=0.5 min_df=5; 英文停用词	同左
分类器参数	tol=0.01; solver=sparse_cg	同左，固定版本默认值
评价方式	同一测试集上的准确率	同左

主动改变处理规则；词表与特征维度是随之改变的下游结果

来源：L1、L2: config_effective.json; 完整参数见讲者备注



改变不等于实验不可比

实现一致性比较

目标：核对两套执行是否一致

操作：remove不同会改变输入条件

应先对齐，再比较预测结果

处理条件比较

目标：研究处理规则的影响

操作：主动改变remove的输入

同时说明哪些条件保持不变

可比不等于一切相同，而是改变一些东西可以回答不同问题



示例：区分指标与不同计算口径

Accuracy

全部样本中，预测正确的比例

真\预	A	G	S	R
A	13	1	3	13
G	2	24	1	3
S	1	5	17	7
R	8	3	3	16

Precision

预测为这一类的样本中，有多少正确？（看混淆矩阵列）

Recall

真实属于这一类的样本中，有多少找对了？（看混淆矩阵行）

示例：真实图形学389条，正确347条； $\text{Recall} = 347 / 389$ 。

先分清分母与评价范围，再比较指标数值



示例： 区分不同处理方法

示例：（模拟场景）二分类100条样本

类别	真实样本数	正确数	召回率
甲类	90	90	100%
乙类	10	0	0%

准确率： $90 / 100 = 90\%$

宏平均召回率：
 $(100\% + 0\%) / 2 = 50\%$

准确率按样本计数；宏平均召回率让每个类别等权



示例： 区分不同数据带来的影响

保留元信息

$$1208 / 1353 = 89.28\%$$

移除指定元信息

$$1026 / 1353 = 75.83\%$$

样本身份与真实标签 → 逐条正确判断 → 正确总数 → 准确率
预测少一条时，应解释缺失

汇总指标应能追溯到对应的逐条输出



比较：百分点与相对变化

准确率之差

下降 13.45 个百分点

相对下降（以保留元信息条件为基准的）

相对下降约 15.07%

百分点： $(1208 - 1026) / 1353 \times 100\%$

相对下降： $(1208 - 1026) / 1208 \times 100\%$

结论：移除指定元信息对结果精度劣化比较严重



复现结果比较：2 用结果对照表的整理

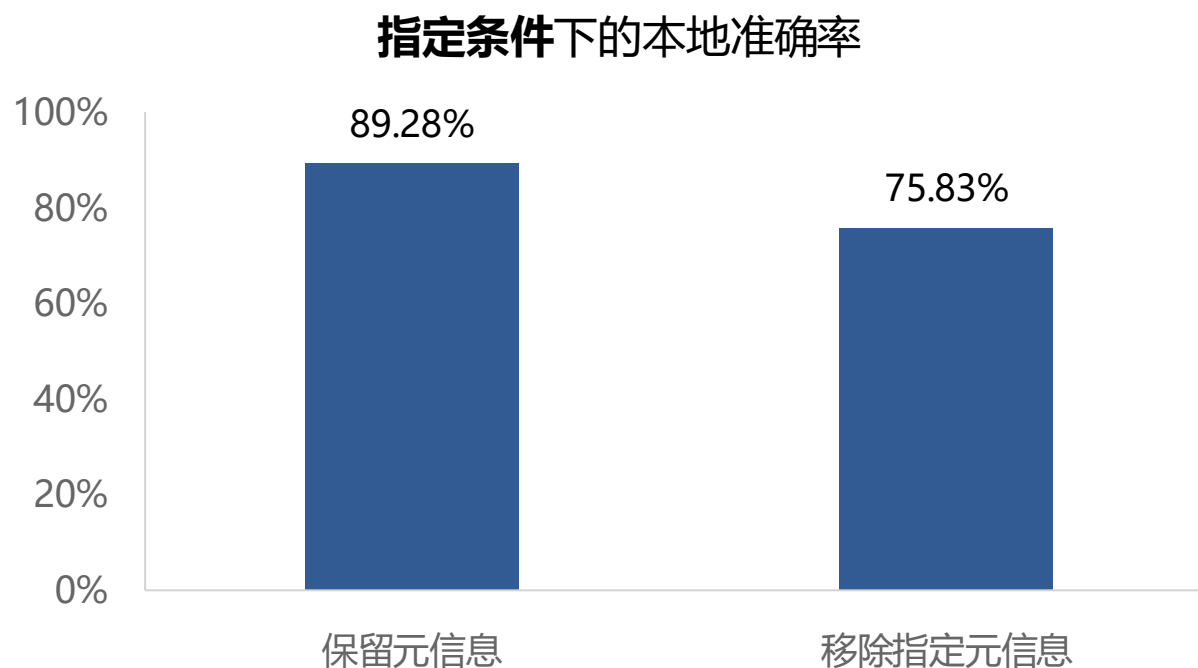
实验条件	结果来源	测试数	正确数	准确率
保留元信息	本地完整运行	1353	1208	89.28%
移除指定元信息	本地完整运行	1353	1026	75.83%

表注：同一四类别、既有划分；条件名称、来源和百分号均须明确。
完整配置可定位，不必把全部参数塞进表格

注意：表格不仅呈现数字，还应保留理解数字的必要条件



图表比较分析的细节信息



不当写法：过于简单

“原方法 / 改进方法”
本例改变的是处理条件。

应当描述细节：图注保留

同一四类别与1353条测试文本
两条件各一次完整运行
不添加未经测量的误差条

条件标签和图注是结果的一部分，不只是图表装饰



复现结果比较：3 单次与重复运行

重复时改变了什么？

1. 相同输入的确定性重跑？
2. 随机初始化或随机划分？

本例的记录信息

两条件各一次完整运行

打乱顺序不等于重新划分

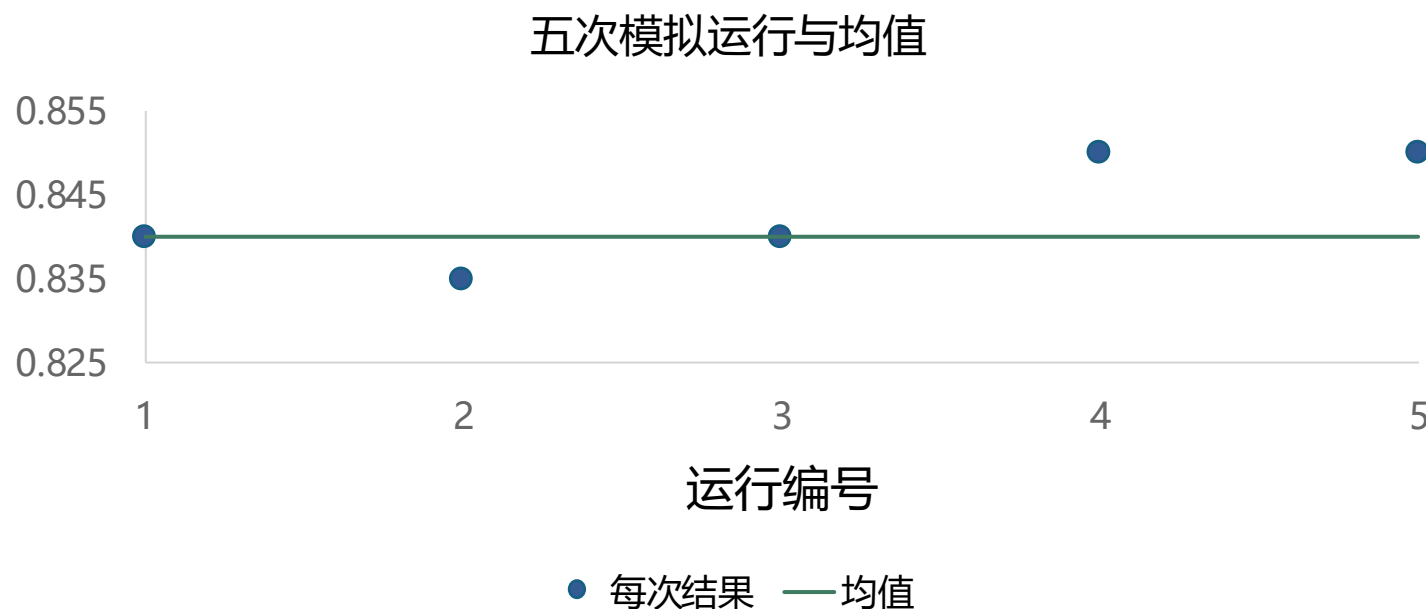
不添加未经测量的误差评测

重复应对应明确的问题，重复次数自己指定



为什么不能只报告最好的一次

模拟数据：五次有效运行



最佳一次
0.850

全部运行均值
0.840

样本标准差
约0.0079

报告全部有效运行，并说明汇总方式



复现结果比较：4 多次运行结果的分析

标准差

描述运行结果的分散程度

模拟：5次，均值0.840

样本标准差约0.0079

置信区间

表达指定估计量的不确定性

需说明抽样或随机来源

误差应标明含义、运行次数与随机来源



总结：结果说明的逐步改进

原笼统表述

“结果比较接近，复现成功。”

补充事实

$M=0.824$ ，原值 0.842 ，差 0.018 ；基线 $B=0.790$ 。

补充对照判据

未满足绝对差 ≤ 0.010 ，但 $M > B$ 方向保持；限于本次观察。

具体写清满足什么，比使用笼统的成功标签更有信息



Take away

来源

数字来自哪里，能否定位？

条件

比较所需的关键条件是否成立？

目标

判断是否对应原定目标与判据？

产出有条件说明的结果对照表，再调查差异如何产生



偏差解释：描述差异与解释原因

观察到的现象

准确率下降13.45个百分点

由对应正确数直接计算

回答：发生了什么？

候选解释（推测）

某类元信息帮助了分类

需要检查内容、过滤或对照

回答：为什么会这样？

把看到的现象与提出的解释分开



偏差解释： 1 实验过程检查-从低成本检查开始



先排除低成本可查的问题；有直接线索时优先处理



实验错误的案例：预测条数相同仍可能算错

预测本身正确，但文件行顺序错了导致错误

真实标签表	类别
样本a	太空
样本b	图形学

预测文件	预测类别
样本b	图形学
样本a	太空

按位置比较：0 / 2

按ID对齐：2 / 2

数量相同，不等于每一行指向同一样本



怎么检查最便捷？ 实际条件的三方核对

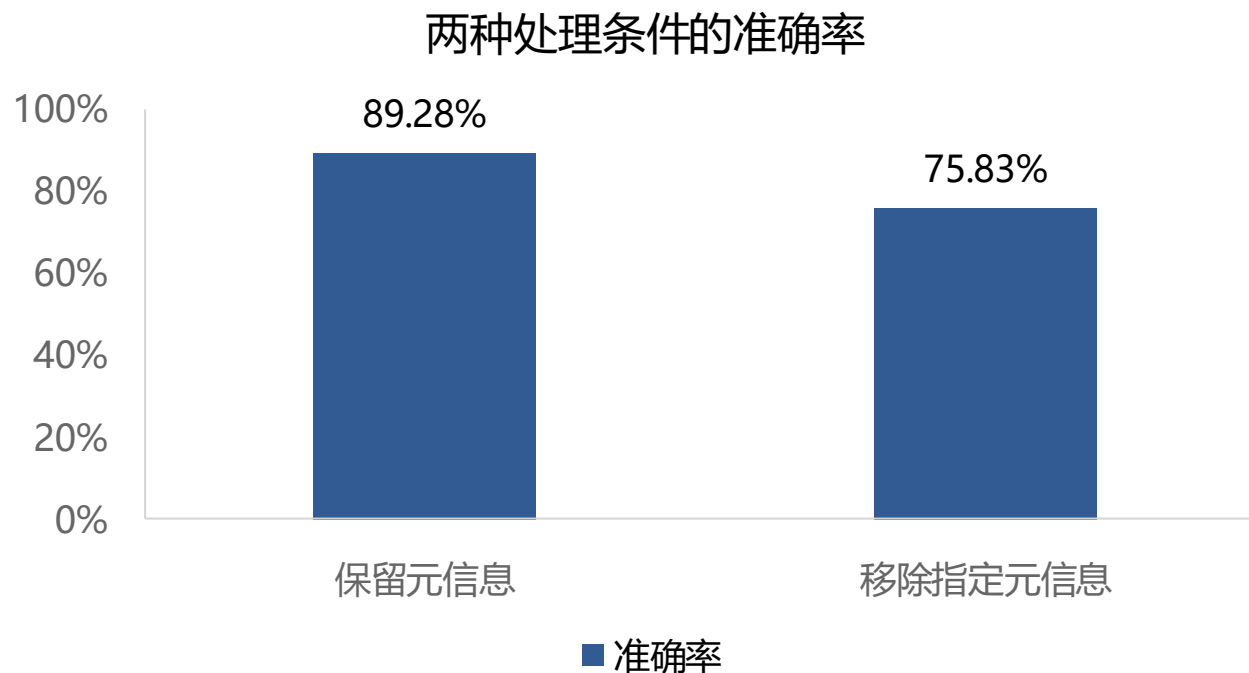
核对对象	原文或文档	参考代码	本次执行
元信息处理	描述过滤范围	检查remove参数	查看有效配置
特征拟合	训练 / 测试角色	检查fit与transform	核对实际输入
分类与评价	方法与指标定义	参数和评价函数	预测与指标记录

方案写“移除”不代表已执行；文件名也不能代替有效配置。

把解释落实到具体设置与代码行为



示例：两种数据处理条件的结果



保持不变的配置

同一四类别与样本划分
同一向量器与分类器参数

改变的配置

文本内容、词表与特征维度
7831维 → 5316维

原始数据：1208 → 1026，净少182条正确预测



先读懂一张混淆矩阵

真 / 预	无神论	图形学	太空	宗教杂谈
无神论	194	16	45	64
图形学	7	347	32	3
太空	22	22	343	7
宗教杂谈	64	15	30	142

行是真实，列是预测

真实太空394条
343条预测正确
22 + 22 + 7条分错

对角线正确数之和
 $194 + 347 + 343 + 142$
 $= 1026$

先说明行列含义，再解释数字，不凭颜色猜测结论

混淆矩阵对比

保留元信息

真 / 预	无神论	图形学	太空	宗教杂谈
无神论	258	7	12	42
图形学	2	380	4	3
太空	1	22	371	0
宗教杂谈	37	9	6	199

移除指定元信息

真 / 预	无神论	图形学	太空	宗教杂谈
无神论	194	16	45	64
图形学	7	347	32	3
太空	22	22	343	7
宗教杂谈	64	15	30	142

同一色标：0–400条；行是真实类别，列是预测类别。
无神论正确数258 → 194；宗教杂谈正确数199 → 142。

总指标发现差异，分项结果帮助定位差异集中在哪里



混淆矩阵对比：比较类别内的统一度量指标

真实类别	样本数	保留时召回率	移除后召回率	下降百分点
无神论	319	80.88%	60.82%	20.06
图形学	389	97.69%	89.20%	8.48
太空	394	94.16%	87.06%	7.11
宗教杂谈	251	79.28%	56.57%	22.71

例：无神论258 / 319 $\rightarrow$ 194 / 319；计数看规模，比例看类别内部

选择计数还是比例，取决于想回答的问题



偏差解释： 2 同批次样本的状态变化分析

正确状态配对	移除后正确	移除后错误	合计
保留时正确	989	219	1208
保留时错误	37	108	145
合计	1026	327	1353

219条由对变错，37条由错变对；净差 = $219 - 37 = 182$ 。

总体下降并不意味着每条样本都变差



示例：两条真实预测变化记录

样本身份	真实类别	保留时预测	移除后预测
test/sci.space/61501	太空	太空：正确	无神论：错误
test/alt.atheism/53473	无神论	宗教杂谈：错误	无神论：正确

能否据此说“第一条因为删除发件人而变错”？
仅仅是单个数据

预测变化是解释线索，不等于已经确认某个字段的作用



个别样本能提供线索，但不能代表全体

全部记录

说明变化规模：219条由对变错，37条由错变对

分项统计

定位类别与错误方向，不只看总准确率

样本阅读

按预定规则分层取样，检查实际内容与处理变化

前页两条只示范不同变化方向，不称为代表性样本。

不能只挑最符合自己解释的文本



现有证据能排查到哪一步

保留元信息

官方脚本与实际运行逐条一致

移除元信息

官方脚本与实际运行逐条一致

核验范围

冻结：版本、选定Ridge流程、两种指定条件

不能据此证明全部代码无错、所有环境一致，或元信息作用机制已查明。

核验范围应与结论范围保持一致

来源：verification/upstream/checks.json：两条件prediction_matches均为true



偏差解释： 3 将解释以补充实验形式描述

邮件头过滤是否是重要影响因素？

设计项	具体计划
保持项	四类别、既有划分、向量器与分类器参数
主动变化	remove=[] 对照 remove=['headers']
拟合规则	各自仅用训练文本拟合向量器
记录输出	准确率、逐条预测、特征维度、实际过滤内容

headers不等于“只有邮箱”；其他过滤项与交互仍需检查



补充实验的不同结果如何解读

明显下降

支持进一步调查邮件头处理，不能归结为全部由邮箱导致

变化很小

检查其他过滤项与联合作用，不能断言邮件头永远无关

未实际执行

先修正执行或记录问题，暂不能评价候选解释

先想清不同结果意味着什么，再决定补充实验是否值得做



偏差解释： 4 随机性与环境

未经核查的解释

“应该是随机种子不同。”

“可能是电脑不同。”

没有指出具体变化与证据

可以继续验证的解释

1. 指出具体变化：例如随机划分
2. 同一划分比较，或研究划分波动
3. 计时则核对硬件、缓存与测量范围

环境与随机性不能成为未经检查的通用解释



敏感性分析与追逐目标数值

敏感性分析

检查min_df取3、5、10

说明范围依据，保留全部结果

考察方向是否依赖某个阈值

事后新增分析标为探索性；不反复用测试集挑选最终模型

来源：教学情境：参数分析尚未执行，无实测结果

科研素养训练II·复现结果比较与偏差解释



偏差解释： 5 偏差解释记录表

记录项	本例内容
观察	下降13.45个百分点；净少182条正确预测
候选解释	指定处理规则可能改变分类线索与词表
已做检查	样本ID、真实标签、计数、配置及参考实现一致性
得到证据	样本可配对、计数相符、选定流程预测一致
限制与下一步	尚未分离各类元信息作用；拟作分项过滤对照

配置不同不等于已解释全部下降；计划不能写成实验证据



课堂练习：选择有依据的解释

假设：预处理不同，环境版本相同

A

原论文错误。

B

差异全部来自预处理。

C

预处理差异是待检验的解释，应对齐或做配套对照。



解释修正：从过度结论改为有限结论

原句

“文本分类都在利用邮箱作弊，删除邮箱导致下降13.45%。”

修订

在指定四类别、既有划分及配置下，移除邮件头、签名尾部与引用内容后，准确率由89.28%降至75.83%，下降13.45个百分点。

修正外推范围、具体因素、单位与行为判断



偏差解释： 6 尚未完成复现时如何报告

工作状态	应提交的材料
运行受阻	阻塞位置、已有尝试与失败记录
可比实验完成，目标未获支持	实际结果、判据与诊断证据
关键条件未知，无法判断	缺失条件、影响及补充计划

例：未取得测试标签，暂不能计算指标；提交已有预测、缺口及下一步。

准确报告工作状态，比强行给出成功或失败更重要



作业4的交付要求

以前三次作业和实际运行材料为基础，提交四部分

1. 目标与实际条件对照
2. 带来源的结果对照表或图
3. 偏差检查及对应证据
4. 有限结论与后续计划

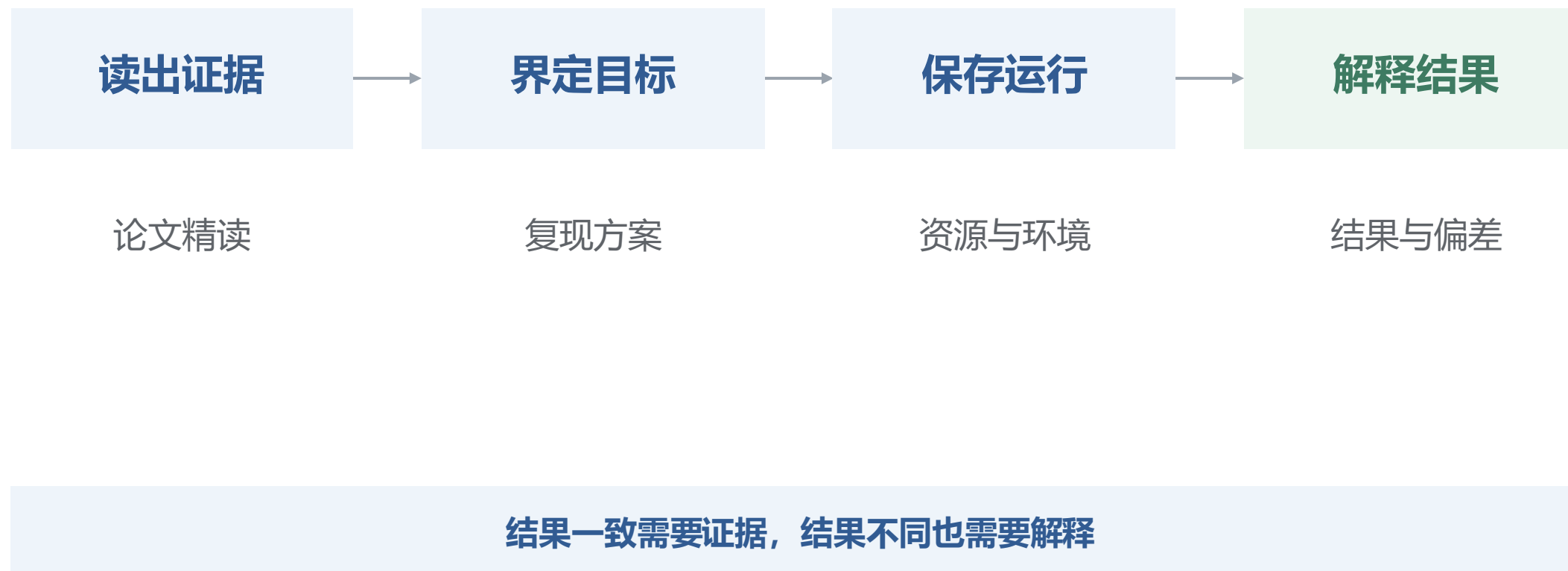
尚未完成正式运行：
提交已有证据与缺口，
明确哪些目标暂不能回答。
不得以模拟数据替代实测。

使用大模型时，附模型用途及完整提示词记录（含追问与修改指令）。

评价比较依据、诊断证据与结论边界，不以指标高低评分



四次训练的共同要求





Project汇报的参考顺序

论文总结

验证什么

如何执行

得到什么

检查了什么

能说明什么

阅读卡与方案

实际运行记录

对照表与图

诊断证据

结论与限制



廈門大學
XIAMEN UNIVERSITY

Thank You!