# 科研素养训练II

## 第1次课：论文复现导论与论文阅读

从结构化阅读走向面向复现的论文精读

宋庆禹

2026.09.07

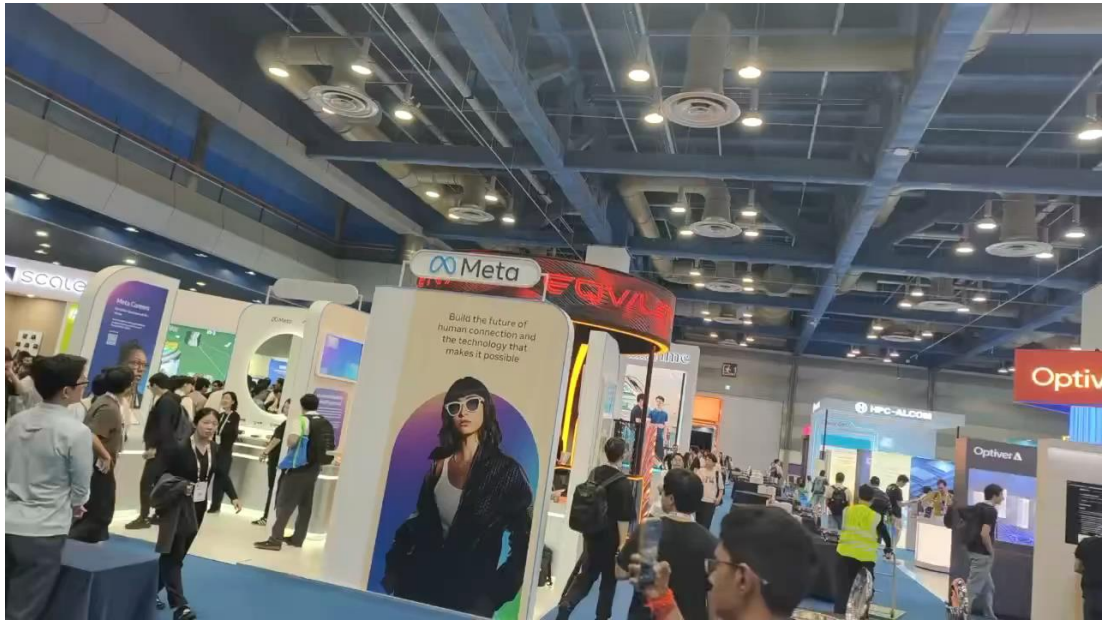
厦門大學
XIAMEN UNIVERSITY

群聊：2026秋-科研素养训练II
课程群

该二维码7天内(9月13日前)有效，重新进入将更新

# ICML 2026 7.5-7.9 韩国 首尔






# ZO-based memory-efficient LLM fine-tuning

## AI for Optimization · LLM fine-tuning

- **Background:** Memory-efficient LLM fine-tuning algorithm on resource constrained scenarios (on-device or hardware specific)
- **Motivation:** Zeroth-order (ZO) optimization
  - Finite difference for gradient estimation avoids storing backpropagation activations
  - Isotropic perturbations waste queries in high dimensions
- **Design:** Biased sampling
  - Derive a compact subspace and restrict perturbations to the space
  - Low-rank subspace: activation-informed during the forward pass

**Activation structure → Guided low-rank perturbations → Sharper gradient estimates**


*Table 5.* Additional Qwen3-8B results.

| Method | OpenMath | SST-2 | COPA | CB |
|---|---|---|---|---|
| AGZO | **0.640** | **0.932** | **0.820** | **0.928** |
| MeZO | 0.560 | 0.916 | 0.800 | 0.910 |
| LOZO | 0.580 | 0.907 | 0.790 | 0.910 |

~10% benchmark improve

- **Achievement**:
  - Optimality: *Tighter gradient estimation bound than SOTA ZO methods*
  - End-to-end performance: *Up to ~10% improvement than SOTA ZO methods*

# 本次课内容

01 **课程定位与整体安排** 科研素养训练II如何衔接前序课程

02 **论文复现导论** 为什么复现，以及复现任务如何分类

03 **面向复现的论文精读** 从论文陈述中追踪研究证据

04 **论文阅读卡** 将可验证信息转化为后续复现入口

# 课程安排

**8 周**

每周一 7–8 节
共 16 学时

## 课程考核

| | | |
|---|---|---|
| **出勤** | **10%** | 课堂参与 |
| **4次作业** | **40%** | 每次10% |
| **Project** | **50%** | 4人一组，组内同分 |

**Project 汇报包含论文理解与复现结果**

# 主讲教师

## 宋庆禹

**研究方向**

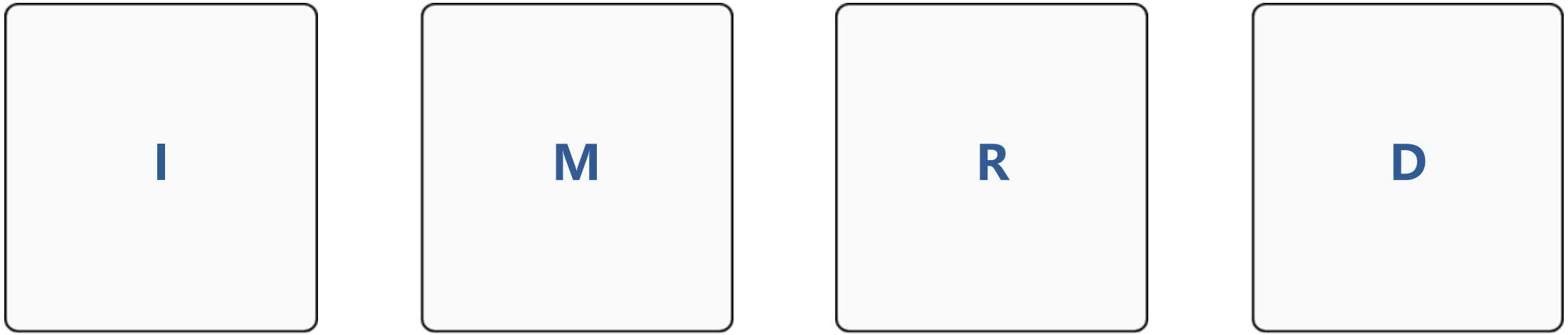
**智能优化算法**
**大语言模型训练与推理优化**

| | |
|---|---|
| 香港中文大学 | **博士** |
| 清华大学 | **硕士** |
| 哈尔滨工业大学（威海） | **学士** |

https://simmonssong.github.io

# 本次课程的学习定位

## IMRaD 回顾




**Introduction**

研究动机与问题

**Methods**

研究如何实施

**Results**

研究发现什么

**Discussion**

发现意味着什么

**本课从论文结构继续向下追踪：每个结论如何由可验证证据支撑**

# 本次课程的学习定位

## 已有基础

- 理解 IMRaD 结构
- 能够概括研究贡献
- 能够定位论文各部分

## 本次训练

- 拆解研究证据
- 提取可验证信息
- 建立复现任务入口

1 研究问题 → 2 方法设计 → 3 数据来源 → 4 实验设置 → 5 评价指标 → 6 结果呈现 → 7 结论解释

# 科研素养训练 I 与 II 的衔接

## 科研素养训练 I

- 理解论文结构
- 掌握 IMRaD 阅读
- 概括研究贡献

## 科研素养训练 II

- 拆解研究证据
- 判断可复现性
- 开展复现实践

**学习目标从“读懂论文”推进到“验证论文”**

# 课程整体结构与学习路径

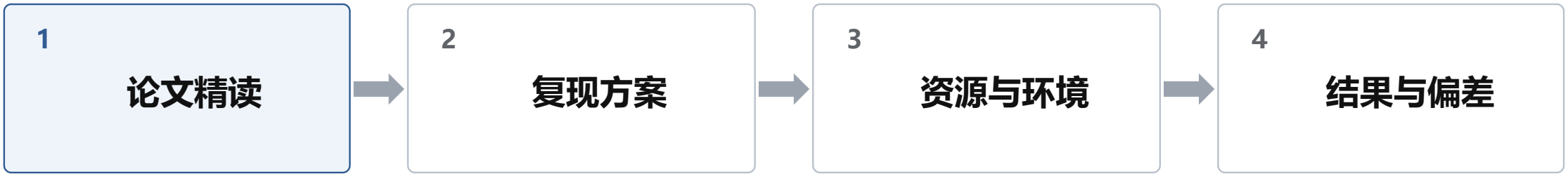
**前四节课**　　围绕一篇论文完成复现实践闭环

1 论文精读 → 2 复现方案 → 3 资源与环境 → 4 结果与偏差

**后四节课**　　科研素养专题讲座与分组汇报

**课程以小组 Project 串联阅读、方案、执行与汇报**

# 前四节课：论文复现实践主线




论文阅读卡　　复现方案书　　可运行实验包　　结果分析报告

**四次课共同完成一次完整、可审计的复现训练**

# 后四节课：科研素养专题模块

厦門大學
XIAMEN UNIVERSITY

05

**论文选题汇报**

说明研究问题与复现价值

06

**专题讲座 1**

拓展研究方法与科研视野

07

**专题讲座 2**

讨论开放科学与研究规范

08

**复现结果汇报**

呈现证据、偏差与反思

**专题讲座服务于 Project 选题、实施与反思**

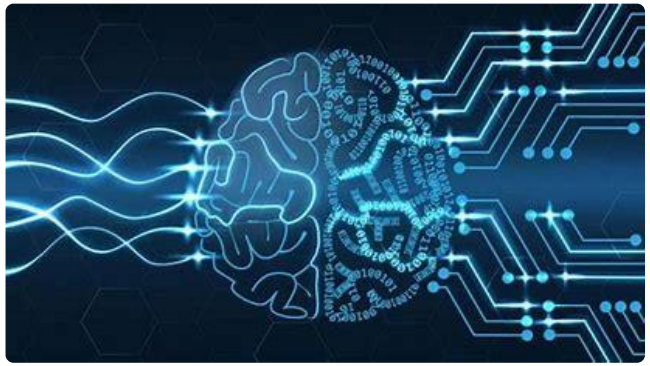
科研素养训练II

# 论文复现导论与论文阅读

从结构化阅读走向面向复现的论文精读

宋庆禹 · 2026.09.07

# 论文可复现性的价值



## 教育价值

掌握技能，训练批判性思维

## 科研价值

检验结论，发现新的研究问题


## 工业价值

评估方法能否迁移到实际系统

**复现训练把论文中的知识转化为能够检查和执行的研究能力**

# 你有多相信一篇论文的结果？

**如果你是作者，你会如何获取对比方法的实验结果？**

## A 直接引用原文结果

- 实验条件可能与本研究不一致
- 单个数字未必代表稳定表现
- 只能确认来源，不能验证过程

## B 自己重新运行实验

- 统一数据、指标与代码版本
- 获得多次运行的均值与波动
- 形成可以审计的实验记录

**A 可以规范引用；B 能提供更强的可比证据**

# 结果引用与学术诚信

## 学术不端

- 把原论文数字写成“我们的实验结果”
- 声称重新运行，实际并未运行
- 不注明来源，制造结果由本研究产生的印象
- 只引用有利数字并隐瞒条件差异

## 合规做法

- 明确标注“原论文报告结果”（reported results）
- 规范引用结果来源
- 如实说明是否重新运行
- 披露实验条件与指标差异

**判断标准：来源是否透明，研究过程是否真实**

# 论文复现的必要性

**复现帮助我们判断一项研究能否进入可查证、可执行、可验证的证据链**


**复现既不是机械重复，也不是专门寻找原文错误**

# 复现结果示例：百项心理学研究统计结果

Reproducibility Project: Psychology（2015）

**97%**

原研究报告了
统计显著结果

**36%**

复现研究得到
统计显著结果

**复现重新评估结论的证据强度，而不是给论文简单判定对错**

资料来源：Open Science Collaboration (2015)

# 课程复现既能成功，也能产生有价值的失败

Stanford CS244: Reproducing Network Research

**200**

五年中参与课程的学生

**16 / 18**

2012年18个小组中
16个至少复现一项结果

**4**

4个小组在复现基础上
获得进一步结果

**交叉验证中，另一组均能在1天内重现课程实验，多数少于1小时**

资料来源：Heller et al. (2017)

# 复现结果不一致并不等于原研究错误

| 01 | 02 | 03 | 04 | 05 |
| --- | --- | --- | --- | --- |
| 抽样波动 | 统计功效不足 | 场景或样本变化 | 方法报告不完整 | 分析选择不同 |

**先比较哪些要素保持一致，再判断差异可能来自哪里**

**一次未复现成功只是调查的起点**

资料来源：LMU Open Science Center

# 可信研究依赖可靠性与可复现性

## 可靠性

- 事前规划研究设计
- 明确假设与判断标准
- 采用适当的设计与分析

+

## 可复现性

- 数据与材料可获得
- 代码和文档完整
- 环境规格描述明确

**两者共同提高可复现性**

资料来源：LMU Open Science Center

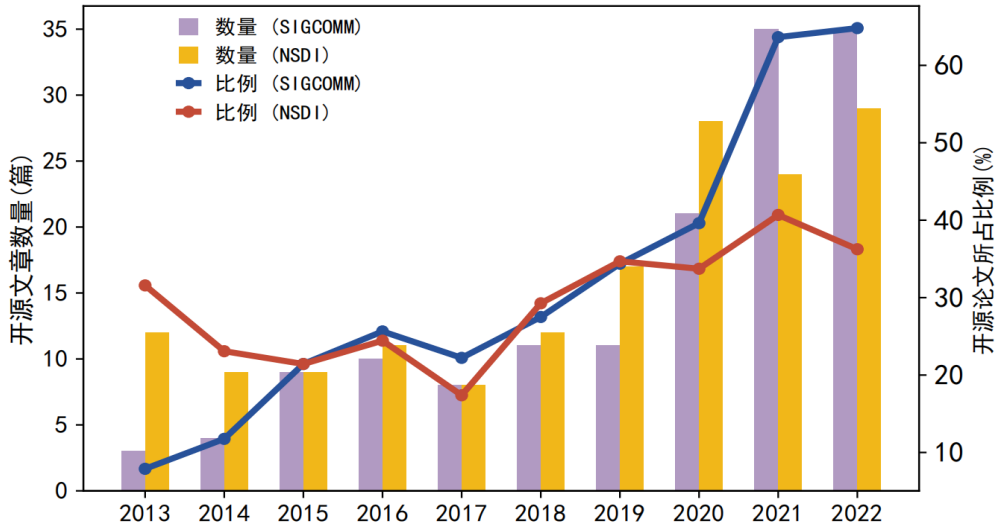
# 本课所说的“论文复现”

| 类型 | 数据与代码 | 主要目的 |
| --- | --- | --- |
| 结果重算 | 原数据 + 原代码 | 核对论文报告数字 |
| 独立实现 | 原数据 + 新代码 | 检验方法描述是否充分 |
| 直接复现 | 新数据 + 近似原方法 | 检验结果稳定性 |
| 概念复现 | 新数据 + 不同操作化 | 检验结论一般性 |

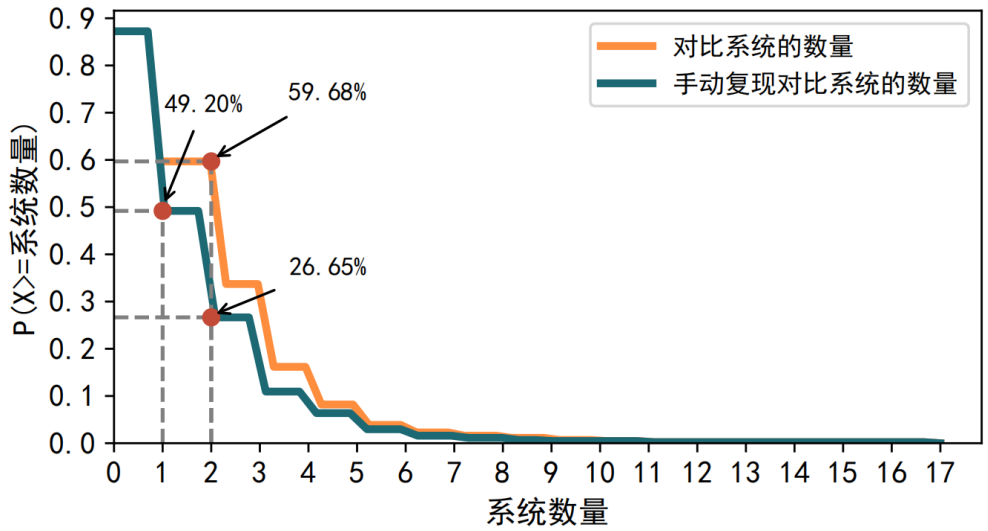
**术语在不同学科中并不完全统一，本课依据数据、代码和方法的变化判断**

资料来源：LMU Open Science Center

# 案例：计算机网络研究中的代码复现




**2013–2022** SIGCOMM / NSDI 论文中开源代码比例平均约 32% / 29%




**2.29** 平均每篇论文需要手动复现的网络系统数量

资料来源：原课件所列网络系统复现统计

# 复现类型应由你的研究目的决定

厦門大學 XIAMEN UNIVERSITY

01

## 核验型

确认原文结论能否重新得到

02

## 实现型

依据论文描述独立完成方法

03

## 稳健型

改变种子、样本或合理参数

04

## 拓展型

增加数据、任务或研究条件

**本课程以核验型或实现型为主**

资料来源：LMU Open Science Center

# 同一篇论文可以形成四种复现任务

| | | | |
|---|---|---|---|
| **核验型** | **原代码 + 原工作输入** | 重跑 DCTCP 的关键图表 | 数字能否重新得到? |
| **实现型** | **依据论文描述复现代码** | 从论文重新实现网络系统 | 描述是否足以落地? |
| **稳健型** | **改变输入：种子、内核或链路速率等** | 比较不同环境下的趋势 | 结论是否依赖某个设置? |
| **拓展型** | **增加场景：拓扑、负载或其他实验场景等** | 将 Outcast 扩展到 Fat Tree | 结论能否推广? |

资料来源：Stanford CS244 课程案例

# 描述变化比贴类型标签更重要

| 复现维度 | 保持一致 | 必要调整 | 重新设计 |
| --- | --- | --- | --- |
| 研究问题 | 原问题 | 限定范围 | 新问题 |
| 数据或样本 | 原数据 | 重新划分 | 新数据 |
| 方法 | 原代码 | 版本适配 | 独立实现 |
| 分析与指标 | 原指标 | 等价实现 | 新增分析 |
| 软件与环境 | 原环境 | 兼容版本 | 新平台 |

**重新划分数据或更换依赖版本，都应明确记录**

资料来源：LMU Open Science Center

# 同一研究往往存在多条合理分析路径


**选择本身未必错误；未说明选择空间并只报告最有利路径，会降低论文可信度**

资料来源：LMU Open Science Center

# 不透明研究过程中的具体选择会削弱研究可信度

**p-hacking** —— 反复尝试分析，只报告显著的结果

**HARKing** —— 看到结果后，再把发现写成事前假设

Hypothesizing After the Results are Known

**选择性报告** —— 只展示最佳指标、最佳种子或成功实验

**诚实错误** —— 表格复制、脚本版本或数据处理出现失误

**精读的重点是辨认哪些信息已经写清楚，哪些仍不可查证**

资料来源：LMU Open Science Center

# 四类研究风险示例

## p-hacking

测试多个指标、协变量和停止时点，只报告 p < 0.05 的组合

## HARKing

先看到“带宽下降”，再把它写成事前研究假设

## 选择性报告

运行10个随机种子，只展示准确率最高的一次

## 诚实错误

脚本读错数据列或混淆 ms 与 s，导致结果偏离

**多种分析自由度叠加时，复现过程中的假阳性率可由 5% 升至 60.7%**

资料来源：Simmons, Nelson & Simonsohn (2011)

# 课堂判断：哪些做法存在复现风险？

**A** —— 运行5个随机种子，只报告最高一次结果

**B** —— 删除异常样本，但未说明阈值及删除数量

**C** —— 翻译实验材料，并完整记录翻译与预测试过程

**判断为透明适配、可疑选择或信息不足，并说明还需要什么信息**

资料来源：LMU Open Science Center

# 论文复现对科研训练的价值

- 数据来源是否明确
- 评价指标是否清楚
- 实验参数是否可追踪
- 结论是否由实验充分支撑


**复现训练把“相信论文”转化为“检查论文”**

# 从结论表述到复现追问

**论文表述**

> **“显著优于已有方法”**

**复现追问**

- **优于谁?**
- **在哪个数据集?**
- **使用什么指标?**
- **差异有多大?**

**复现问题必须把比较对象、条件、指标和判断标准说清楚**

# 本次课程的核心问题

| | | |
|---|---|---|
| 01 | **研究问题是什么?** | 论文试图解释或解决什么问题 |
| 02 | **核心方法是什么?** | 哪些步骤产生了关键结果 |
| 03 | **支撑结论的证据是什么?** | 结论对应哪张图、哪张表或哪项检验 |
| 04 | **哪些因素可能影响复现?** | 数据、参数、环境和分析选择是否清楚 |

1 问题 → 2 方法 → 3 数据 → 4 配置 → 5 指标 → 6 结果 → 7 结论

# 三种论文阅读层次

| | | | |
|---|---|---|---|
| 01 | 普通阅读 | **知道论文讲什么** | 一句话复述主结论 |
| 02 | 结构化阅读 | **知道各部分如何组织** | 定位方法、数据与指标 |
| 03 | 面向复现的精读 | **知道结论如何被证据支撑** | 写出可以执行的核验步骤 |

**本课重点训练第三层：把阅读结果转化为复现线索**

# 面向复现的论文精读原则

## 低效做法

- 逐字翻译
- 摘抄摘要
- 全文划重点

## 有效做法

- 带着问题阅读
- 追踪证据位置
- 记录复现线索

**精读关注证据关系，而不是文字覆盖率**

# 论文精读中的关键追问

1

核心假设在哪里?

2

关键方法是什么?

3

实验依据是否充分?

4

哪些细节会影响复现?

5

哪些信息需要进一步查证?

**问题越具体，阅读记录越容易进入后续复现方案**

# 论文证据链的构成

1 研究问题 → 2 方法设计 → 3 数据来源 → 4 实验设置 → 5 评价指标 → 6 结果呈现 → 7 结论解释

**每个节点都对应一个阅读任务，也对应一个后续复现风险点**

精读的核心：沿着证据链检查信息是否完整、可执行、可比较

# 论文证据链示例


**结论能否成立，取决于前六个节点能否形成完整证据**

# 证据链缺失带来的复现风险

| | | |
|---|---|---|
| 数据来源不清 | → | 无法获得相同输入 |
| 实验设置不清 | → | 无法还原运行条件 |
| 指标定义不清 | → | 无法判断结果是否一致 |
| 关键参数缺失 | → | 无法解释复现偏差 |

**信息缺失并不自动否定结论，但会限制复现能够回答的问题**

# 可验证信息的定义

## 判断标准

- 能够被查证
- 能够被执行
- 能够被比较
- 能够被复现

## 操作化示例

- 数据集名称、版本与下载地址
- 执行命令、参数与随机种子
- 相同数据划分与指标定义
- 他人按步骤得到一致趋势

**可验证信息必须能够落到具体对象、操作和判断依据**

# 不可直接复现的模糊表述

| | | |
|---|---|---|
| 01 | **方法表现更好** | 更好多少？使用什么指标？ |
| 02 | **实验效果明显提升** | 提升体现在哪个数据集？ |
| 03 | **具有较强泛化能力** | 泛化到哪些对象和条件？ |
| 04 | **采用常规设置** | 常规设置具体是什么？ |
| 05 | **进行了充分实验** | 充分包括多少次实验？ |

**模糊表达需要继续追问，不能直接写入复现方案**

# 可用于复现的明确表述

| | | |
|---|---|---|
| 结果 | —— | F1 从 81.3 提升到 84.6 |
| 训练 | —— | 训练 50 epoch，batch size 为 32 |
| 数据 | —— | 数据划分比例为 8:1:1 |
| 消融 | —— | 去掉模块 A 后性能下降 2.4 |
| 资源 | —— | 代码仓库提供训练和测试脚本 |

**明确表述可以直接进入论文阅读卡与复现计划**

# 实验结果图表的精读方法

- 基线方法是谁
- 提升是否稳定
- 是否报告方差或置信区间
- 比较条件是否公平

| 方法 | Dataset A | Dataset B | 指标 | 阅读判断 |
| --- | --- | --- | --- | --- |
| Baseline | 81.3 | 79.5 | F1 | 原文对照 |
| Method X | 84.6 | 82.1 | F1 | 待复现目标 |
| 去掉模块 A | 82.2 | 80.0 | F1 | 消融线索 |

蓝色行为阅读时优先定位的复现目标

**优先定位能够支撑论文核心结论的图、表与实验设置**

# 论文复现中的常见风险来源

**代码** —— 未公开 / 入口不清

**数据** —— 不可下载 / 划分未知

**环境** —— 依赖过旧 / 版本冲突

**参数** —— 缺学习率 / 随机种子

**评价** —— 脚本不一致 / 指标模糊

**成本** —— 算力与时间过高

**12 / 40**

原系统代码
直接开源

**9 / 40**

实验工作负载
直接开源

资料来源：Stanford CS244 课程统计

# 面向复现的论文阅读卡

**论文阅读卡不是摘要，而是后续复现工作的入口**

01 论文信息

02 研究问题

03 核心方法

04 数据 / 代码

05 评价指标

06 主要结果

07 可验证信息

08 复现风险

09 待查证问题

**阅读卡同时记录明确证据、待查证问题与潜在风险**

# 论文阅读卡的核心栏目

| | | | |
|---|---|---|---|
| 论文基本信息 | 论文信息 | 研究问题 | 核心方法 |
| 复现实验信息 | 数据 / 代码 | 评价指标 | 主要结果 |
| 复现判断 | 可验证信息 | 复现风险 | 待查证问题 |

**前两组记录论文明确报告的内容，最后一组记录复现判断**

# 论文阅读卡示例

| 栏目 | 面向复现的具体记录 |
| --- | --- |
| 论文信息 | 某方法用于文本分类 |
| 研究问题 | 提高短文本语义表示能力 |
| 核心方法 | 预训练模型 + 对比学习目标 |
| 评价指标 | Accuracy、Macro-F1 |
| 主要结果 | 文本总结困惑度 |
| 可验证信息 | 数据集、主实验与部分参数 |
| 复现风险 | 代码完整性、数据划分与训练成本 |

**阅读卡记录的不是评价，而是后续能够执行与核查的信息**

# 作业 1：阅读论文并提取可验证信息

## 个人作业：选择一篇论文，提交论文阅读卡 v1

| | | |
|---|---|---|
| 01 | **填写阅读卡前半部分** | 论文信息、研究问题、核心方法 |
| 02 | **定位实验信息** | 数据来源、评价指标、主要结果 |
| 03 | **记录复现判断** | 可验证信息、复现风险、待查证问题 |

**提交物：一页阅读卡 + 原文中对应证据位置**
**注：如果你使用大模型辅助，请给出相应的模型以及完整提示词**

# 本次课程小结与课后任务

结构阅读 → 证据链分析

贡献概括 → 可验证信息提取

论文理解 → 复现准备

**课后任务：提交论文阅读卡 v1**

| 1 问题 | 2 方法 | 3 数据 | 4 设置 | 5 指标 | 6 结果 | 7 结论 |
|---|---|---|---|---|---|---|

厦門大學
XIAMEN UNIVERSITY

群聊：2026秋-科研素养训练II
课程群

该二维码7天内(9月13日前)有效，重新进入将更新