

软标签：复盘为什么要记下当时的次优选项

Soft labels: why a review should record the options not taken

庚辛研究院 / Glacier Institute^{*†}

出自 维特根斯坦研究所 · The Wittgenstein Institute

庚辛研究院 (Glacier Institute) | 通讯论文 Working Paper GI-WP-2026-P13 | 2026-09-22 | 中文正文，附英文标题与摘要

建议引用格式 Cite as:

Glacier Institute (2026). Soft labels: why a review should record the options not taken. Glacier Institute Working Paper GI-WP-2026-P13.

<https://glacier.mba/research/GI-WP-2026-P13.html>

庚辛研究院 (2026). 《软标签：复盘为什么要记下当时的次优选项》. 庚辛研究院通讯论文 GI-WP-2026-P13.

<https://glacier.mba/research/GI-WP-2026-P13.html>

要点 Key Takeaways

- 1 只记成败，等于把一场会压缩成一个字；软标签是那些被压掉的东西。
- 2 「这一单没成」是一个字；「当时还有哪两条路」是一段可以被别人用的信息。
- 3 频率写了五档，内容只写了七种，而且七种全是结论——洞察在，仪表不在。
- 4 软标签必须在结果出来之前写——结果一揭晓，当初的不确定就被抹掉了。

摘要

一次复盘如果只产出成败，组织学到的远少于它产出「当时还有哪些选项、为什么没选」。前者是硬标签，后者是软标签。本文主张：复盘的学习效率由软标签的比例决定，而软标签不会自己出现，它必须被一张表格逼出来。底本是公开网页：一家精品投行的官网上有六十五处关于复盘的公开表述，本文做了三次机械计数——时间粒度五档、「每周」被十处一致确认；产出字段只有七种被命名过，且最多的两种指同一件东西；而在全文一万零七十九行上做的缺席检查显示，「次优」「第二好」「没选」「替代方案」「当时还有」全部零处。也就是说，这套公开表述规定了多久复盘一次，却没有任何一处规定复盘要记下当时被排除的选项——而同一份文本里已经把问题描述得很准：放过一夜，细节就被压缩成一句「他们没兴趣」。机制部分用五组文献：知识蒸馏里的暗知识、自蒸馏与标签平滑、事后复盘的实证、组织学习中的样本不对称，以及后见之明、自我评估失真、心理安全与错误管理训练。工具是一张十一行的软标签复盘登记表，外加一条软标签比例的计数规则。主张写成可以被推翻的命题，文末给出五类边界。本文不出现任何客户公司的名字，不出现这家机构自身的任何经营数字。

关键词：组织学习；事后复盘；知识蒸馏；暗知识；后见之明偏差；心理安全；失败学习

JEL 分类 JEL Classification: D83 (搜寻、学习与信息), M10 (一般商业管理), L84 (专业与商业服务), G24 (投资银行与创业投资)

^{*} 利益披露 Disclosure: 庚辛研究院是庚辛资本的研究机构，本文由庚辛研究院署名。庚辛资本在其业务中担任部分科技企业的财务顾问，并以自有资金参与部分企业的股权投资；这类关系可能与本文讨论的行业范围重合。本文不涉及任何具体在投项目，不使用任何未公开信息；文中引用的企业、行业与数据信息一律取自公开来源并逐条注明出处。作者未因本文获得任何第三方报酬。

[†] 免责声明 Disclaimer: 本文为方法讨论性质的通讯论文，仅代表作者在写作时点的分析。本文不构成投资建议，不构成任何证券、基金份额或其他权益的要约或要约邀请，亦不构成对任何公司估值、融资结果或投资回报的承诺或预测。本文未经同行评议，内容可能在后续版本中修订。

Abstract

A review that produces only success or failure teaches an organisation far less than one that also produces "what else was on the table, and why it was not chosen". The first is a hard label; the second is a soft label. This paper argues that the learning rate of a review is set by the proportion of soft labels it carries, and that soft labels do not appear on their own — a form has to force them out.

The source text is public. A boutique investment bank's website contains sixty-five distinct public statements about its review practice. Two mechanical counts are run on them. The first counts time granularity: five cadences are specified in detail, and the weekly one is confirmed by ten separate statements. The second counts output fields — what a review is supposed to record — and finds only seven ever named, of which the two most frequent (writing a lesson into a prohibition, eleven statements; the stop-doing list, eight) refer to the same thing.

A third count is an absence check, run over all 10,079 lines: "second best" appears zero times, "the option not taken" zero, "alternative" zero, "what else was available" zero. The public text prescribes how often to review, and prescribes that a review be written into a prohibition, but nowhere prescribes recording the options that were ruled out at the time.

The gap deserves a paper because the same public text already describes the problem precisely: the most valuable thing in a meeting is often not the conclusion but the wording of the other side's hesitation and the order in which they probed; leave it overnight and the detail compresses into "they weren't interested". The insight is there. The field that would carry it is not.

The mechanism section draws on five bodies of work: dark knowledge in knowledge distillation, where the relative probabilities a teacher assigns among wrong classes carry information no hard label holds; self-distillation and label smoothing, which describe what happens when the teacher is the student; the empirical literature on after-event reviews, which finds that reviews work but that the effect depends on the kind of review rather than on their number; asymmetric sampling in organisational learning, where failure teaches more but failures are systematically undersampled; and a set of boundary references — hindsight bias, flawed self-assessment, psychological safety, and error management training.

The paper offers one tool: an eleven-row soft-label review record, each row written as field, what the field asks, what a good entry looks like, what a bad one looks like, together with a counting rule for the soft-label proportion that anyone can compute from ten records. The central claim is stated as a falsifiable proposition together with the evidence

that would refute it. Five boundary cases close the paper, including one that must be stated: the premise that the teacher is stronger than the student often fails inside an organisation.

No client company is named, no description that could identify one is used, and none of the firm's own financial figures appear.

Keywords: organisational learning; after-event review; knowledge distillation; dark knowledge; hindsight bias; psychological safety; learning from failure

1. 问题：「你们复盘到底记什么」

这个问题有两个版本，一个来自外面，一个来自里面。

外面的版本通常这样问：你们说每会一小盘、每天一大盘、每周一总盘，那和开会有什么区别？盘完之后留下什么，别人拿得到吗？

里面的版本更扎人：复盘写了这么多年，为什么新人还是要重新踩一遍坑？

本文认为这两个问题是同一个问题，而且答案不在频率上。

一次复盘写下「这一单谈成了」或者「这一单没成」，是一个硬标签。一次复盘写下「当时还考虑过另外两条路，第二条差在哪里，为什么最后没走」，是软标签。两者的成本差不了多少，信息量差很远。

这个区分不是本文发明的。它来自一篇被引用了十年的机器学习论文：训练一个小模型去拟合大模型输出的完整概率分布，比让它去拟合正确答案本身学得更好，因为大模型在错误答案之间给出的相对概率里，藏着「这道题错在哪一档」的信息。那篇论文的作者把这部分信息叫做暗知识。

一家机构的复盘制度，恰好就是把上一段经历当成教师、把下一段动作当成学生的一次蒸馏。如果教师只输出成败，学生学到的就是一个字。

本文的路线：第2节把公开底本摆出来，做三次机械计数；第3节给出五组机制文献；第4节把蒸馏的三个前提逐条对上一家机构的现实，说明哪一个能靠制度保证；第5节给出工具；第6节把主张写成可以被推翻的命题；第7、8节划边界。

与本系列其他各篇的分工。GI-WP-2026-P11《借来的理论》第4.4节对「数据质量与教师模型 → 复盘制度」这一对做过一次三段式判断，并提出「在复盘纪要里加一栏」；本文把那一句展开成一篇独立论文，补上它没有的三样东西：公开表述的三次计数、三个前提的逐条判

表 2 / Exhibit 2 官网公开文本里被命名过的复盘产出字段（续表 1）。

产出字段	官网里的说法	处数	代表行号
禁令	「踩过的坑写成禁令」	11	L932、L3065、L4067
只做减法的清单	「Stop Doing List（只做减法的清单）」	8	L932、L3065、L5950
方法	「蒸馏过的经验，才是方法」	5	L1524、L5874、L5903
尺子	「把『我感觉』还原成『指标显示』」	3	L3142、L5907、L8326
备忘录	「机构反馈逐条落成备忘录」	2	L7974、L8218
缺口	「评估上一周的判断质量、找出缺口」	1	L3889
名录	「能在几天内触达决策人的名录」	1	L8118
当时的次优选项	全文无对应说法	0	—

来源与说明：计数规则同表 1，分母同为 65 行。末行不是遗漏，是本文这次计数的结果。计数由本文做出。

定、以及一张可以直接照抄的登记表。GI-WP-2026-P12《换手处的损耗》的结论是「端到端必须配一个外部评估者」；本文讲的正是这个评估者要产出什么。GI-WP-2026-P9《校准》讲的是在发问之前把预期对齐，是事前；本文是事后。

「这一单没成」是一个字；「当时还有哪两条路」是一段可以被别人用的信息。

2. 底本：官网关于复盘写了什么

2.1 底本怎么取的

底本是公开网页，任何人可以复现。取法：`curl` 拉取 <https://glacier.mba/institute.html>，依次剥掉 HTML 注释、`<script>`、`<style>`，再把全部标签换成换行、反转义实体、去空行，得到纯文本，共 10,079 行。抓取日期 2026-09-22。下文行号均指这份纯文本的行号，与 GI-WP-2026-P11、GI-WP-2026-P12 用的是同一份底本、同一套行号。

2.2 第 18 章第 4 对：复盘会就是教师模型

官网第 18 章把知识蒸馏对上了自己的复盘制度：

「决定模型上限的不是数据量，是数据质量；要有教师模型持续评估：缺哪一类泛化数据，就补哪一类场景。」——理论句，L3887 [1]

「复盘会就是那个教师模型——评估上一周的判断质量、找出缺口，指挥下一周去补缺的那类场景。」——现象句，L3889 [1]

这一对标注的来源是「Hinton, Vinyals & Dean (2015), Distilling the Knowledge in a Neural Network, arXiv:1503.02531」（L3919）[1][2]。

这是一次相当准确的类比。教师模型确实在做「评估质量、找出缺口、指挥补缺」这三件事。但原论文里教师模型还做了第四件事，而现象句没有提到它——教师输出的是一个分布，不是一个答案。第 3.1 节说明这第四件事为什么才是蒸馏的核心。

2.3 官网自己已经把问题描述准了

这一节要先把一件事说清楚：本文指出的缺口，不是这家机构没看见的东西。同一份公开文本里有两句话，把硬标签的形成过程描述得比多数管理文献都准：

「会上最值钱的往往不是结论，是对方犹豫时的措辞、追问的顺序——那是他还没摊出来的底牌。」——L5886 [1]

「放过一夜，细节就被压缩成一句『他们没兴趣』。当天盘完，第二天那场会才有修正过的讲法。」——L5887 [1]

第二句就是硬标签的定义。一场两小时的会里有几十条信号——对方在哪一页停下、问题问到第几层、哪个数字被追了两遍——过一夜之后全部塌缩成五个字。官网用的动词是「压缩」，接着给出三个字的结论：

「信号会饿。」——L5888 [1]

同一篇里还写了一句方法上的要求：

「盘信号，不只盘结论」——L5893 [1]

所以洞察是有的。本文要问的是下一个问题：这条洞察有没有落成一个字段。

2.4 三次计数

下面三张计数由本文做出，规则写在写作说明 6.1，任何人可以对同一份纯文本复算。

第一次：时间粒度。规则是：在同一行里同时出现「复盘」与某个时间粒度词的去重行数。

表 1 / Exhibit 1 官网公开文本里复盘的时间粒度。

时间粒度	处数	行号
每场会	5	L912、L3228、L5882、L9260、L9265
当天／当晚	5	L912、L3228、L5882、L9260、L9265
每周	10	L912、L3228、L3877、L3889、L5256、L5414、L5882、L8218、L9260、L9265
每月	1	L3228
每季	2	L3228、L5882

来源与说明：计数规则：同一行同时出现「复盘」与该粒度词，整行逐字去重后计一处。分母为含「复盘」且非第三方原话的去重行 65 行。计数由本文做出。

第二次：产出字段。同样的规则，数的是「复盘应当记下什么」被命名过几种。

第三次：缺席检查。这一次的的范围不是那 65 行，是全部 10,079 行——因为要证明「没有」，就不能只在自己划的范围里找。

表 3 / Exhibit 3 缺席检查（续表 2）。

检索词	命中	说明
次优	0	
第二好	0	
没选／未选择／没 有选	0	
替代方案	0	
当时还有	0	
为什么不选／为什么 没选	0	
被排除的选项／排除 的理由	0	
差在哪一档／差多少	0	
备选	1	L5838，指创始人手上的备选方案，与复盘字段无关
否在哪里，为什么 否	1	L8622，见下

来源与说明：检索范围为纯文本全部 10,079 行，不限于含「复盘」的行。检索词逐条列出，命中数为原始命中行数（未去重）。计数由本文做出。

最后一行是这次检查里唯一一处接近的表述，值得逐字摆出来：

「一次否决往往比十次赞美更有信息量。」——L8621 [1]

「被否的时候先别辩解，先记账：否在哪里，为什么否。这是别人替我们量过的尺子。」——L8622 [1]

这两句几乎就是本文要的东西，但方向是反的。它说的是别人否我们的时候要记账。本文说的是我们自己否掉一个选项的时候也要记账。前者的教师在外部，后者的教师在自己身上——而后者在数量上多得多：一个组织被别人否的次数有限，自己否掉某个做法的次数是它的很多倍。

2.5 三次计数合起来说了什么

第一，频率被写得很细，内容被写得很粗。五个时间粒度，「每周」这一档被十处表述一致确认。而产出字段只有七种被命名过，其中最多的两种指的是同一件东西（写成禁令／只做减法的清单），合计占了被命名处数的一大半。

第二，被命名的七种产出，全部是结论型的。禁令、方法、尺子、备忘录、缺口、名录——它们都是复盘得出的东西，不是复盘当时看见的东西。用第 3.1 节的话说：它们都是硬标签。

第三，洞察与仪表之间有一个缺口。L5886 说会上最值钱的不是结论，L5893 说要盘信号不只盘结论；但七种被命名的产出字段里，没有一种承载信号。一条被写下来的洞察，如果没有一个字段接着它，它下周就会变回一句好话。

频率写了五档，内容只写了七种，而且七种全是结论——洞察在，仪表不在。

3. 机制：硬标签与软标签

3.1 暗知识：教师真正教的是错在哪一档

Hinton、Vinyals 与 Dean [2] 做的事情是：先训练一个强的教师模型，再让一个小的学生模型去拟合教师输出的完整概率分布，而不是拟合正确答案。他们把关键洞察写得很清楚：教师在错误类别之间给出的相对概率，携带了正确答案本身没有的信息。一张图片被判为宝马车的概率最高，但它被判成垃圾车的概率如果远高于被判成胡萝卜，那就说明模型学到了「宝马和垃圾车都是车」这件事。作者把这部分信息叫做暗知识。

这条路更早有一个前身。Bucilua、Caruana 与 Niculescu-Mizil [3] 提出模型压缩：用一个大的集成模型去给数据打标，再用这些标去训练一个小模型，小模型

能达到接近集成的表现。两篇合起来给出的命题是：学生学得快，不是因为答案更准，是因为标签更软。

搬到组织里，对应关系是直接的：

- 硬标签 = 这一单成了 / 没成；这个判断对了 / 错了。
- 软标签 = 当时排在第二位的选项是什么；它差在哪一档；什么情况下它会变成第一位；当时手上缺的是哪条信息。

软标签之所以值钱，是因为它教的是决策空间的形状，而不是决策空间里的一个点。一个点只在完全相同的局面里可用；形状可以外推。

3.2 自蒸馏：教师就是学生的时候会发生什么

蒸馏预设教师比学生强。组织里的复盘常常不满足这一条——评估上周判断的人，就是上周做判断的人。

机器学习里对这种情形有专门的研究。Furlanello 等 [4] 发现，用一个模型去教一个结构完全相同的新模型（他们叫它「重生网络」），学生有时反而超过教师。这似乎是个好消息，但它的成因需要小心：Müller、Kornblith 与 Hinton [5] 指出，这一类增益的一部分来自标签变软本身的正则化效应，而不是来自教师更强。他们同时报告了一个反向结果：用标签平滑训练出来的教师，它的表征会把同类样本压得更紧、丢掉类间的相对结构，因此它作为教师去蒸馏时反而更差——软得太均匀，暗知识就没了。

两条合起来给组织一个很具体的提醒：

1. 自己评估自己不是完全无效的，因为把输出变软本身就有价值。
2. 但软化必须保留结构。复盘纪要如果每次都写「大家都有责任、以后注意」，那是把标签抹平，不是变软。真正的软标签必须是有序的：第二名是谁，差多少。

官网那句「复盘先复自己，再复别人」（L5899）[1] 在这个框架里是成立的，条件是第 2 条。

3.3 复盘的实证：有效，但取决于种类

事后复盘不是一个只有故事的做法，它有实验证据。

Ellis 与 Davidi [6] 做过一项关键的研究：让受训者不仅复盘失败的经历，也复盘成功的经历。结果是，同时复盘成败两种经历的人，构建出的心智模型比只复盘失败的人更完整，后续表现也更好。这条结论对本文非常重要——它说明复盘的价值不在于找错，而在于覆盖更多的可能性。只盘失败，等于只采样了分布的一侧。

Ellis、Mendel 与 Nir [7] 接着做了一个更细的实验，比较不同种类的事后复盘。他们发现复盘的效果取决于它引

导人去做什么，而不是取决于有没有开这个会。Villado 与 Arthur [8] 用一个复杂任务比较了主观复盘与基于客观记录的复盘，发现后者带来的团队表现改善更大——手上有记录的复盘，和凭印象的复盘，不是同一件事。

Tannenbaum 与 Cerasoli [9] 把这一支做了一次元分析，结论是复盘（debrief）确实能改善绩效，而且这个效果在个体与团队、模拟与真实场景里都站得住。

四条合起来：复盘有效是有证据的；但有效的是有结构、有记录、覆盖成败两侧的复盘，不是复盘这个动作本身。频率再高，如果每次只记一个硬标签，加起来还是一串硬标签。

3.4 组织学习：失败教得更多，但失败被欠采样

Madsen 与 Desai [10] 研究了一个失败后果极其严重、记录又极其完整的行业——轨道运载火箭发射。他们发现组织从失败中学到的东西比从成功中学到的更多，而且从失败中学到的知识衰减得更慢。Haunschild 与 Sullivan [11] 用航空业的事故与事件数据得到一条互补的结论：原因异质的事故比原因单一的事故带来更多学习，因为它们迫使组织去检查更多的可能解释。

但这里有一个陷阱，Denrell [12] 把它写得最清楚：观察别人的做法来学习时，人看到的样本是被筛选过的——失败的组消失了、不被报道、不被访谈。于是同一套冒险策略，从幸存者身上看起来像是有效的。他把这叫做失败的欠采样，并指出很多管理上的「成功法则」是这个偏差的产物。

Argote 与 Miron-Spektor [13] 把这一支整合成一个框架：组织学习是经验转化为知识的过程，而这个过程发生在什么样的情境里，决定了它转化得出还是转化不出。

对本文的用处是一条很硬的推论：如果复盘只记录已经发生的那条路，组织的学习样本就永远是被自己的选择筛过的。记下当时的次优选项，是这个组织唯一能自己制造的反事实样本——它不需要等别人失败，也不需要等自己失败。

3.5 边界文献：软标签为什么不会自己出现

这一小节讲的是为什么需要一张表格，而不是提醒大家「多写一点」。

第一，后见之明偏差。Fischhoff [14] 的经典实验说明：知道了结果之后，人会系统性地高估自己当初对这个结果的预期。结果一旦揭晓，当时真实的不确定性就被抹掉了。这意味着软标签必须在结果出来之前写，事后补写的次优选项已经被结果污染了。

第二，自我评估靠不住。Kruger 与 Dunning [15] 的研究说明，能力较弱的人系统性地高估自己，部分原因是识

表 4 / Exhibit 4 软标签复盘登记表。

序号	栏位	这一栏在问什么	合格的填法	不合格的填法
1	这次的判断	最后定了什么	一句话写清动作与对象	写成结果而不是动作
2	当时排第二的选项	差一点被选中的是哪一条	具体到可执行的程度	「还有别的方案」
3	第二名差在哪一档	差距是大是小，差在哪个维度	指出差的那个维度并给出方向	「综合考虑」
4	什么情况下第二名会变第一	翻转这个判断需要什么	写出一个可观察的条件	「看情况」
5	最早被排除的那条	一开始就被否掉的是什么，为什么	写出排除它的那条理由	留空
6	当时缺的那条信息	手上没有、但会改变判断的是什么	指名那条信息与它的来源	「信息不够」
7	这次用了哪条既有禁令	有没有调用过去的经验	写出禁令编号或原句	留空
8	什么时候能验证	这个判断多久之后见分晓	写一个具体日期	「以后就知道了」
9	如果结果相反，先怀疑哪一步	事前指定归因的起点	指名一个具体环节	事后再补
10	本次是否产生新禁令	有没有沉淀	有就写出来，没有写「无」	留空
11	记录人与复核人	谁写的，谁看过	复核人不是做判断的那个人	两栏同一个名字

来源与说明：十一栏，随复盘纪要一同填写。第 1 栏是硬标签，第 2 至 7 栏是软标签，第 8 至 11 栏是让这张表不退化的四条约束。本表为本文归纳，官网没有对应物。

别自己不足所需的正是他们缺乏的那种能力。Dunning、Heath 与 Suls [16] 把这条推广到工作场景与健康、教育等领域，结论一致：自我评估与客观表现之间的相关通常很弱。所以「让每个人自己写下判断质量」这条路，先天有一个上限。

第三，说出来要有安全。Edmondson [17] 的研究说明，团队里的心理安全决定成员愿不愿意说出错误与疑问，而学习行为正是绩效的中介。写下「我当时差点选了另一条」是一种暴露。没有安全，这一栏会被填成空话。官网那句「错处得先摆上桌——看得见的错，才改得掉」（L5911 – L5912）[1] 说的正是这件事。

第四，把错误当成学习材料是可以训练的。Keith 与 Frese [18] 的元分析比较了错误管理训练（明确鼓励在训练中犯错并引导人从中提取信息）与传统训练，发现前者效果更好，尤其是在迁移到新任务时。这条给出一个乐观的推论：软标签不是一种天赋，它是一个可以被安排出来的动作。

四条合起来，就是本文要一张表格而不是一句倡议的理由：后见之明要求它在结果之前写；自我评估的失真要求它有固定的栏位而不是自由发挥；心理安全要求它被规定成常规动作而不是临时要求；而错误管理训练说明这件事教得会。

软标签必须在结果出来之前写——结果一揭晓，当初的不确定就被抹掉了。

4. 蒸馏的三个前提，逐条判

一个借来的概念要能用，得先看它的前提在这儿成不成立。知识蒸馏有三个前提。

4.1 前提一：教师比学生强

在组织里常常不成立。复盘会上评估上周判断的，往往就是上周做判断的那批人。这不是自蒸馏，因为自蒸馏里教师至少是一个训练好的、固定的模型；组织里的「教师」和「学生」是同一批人在同一个星期里的两个身份。

更要紧的是，教师的错误会被完整地传给学生。一个判断错了但没人发现，它会作为「正确的教师输出」被写进禁令清单，然后被全员重读。自蒸馏放大既有的偏差，而不是纠正它。

能做的补救有限，但不是没有：第 3.3 节的 Villado 与 Arthur [8] 指出客观记录比主观印象更有效。把「当时手上有什么」这条记录留下来，教师就不是纯粹的自己，而是自己加上一份不会撒谎的时间戳。

4.2 前提二：教师是先独立训练好的

在组织里常常不成立。原论文里的教师是在独立的数据上先训练好的。一场只用本周自己产生的材料的复盘，不是一个独立的教师，是同一个模型的回声。

这一条正是 GI-WP-2026-P12 的结论落地的地方。那一篇论证端到端不换手的代价是归因变难，因此必须配一个外部评估者。外部评估者在这里的技术含义是：教师的一部分输入必须来自没有参与这件事的人。

官网那句「被否的时候先别辩解，先记账：否在哪里，为什么否。这是别人替我们量过的尺子」（L8622）[1]是本文找到的、唯一一处明确把教师放在外部的表述。它的对象窄了一点——只覆盖别人否我们的场合——但方向是对的。

4.3 前提三：教师的输出是软的

这一条是三条里唯一能靠制度保证的。前两条取决于人，取决于组织有没有外部输入；第三条只取决于表格上有没有那一栏。

而按第 2.4 节的三次计数，这一栏目前不在。七种被命名的产出字段全部是结论型的，全文对「当时的次优选项」零处提及。

所以本文的建议落在第三条上，不落在前两条上。不是因为前两条不重要，是因为前两条改不动，第三条明天就能改。一条改得动的规则，胜过两条正确但没人执行的原则。

4.4 软标签比例：一个可以数出来的量

把第三条变成一个数：

软标签比例 = 复盘纪要中同时填写了「次优选项」「差在哪一档」「当时缺的那条信息」三栏的份数 ÷ 抽查的纪要总份数。

规则要写死三件事，否则这个数会自己漂：

1. 抽样必须随机，不能挑写得好的那几份。
2. 三栏必须同时非空才计入分子。只填一栏不算——只填「次优选项」而不填「差在哪一档」，等于给了第二名的名字却没给分数，暗知识还是没有。
3. 「无」也是一种合格的填法，但必须写出来。真的想不出第二个选项，就写「当时只看到这一条」——这本身是一条极有价值的记录，它说明当时的视野窄到了什么程度。

这个比例可以按季度算，趋势比绝对值有用。第一次算出来大概率很低，这不是问题；它是基线。

记忆点 · KEY POINT

三个前提里只有「把输出变软」是明天就能改的——所以先改它。

5. 工具：软标签复盘登记表

5.1 十一栏

给复盘纪要加一页。用法：每一次做完判断当场填，不等结果（第 3.5 节第一条）。

四条约束值得单独说。

第 8、9 栏是为了对付后见之明 [14]。事前写下验证时间与归因起点，事后就没法说「我早就觉得不对」。

第 10 栏的「无」是关键设计。允许写「无」，这一栏才不会被硬凑；而一个季度下来如果全是「无」，那说明的不是没有教训，是没有人看。

第 11 栏是为了对付前提一（第 4.1 节）。复核人不是判断人，教师就多了一点独立性。这一栏不需要复核人同意判断，只需要他确认第 2 至 6 栏不是空话。

5.2 这张表的成本

一页表格，每次多花几分钟。本文不认为它是免费的：它的真实成本不是时间，是第 2、5、6 栏要求人承认自己当时看见了别的路却没走，而第 11 栏还要让另一个人看见。这就是为什么它需要心理安全 [17]，也是为什么它必须被写成常规动作而不是临时要求——常规动作不针对任何人。

记忆点 · KEY POINT

允许填「无」，这一栏才不会被硬凑；一个季度全是「无」，说明的不是没教训，是没人看。

6. 一个可以被推翻的命题

主命题。如果两个组织的复盘频率相同，而其中一个的复盘产出中软标签比例更高（按第 4.4 节的规则计），那么在一段时间之后，软标签比例高的那个组织在新的、与过去不完全相同的局面里表现更好；而在与过去高度相似的局面里，两者没有差别。

这个命题的形状是有意这样写的：软标签的收益应当体现在外推上，不体现在重复上。如果一个组织只做重复性的事，硬标签就够了——这也是第 7 节反例二。

什么证据能推翻它。在一组复盘频率相当的组织里，按软标签比例分成两组，比较它们在新局面与旧局面下的表现。如果软标签比例高的那一组在新局面里也没有优势，主命题不成立——那说明软标签只是记录成本，不是学习材料。如果两组在旧局面里出现差异，命题同样

要修正：那说明软标签的作用机制不是外推，是别的东西。

推论一（缺口检验，本文已做了一次）。如果一个组织的复盘制度是自发形成的、没有被一张表格规定过，那么它的公开表述里对「多久盘一次」的规定应当细于对「盘的时候记什么」的规定。表 1、表 2 显示：时间粒度五档、「每周」被十处一致确认，而产出字段只有七种被命名，且最多的两种指同一件东西。推论一在这一个样本上没有被推翻。如果换一家机构，发现它对复盘内容的规定比对频率的规定更细，推论一被推翻——那说明这个不对称不是自发制度的普遍特征。

推论二（时序检验，尚未检验）。如果软标签必须在结果之前写 [14]，那么事后补写的次优选项应当与事前当场写下的系统性不同：事后补写的会更接近最终结果。检验方法是同一批判断分两组，一组当场写、一组事后补，比较两组「第二名」的分布。本文没有做，因为它需要一次有对照的内部试验。明确留空，不用推测去补。

推论三（软化不等于抹平）。第 3.2 节引 Müller 等 [5] 指出，软得太均匀会丢掉暗知识。可检验的形式：如果一个组织的复盘纪要里第 2、3 栏长期填成没有区分度的套话，那么它的软标签比例即使很高，也不应当带来第 6 节主命题预测的外推优势。这条给出了一个诊断：比例高而无效，先查第 3 栏有没有区分度。

记忆点 · KEY POINT

软标签的收益应当出现在新局面里；旧局面里也有差别，那就不是软标签在起作用。

7. 边界与反例

一份只有正面例证的分析不值得信赖。以下五类，是这套做法已知会失灵的地方。

反例一：教师比学生强这个前提，在组织里常常不成立。这是最要紧的一条，第 4.1 节已经逐条说过，这里要把它的后果写完整：自蒸馆会把教师的偏差原样传下去，而且传得很整齐——因为禁令是全员重读的。一条错误的禁令比一条错误的判断危害大得多，它会被反复宣读。所以这张表必须配一条退出机制：任何一条禁令都要有它是基于哪次判断产生的记录，以便在那次判断被证明是错的时候，能把它撤回。这条在本文的建议里是与登记表并列的第二件事。

反例二：不是所有工作都需要软标签。高度重复、规则固定的工作，硬标签就够了，甚至更好——多写的每一栏都是纯成本。软标签的收益来自外推，而重复性工作不需要外推。判断的方法很简单：问下一局的局面和这

一次会不会不一样。答案是「基本一样」，这张表就该简化到第 1、8、10 栏。

反例三：写下来不等于被读到。第 2.4 节数出来的七种产出字段里，禁令与只做减法的清单占了大半，而官网反复强调的是「全员重读」（L932、L3065、L4067 等）[1]。这个强调是对的：一份没有被重读的记录，和没有记录的差别只在存储成本上。本文的登记表同样面临这个问题，而且更严重——软标签比禁令长、比禁令具体，因此更难被反复读。可能的解法是按第 4 栏（什么情况下第二名会变第一）建索引：不读全部，只在遇到相似局面时读那一条。本文没有验证这个解法。

反例四：这一栏会被优化。一旦软标签比例变成一个被看的数，它就会被凑。Villado 与 Arthur [8] 的结论提醒我们真正起作用的是基于客观记录的复盘，而不是复盘表格填得满。所以这个比例不适合当考核指标，只适合当诊断指标——看趋势，不排名次，不与任何人的评价挂钩。这一条是硬的：一旦挂钩，第 2 至 6 栏立刻变成作文。

反例五：本文的缺口来自公开文本，不等于实际没做。表 3 的缺席检查证明的是「这套公开表述里没有规定这一栏」，不是「这家机构从来不记次优选项」。内部纪要本文看不到，也不应当看到。公开文本的缺席是一个关于表述的结论，不是一个关于行为的结论。这一点与 GI-WP-2026-P11 第 7 节反例三、GI-WP-2026-P12 第 7 节反例五是同一条：没写下来的那部分不进分母。

一条错误的禁令比一条错误的判断危害大——它会被全员反复重读。

8. 这个分析在什么范围内成立

一、对象范围。本文分析的是一份公开文本里关于复盘制度的六十六处表述，出自一家专业服务机构。结论适用于「判断密集、反馈周期长、每次局面不完全相同」这一类知识工作的复盘设计。它不外推到高度重复的操作性工作（见第 7 节反例二），也不外推到有即时客观胜负判定的场景。

二、证据的性质。表 1、表 2、表 3 是机械计数，规则写在写作说明 6.1，任何人可以对同一份纯文本复算，结果应当一致。这三张表是计数，不是评价——它们说的是「这个词在公开文本里出现了几次」，不是「这家机构做得好不好」。表 4 是本文归纳的工具，不是测量。

三、计数规则的已知弱点。表 1、表 2 的规则是「同一行同时出现」，而纯文本的分行取决于原网页的标签结构。一个跨了两行的句子，它的两半会被分别计数或都不计数。本文选择这条规则是因为它机械、可复算；代

价是它对长句不友好。换一条规则（例如按段落而不是按行），个别数字会变，但第 2.5 节的三个读法依赖的是量级差异（十处对一处、七种对零种），不依赖单个数字。

四、缺席检查的局限。表 3 检索的是词，不是意思。一个组织完全可能用本文没有想到的说法来表达「当时的次优选项」。本文列出了检索词全表以便复核与增补，但不能排除漏检。这是一条真实的局限，写在这里而不是藏起来。

五、没有做因果推断。本文不主张加了这张表就会改善任何结果。第 6 节的主命题是一个待检验的命题，不是一个已经成立的结论。

六、一个诚实的空缺。推论二（事前写与事后补的差异）需要一次有对照的内部试验，本文做不了，明确留空。在这项工作完成之前，「软标签必须在结果之前写」这条规则的依据是 Fischhoff [14] 的实验证据加上一个类比，不是在这个场景里被直接验证过的结论。

七、本文不做什么。不评价这家机构的任何业务结果，不构成投资建议，不给出任何可回溯到具体交易或具体公司的信息。本文的三张计数表只数公开文本里的词，不涉及任何经营数据。

表 3 数的是词不是意思——检索词全表列出来，就是为了让别人来补我漏掉的说法。

9. 结论

回到第 1 节的两个问题。

「你们复盘到底记什么」。这个问题的答案不该是一张频率表。频率决定信息有多新鲜，内容决定信息有多厚。本文数出来的是：公开表述里频率被写了五档、「每周」被十处一致确认，而产出字段只有七种被命名过，且全部是结论型的。该补的不是频率，是字段。

「为什么新人还是要重新踩一遍坑」。因为他继承的是一串结论，不是一个决策空间。禁令告诉他「别做这个」，但没告诉他「当时还有哪两条路、为什么这条不行」。一个只拿到结论的人，遇到一个稍微不同的局面，就只能重新试一遍。结论可以继承，判断力不能——除非把当时看见的东西一起留下。

本文提了两件事，都很小：

第一，给复盘纪要加一页十一栏的登记表（第 5 节），并且当场填、不等结果。三栏同时非空的比例就是软标签比例，按季度算，看趋势，不排名次、不挂钩考核。

第二，给每一条禁令记上它出自哪一次判断（第 7 节反例一），以便那次判断被证明是错的时候能把它撤回。一条禁令一旦进了全员重读的清单，撤回它的难度远大于写下它。

两件事都不需要新的会议、新的角色、新的系统。它们只需要一张表格和一个字段。

三句话可以概括全文：

1. 只记成败，等于把一场会压缩成一个字；软标签是那些被压掉的东西。
2. 蒸馏的三个前提里，只有「把输出变软」是明天就能改的——所以先改它。
3. 软标签必须在结果出来之前写；结果一揭晓，当初的不确定就被抹掉了。

结论可以继承，判断力不能——除非把当时看见的那些选项一起留下。

参考文献 / References

体例：作者（年份）·题名·期刊/出版方，卷(期)，页码，DOI 或可访问链接。按来源等级分组：一手材料为庚辛研究院对外公开表述；二手文献为同行评议文献与公开预印本。下列各条于 2026-09-22 用 `_仓储/_build/refcheck.py` 的 `check()` 逐条真实请求核验，18 条全部可达，且每一条的题名与 Crossref 或 arXiv 返回的题名逐字比对一致。核验明细见写作说明第四节。本表不列任何未经核实的条目。

一手材料（庚辛研究院对外公开表述；本文引用的原句均出自 [1]，一字未改）

[1] 庚辛研究院（2026）·庚辛研究院官网：致企业家与投资人（L907–L912）、只做减法的清单（L932、L3064–L3065）、复盘的节拍（L3227–L3228）、第 18 章第 4 对「数据质量与教师模型/复盘制度」（L3886–L3889、L3919）、随笔《复盘是节拍器，不是救火队》（L5864–L5918）、随笔《把投资人当上游》（L8615–L8622）·庚辛资本官方网站。
<https://glacier.mba/institute.html>（访问日期 2026-09-22）

二手文献（知识蒸馏与软标签）

- [2] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. <https://arxiv.org/abs/1503.02531>（访问日期 2026-09-22）
- [3] Bucilua, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model Compression. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 535–541. <https://doi.org/10.1145/1150402.1150464>
- [4] Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., & Anandkumar, A. (2018). Born-Again Neural Networks. <https://arxiv.org/abs/1805.04770>（访问日期 2026-09-22）
- [5] Müller, R., Kornblith, S., & Hinton, G. (2019). When Does Label Smoothing Help? <https://arxiv.org/abs/1906.02629>（访问日期 2026-09-22）

二手文献（事后复盘的实证）

- [6] Ellis, S., & Davidi, I. (2005). After-Event Reviews: Drawing Lessons from Successful and Failed Experience. *Journal of Applied Psychology*, 90(5), 857 – 871. <https://doi.org/10.1037/0021-9010.90.5.857>
- [7] Ellis, S., Mendel, R., & Nir, M. (2006). Learning from Successful and Failed Experience: The Moderating Role of Kind of After-Event Review. *Journal of Applied Psychology*, 91(3), 669 – 680. <https://doi.org/10.1037/0021-9010.91.3.669>
- [8] Villado, A. J., & Arthur, W., Jr. (2013). The Comparative Effect of Subjective and Objective After-Action Reviews on Team Performance on a Complex Task. *Journal of Applied Psychology*, 98(3), 514 – 528. <https://doi.org/10.1037/a0031510>
- [9] Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do Team and Individual Debriefs Enhance Performance? A Meta-Analysis. *Human Factors*, 55(1), 231 – 245. <https://doi.org/10.1177/0018720812448394>

二手文献（组织学习与失败样本）

- [10] Madsen, P. M., & Desai, V. (2010). Failing to Learn? The Effects of Failure and Success on Organizational Learning in the Global Orbital Launch Vehicle Industry. *Academy of Management Journal*, 53(3), 451 – 476. <https://doi.org/10.5465/amj.2010.51467631>
- [11] Haunschild, P. R., & Sullivan, B. N. (2002). Learning from Complexity: Effects of Prior Accidents and Incidents on Airlines' Learning. *Administrative Science Quarterly*, 47(4), 609 – 643. <https://doi.org/10.2307/3094911>
- [12] Denrell, J. (2003). Vicarious Learning, Undersampling of Failure, and the Myths of Management. *Organization Science*, 14(3), 227 – 243. <https://doi.org/10.1287/orsc.14.2.227.15164>
- [13] Argote, L., & Miron-Spektor, E. (2011). Organizational Learning: From Experience to Knowledge. *Organization Science*, 22(5), 1123 – 1137. <https://doi.org/10.1287/orsc.1100.0621>

二手文献（边界：后见之明、自我评估、心理安全、错误管理）

- [14] Fischhoff, B. (1975). Hindsight Is Not Equal to Foresight: The Effect of Outcome Knowledge on Judgment under Uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 288 – 299. <https://doi.org/10.1037/0096-1523.1.3.288>
- [15] Kruger, J., & Dunning, D. (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology*, 77(6), 1121 – 1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- [16] Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed Self-Assessment: Implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest*, 5(3), 69 – 106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
- [17] Edmondson, A. (1999). Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly*, 44(2), 350 – 383. <https://doi.org/10.2307/2666999>
- [18] Keith, N., & Frese, M. (2008). Effectiveness of Error Management Training: A Meta-Analysis. *Journal of Applied Psychology*, 93(1), 59 – 69. <https://doi.org/10.1037/0021-9010.93.1.59>

核验说明：[1] HTTP 200；[2][4][5] arXiv 页 200，并以 arXiv API 返回题名逐字核对（"Distilling the Knowledge in a Neural Network"、"Born Again Neural Networks"、"When Does Label Smoothing Help?"）；[6][7][8][14][15][18] DOI 直接解析 200，Crossref 题名一致；[3][9][10][11][12][13][16][17] 出版方对脚本返 403（反爬，不是不存在），以 Crossref 元数据核对题名、卷、期、页码，全部一致。一处已修正并记录在案：[12] Denrell 2003 最初按卷期推写成 DOI [10.1287/orsc.14.3.227.15164](https://doi.org/10.1287/orsc.14.3.227.15164)，该地址取不到元数据；按题名在 Crossref 检索得到正确 DOI 为 [10.1287/orsc.14.2.227.15164](https://doi.org/10.1287/orsc.14.2.227.15164)（题名、卷 14、期 3、页 227 – 243 全部一致），已换。

本系列内部引用（尚无 DOI，按编号引用）

- GI-WP-2026-P9 《校准：在发问之前把预期对齐》
- GI-WP-2026-P11 《借来的理论：前沿技术的概念如何解释一家投行的日常》
- GI-WP-2026-P12 《换手处的损耗：全托管的边界画在哪里》

引用的庚辛官网原文出处

全部官网引文取自 [1] <https://glacier.mba/institute.html>（抓取日期 2026-09-22，剥去注释、脚本、样式与全部标签后的纯文本，共 10,079 行）。逐句出处：

- 第 18 章第 4 对理论句「决定模型上限的不是数据量，是数据质量；要有教师模型持续评估：缺哪一类泛化数据，就补哪一类场景。」——L3887
- 第 18 章第 4 对现象句「复盘会就是那个教师模型——评估上一周判断质量、找出缺口，指挥下一周去补缺的那类场景。」——L3889
- 第 18 章第 4 对来源行「Hinton, Vinyals & Dean (2015), Distilling the Knowledge in a Neural Network, arXiv:1503.02531」——L3919
- 「实行每会小复盘、每日完整复盘、每周策略复盘。」——L912
- 「每场会一小盘，当天一大盘，每周一次总盘，每月完整复盘，每季规划后两个季度的资本市场竞赛方案。」——L3228（本文只引这一句的前半，后半与本文论点无关）
- 「最朴素的一件东西，是一张『不做什么』的清单：踩过的坑写成禁令，复盘会全员重读。方法可以变，边界只加不减。」——L3065
- 「会上最值钱的往往不是结论，是对方犹豫时的措辞、追问的顺序——那是他还没摊出来的底牌。」——L5886
- 「放过一夜，细节就被压缩成一句『他们没兴趣』。当天盘完，第二天那场会才有修正过的讲法。」——L5887
- 「信号会饿。」——L5888
- 「盘信号，不只盘结论」——L5893
- 「复盘先复自己，再复别人。」——L5899
- 「复盘不是账本上的事后加总，它给的是一把尺子：把『我感觉』还原成『指标显示』。」——L5907
- 「错处得先摆上桌——看得见的错，才改得掉。」——L5911、L5912
- 「经验不复盘，只是经历；蒸馏过的经验，才是方法。」——L5874
- 「一次否决往往比十次赞美更有信息量。」——L8621
- 「被否的时候先别辩解，先记账：否在哪里，为什么否。这是别人替我们量过的尺子。」——L8622

- 表 1、表 2、表 3 的计数覆盖到的其余行号，逐条列在表内。

一处标点说明：L3065、L5887、L5907 原文中的内层引号在纯文本里是半角双引号或「」，本文放进外层「」引用时，内层统一按中文惯例写作『』。只改了引号形状，一个字没改。

脱敏说明：本文引用的十六句官网原文中没有出现任何客户公司的名称，因此引用不需要做任何「一家…公司」式的改写；改写处计 0。表 1 至表 3 的计数分母里，已按写作说明 5.1 剔除五行客户与第三方原话，剔除的原因是它们不是这家机构自己的表述，不是因为它们带有公司名。本文自己的行文中也不描述任何可反推到具体公司的特征。详见写作说明第五节。

利益披露 Disclosure：庚辛研究院是庚辛资本的研究机构，本文由庚辛研究院署名。庚辛资本在其业务中担任部分科技企业的财务顾问，并以自有资金参与部分企业的股权投资；这类关系可能与本文讨论的行业范围重合。本文不涉及任何具体在投项目，不使用任何未公开信息；文中引用的企业、行业与数据信息一律取自公开来源并逐条注明出处。作者未因本文获得任何第三方报酬。

免责声明 Disclaimer：本文为方法讨论性质的通讯论文，仅代表作者在写作时点的分析。本文不构成投资建议，不构成任何证券、基金份额或其他权益的要约或要约邀请，亦不构成对任何公司估值、融资结果或投资回报的承诺或预测。本文未经同行评议，内容可能在后续版本中修订。