大模型推理系统与实践

第 10 讲论文比较方法与实验可信性

张书豪

华中科技大学计算机学院

2027 · 研究生课程课件

教学目标

核心结论

本讲训练系统论文阅读能力，重点在于从论文中提炼真实研究问题与比较维度。

- 识别瓶颈假设和控制点变化。
- 分析机制代价与适用边界。
- 判断实验是否足以支撑结论。

系统论文的阅读重点

核心结论

阅读系统论文时，首要任务是重建问题定义而非简单记忆图表。

- 论文究竟重新定义了什么问题？
- 改变了哪个控制点或状态对象？
- 引入了哪些额外代价？

五段式精读框架

核心结论

一个清晰的精读报告至少应包含五个部分。

1. 问题假设
2. 核心机制
3. 额外代价
4. 适用边界
5. 实验支撑度

什么叫“问题假设”

核心结论

很多学生读论文只看作者做了什么，却忽略了作者默认哪些前提成立。

- workload 是否是在线动态到达。
- 瓶颈是否真的在作者声称的部分。
- 硬件、模型和系统配置是否会改变结论。

忽略问题假设的风险

核心结论

曲线和表格通常建立在特定假设、工作负载与平台条件之上，脱离前提容易导致误判。

- 看似很大的加速可能只在窄场景成立。
- 看似中性的结果可能在更公平基线下并不成立。
- 论文真正的贡献常常在问题重构而不是数值本身。

如何识别真正的机制贡献

核心结论

所谓“机制贡献”，应当表现为控制点、状态组织或执行路径被实质性改变。

- 只是换了术语，不算机制。
- 只是重新组合已有做法，也未必构成实质贡献。
- 真正有价值的机制应当改变系统行为边界。

如何判断实验支撑度

核心结论

实验的任务在于论证问题假设与机制收益是否成立。

- workload 是否贴合目标场景。
- 对照组是否公平。
- 指标是否覆盖端到端效果与关键探针。

讲解：论文比较方法课真正训练的能力

核心结论

这讲不是教学生复述论文摘要，而是训练把论文拆成“问题、机制、代价、边界、证据”的能力。

- 先问作者到底重构了哪个系统问题，而不是先看图表大不大。
- 再问机制改变了什么控制点，为什么这个点值得改。
- 最后问实验到底支撑了哪些结论，还有哪些只是作者的延伸解释。

讲解：为什么系统论文不能只比“数字是否更高”

核心结论

论文比较的核心不是找一个冠军，而是判断一篇工作是否真的把问题讲清、做实、证实。

- 工程量很大，不等于研究判断就清楚。
- 加速很明显，不等于对照就公平、边界就被说明。
- 所以课程希望学生形成的不是“排名思维”，而是“证据与边界思维”。

例子：同一篇系统论文应该先看哪里（1）

核心结论

论文比较方法课不是教学生背结论，而是教他们迅速识别论文的真正控制点和证据边界。

第一步：找问题定义

这篇论文到底在解决哪一个系统对象的什么痛点？

例子：同一篇系统论文应该先看哪里（2）

核心结论

论文比较方法课不是教学生背结论，而是教他们迅速识别论文的真正控制点和证据边界。

第二步：找机制变化

它改了控制点、状态组织、执行路径，还是实验方法本身？

例子：同一篇系统论文应该先看哪里（3）

核心结论

论文比较方法课不是教学生背结论，而是教他们迅速识别论文的真正控制点和证据边界。

第三步：找证据边界

它的实验到底支撑了哪部分结论，还有哪部分只是推测？

例子：把 vLLM 论文拆成五段会看到什么（1）

核心结论

经典论文最好拿来做结构拆解，因为它能帮助学生建立统一的阅读框架。

假设

动态 serving 的关键瓶颈之一是 KV 管理方式不支持高效 continuous batching。

例子：把 vLLM 论文拆成五段会看到什么（2）

核心结论

经典论文最好拿来做结构拆解，因为它能帮助学生建立统一的阅读框架。

机制

用 PagedAttention 改写状态组织方式，而不是只调度得更频繁。

例子：把 vLLM 论文拆成五段会看到什么（3）

核心结论

经典论文最好拿来做结构拆解，因为它能帮助学生建立统一的阅读框架。

边界

论文证明了什么条件下收益成立，但不等于所有 serving 问题都因此解决。

代码例子：把论文机制映射回真实代码（1）

核心结论

论文阅读真正有价值的时刻，是你能指出机制落到代码里大概对应哪类对象。

代码映射示例

```
BlockManager / block_table -> KV organization
Scheduler -> control point
bench.py / scripts -> evidence path
```

代码例子：把论文机制映射回真实代码（2）

核心结论

论文阅读真正有价值的时刻，是你能指出机制落到代码里大概对应哪类对象。

课堂应追问

如果一篇论文的核心机制在代码里找不到等价控制点，你该如何判断它是否被真实实现？

代码例子：把论文机制映射回真实代码（3）

核心结论

论文阅读真正有价值的时刻，是你能指出机制落到代码里大概对应哪类对象。

课程结论

论文比较不能停在图表和术语层，必须回到对象、控制点和可验证实现。

代码例子：系统论文精读报告模板（1）

核心结论

这类课程最好给学生一份“可直接套用但不偷懒”的读论文结构模板。

模板片段

```
1. Problem and system object
2. Mechanism and changed control point
3. Cost / boundary / hidden assumptions
4. Evidence and what is still unsupported
```

代码例子：系统论文精读报告模板（2）

核心结论

这类课程最好给学生一份“可直接套用但不偷懒”的读论文结构模板。

课堂应追问

为什么“论文摘要改写一遍”不能算合格精读报告？

代码例子：系统论文精读报告模板（3）

核心结论

这类课程最好给学生一份“可直接套用但不偷懒”的读论文结构模板。

教学意义

模板的作用不是降低难度，而是强制学生用系统语言组织判断。

例子：什么叫 synthetic gain（1）

核心结论

课程里必须训练学生识别“图很好看，但比较并不公平”的论文叙事。

迹象一

baseline 选择过弱，或者没有把现有系统最佳实践纳入比较。

例子：什么叫 synthetic gain（2）

核心结论

课程里必须训练学生识别“图很好看，但比较并不公平”的论文叙事。

迹象二

workload 过于单一，恰好只适合提出的方法。

例子：什么叫 synthetic gain（3）

核心结论

课程里必须训练学生识别“图很好看，但比较并不公平”的论文叙事。

迹象三

只给出端到端大图，不给局部探针与失败条件，导致因果链断裂。

课程研究案例（内部实验）：主张强度

核心结论

系统论文写作不是把结果写满，而是把 claim 压到证据能够承受的范围。

- KV-Delta 可说：repeated movement 下减少重复 KV image，并在 m 增大时改善 E2E。
- 不能直接说：所有迁移指标都改善，因为 TTFT / disruption 仍有反例。
- 课堂问题：如果只保留一句贡献，该写成 protocol win、transport win 还是 install-path win？

说明：本页为课程研究案例或教学假设，不代表已发表论文结论；只有在实验包给出模型、硬件、工作负载和原始记录时，数字才可作为结果引用。

课程研究案例（内部实验）：负结果

核心结论

负结果不是失败汇报，而是帮助读者理解系统边界的证据。

- Segment Reuse：能 commit 但输给 native prefix，说明 committed path 成本尚未压过强 baseline。
- Tiered KV：movement/restore 触发条件不足或代价过高，说明 admission policy 是核心问题。
- MLP Materialization：单算子占比过小，说明优化目标应转向边界与写回路径。

说明：本页为课程研究案例或教学假设，不代表已发表论文结论；只有在实验包给出模型、硬件、工作负载和原始记录时，数字才可作为结果引用。

课程研究案例（内部实验）：Semantic MapReduce

核心结论

SAGE 案例适合训练学生区分 orchestration claim、coverage claim 和 LLM quality claim。

- 可以说：workflow / stream / semantic reduce 能组织大规模分析证据链。
- 需要谨慎：coverage 不足时 F1 下降，不能简单归因于 LLM reducer 不好。
- 不应声称：SAGE 已经替代 Spark/Flink/Ray 或已经证明所有 LLM reducer 稳定获胜。

说明：本页为课程研究案例或教学假设，不代表已发表论文结论；只有在实验包给出模型、硬件、工作负载和原始记录时，数字才可作为结果引用。

例子：工程量大不等于研究贡献大（1）

核心结论

系统论文比较方法课另一个关键目标，是帮助学生区分“做了很多”和“回答了好问题”。

工程量大

代码多、模块多、图表多，但问题定义可能仍然模糊。

例子：工程量大不等于研究贡献大（2）

核心结论

系统论文比较方法课另一个关键目标，是帮助学生区分“做了很多”和“回答了好问题”。

研究贡献大

可能只改变一个关键控制点，却清楚说明了为什么这个点改变了系统边界。

例子：工程量大不等于研究贡献大（3）

核心结论

系统论文比较方法课另一个关键目标，是帮助学生区分“做了很多”和“回答了好问题”。

课堂结论

论文比较首先比较问题与证据，其次才比较工程复杂度。

课堂练习：用五句话否定一篇论文（1）

核心结论

真正的精读能力之一，是能用短而硬的判断指出论文的边界。

练习一

用一句话写出论文假设最脆弱的地方。

课堂练习：用五句话否定一篇论文（2）

核心结论

真正的精读能力之一，是能用短而硬的判断指出论文的边界。

练习二

用一句话写出最需要补充的实验。

课堂练习：用五句话否定一篇论文（3）

核心结论

真正的精读能力之一，是能用短而硬的判断指出论文的边界。

练习三

用一句话写出：如果这个实验结果反过来，整篇论文最先倒塌的结论是什么。

课堂练习

核心结论

课堂练习可围绕代表论文的独立拆解展开，以训练结构化比较能力。

- 每组选择一篇代表论文。
- 用五段式框架做 10 分钟汇报。
- 汇报中必须包含一个会否定作者结论的反例设想。

课堂讨论

核心结论

系统论文阅读的重点在于判断问题是否真实、机制是否必要以及结果是否可信。

- 如何区分“新机制”和“新命名”？
- 如何判断一个对照实验是否公平？
- 什么样的结果最应该要求作者额外解释？

本讲总结

核心结论

规范的论文阅读能力能够直接提升研究选题质量与实验设计质量。

- 论文是问题建模与机制设计的分析样本。
- 论文阅读的最终目标是形成稳定的研究判断能力。

本讲参考资料

- **vLLM / PagedAttention**（SOSP 2023） 原文链接
- **Revisiting Disaggregated LLM Serving**（2025 preprint, arXiv:2601.08833） 原文链接
- **Rethinking KV Cache Compression**（MLSys 2025） 原文链接

课件中的论文数字均指作者在特定模型、硬件和工作负载下报告的结果，不代表无条件性能。

对应 Tutorial

核心结论

本讲对应 Tutorial 9。先独立作答，再使用对应 Tutorial 的教师讨论要点核对。

- 文件：`build/tutorials/Tutorial_09_ 论文比较方法.pdf`
- 动手运行、日志记录与提交要求统一查看 `build/experiments/` 和对应 experiment 源目录下的 `assignment_spring-2027/`。