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

## 软标签：复盘为什么要记下当时的次优选项

庚辛研究院 / Glacier Institute*†

**From The Wittgenstein Institute**

Glacier Institute（庚辛研究院） | Working Paper No. GI-WP-2026-P13-EN | Date: 2026-09-22 | Language: English, with Chinese title and abstract

**Cite as:**

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

### Key Takeaways

1. **Recording only the outcome compresses a whole meeting into one word. Soft labels are what got compressed away.**
2. **"It didn't close" is one word. "Here were the other two routes" is something another person can use.**
3. **Five cadences specified, seven output fields named, and all seven are conclusions. The insight is there; the instrument is not.**
4. **Soft labels have to be written before the outcome. The moment the result lands, the uncertainty that was there is erased.**

### 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

**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）

### 摘要 (Chinese abstract)

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。

**Keywords (Chinese):** 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

## 1. The Problem: "What Does a Review Actually Record?"

The question comes in two versions, one from outside and one from inside.

The outside version usually goes: you say you review after every meeting, again at the end of the day, again at the end of the week — how is that different from just having meetings? What is left afterwards, and can anyone else use it?

The inside version stings more: we have been writing reviews for years, so why does every new colleague still have to step into the same hole?

**This paper argues that these are the same question, and that the answer is not about frequency.**

A review that writes down "this one closed" or "this one did not" is a hard label. A review that writes down "we also considered two other routes; the second one fell short here, and this is why we did not take it" is a soft label. The two cost about the same to write. They carry very different amounts of information.

The distinction is not this paper's invention. It comes from a machine learning paper that has been cited for a decade: training a small model to fit the full probability distribution a large model outputs teaches it more than training it on the correct answer alone, because the relative probabilities the large model assigns **among the wrong answers** encode how wrong each of them is. The authors call that part dark knowledge.

A firm's review practice is exactly a distillation in which the previous stretch of experience is the teacher and the next set of actions is the student. If the teacher emits only outcomes, the student learns one word.

The route: Section 2 lays out the public source text and runs three mechanical counts. Section 3 gives five bodies of mechanism literature. Section 4 takes the three premises of distillation one by one and asks which of them an organisation can guarantee. Section 5 gives the tool. Section 6 states the claim as a proposition that can be refuted. Sections 7 and 8 draw the boundaries.

**How this differs from the other papers in the series.** GI-WP-2026-P11, *Borrowed Theories*, §4.4, gave a three-paragraph verdict on the "data quality and teacher model → review practice" pair and proposed adding a column to the review record. This paper expands that one

sentence into a standalone study and adds three things it did not have: three counts over the public statements, a premise-by-premise verdict, and a record sheet that can be copied as it stands. GI-WP-2026-P12, *Losses at the Handoff*, concludes that end-to-end must be paired with an outside assessor; this paper is about what that assessor has to produce. GI-WP-2026-P9, *Calibration*, is about aligning expectations **before** the ask; this paper is about **after**.

> "It didn't close" is one word. "Here were the other two routes" is something another person can use.

## 2. The Source Text: What the Website Says About Reviews

### 2.1 How the source text was taken

The source is a public web page and anyone can reproduce it. Method: curl the page https://glacier.mba/institute.html; strip HTML comments, then <script>, then <style>; replace every remaining tag with a newline; unescape entities; drop blank lines. The result is 10,079 lines of plain text, retrieved 2026-09-22. All line numbers below refer to that plain text — the same source text and the same line numbering used by GI-WP-2026-P11 and GI-WP-2026-P12.

### 2.2 Chapter 18, pair 4: the review meeting is the teacher model

Chapter 18 maps knowledge distillation onto the firm's own review practice:

> 「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。」
>
> *What caps a model is not the volume of data but its quality. You need a teacher model assessing continuously: whichever class of generalisation data is missing, go and collect that class of scenario.* — theory sentence, 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。」
>
> *The review meeting is that teacher model — it assesses the quality of last week's judgements, finds the gaps, and directs next week towards the class of scenario that fills them.* — phenomenon sentence, L3889 [1]

The pair cites "Hinton, Vinyals & Dean (2015), Distilling the Knowledge in a Neural Network, arXiv:1503.02531" (L3919) [1][2].

**This is an unusually accurate analogy.** A teacher model does assess quality, find gaps and direct the filling of them. But in the original paper the teacher does a fourth thing, and the phenomenon sentence does not mention it — **the teacher emits a distribution, not an answer.** §3.1 explains why that fourth thing is the core of distillation.

### 2.3 The website has already described the problem accurately

One thing has to be said before anything else: **the gap this paper identifies is not something the firm failed to see.** The same public text contains two sentences that describe the formation of a hard label more precisely than most management writing:

> 「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。」
>
> *What is most valuable in a meeting is often not the conclusion, but the wording of the other side's hesitation and the order in which they probed — that is the hand they have not yet shown.* — 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。」
>
> *Leave it overnight and the detail compresses into "they weren't interested". Review it the same day, and the next day's meeting has a corrected version to work from.* — L5887 [1]

The second sentence is the definition of a hard label. A two-hour meeting contains dozens of signals — which page they stopped on, how many layers deep the questions went, which figure was queried twice — and overnight they all collapse into five words. The verb the website uses is "compress". It then states the consequence in three characters:

> 「信号会馊。」 *Signal goes stale.* — L5888 [1]

The same essay also states a methodological requirement:

> 「盘信号，不只盘结论」 *Review the signals, not just the conclusion.* — L5893 [1]

**So the insight is there.** The question this paper asks is the next one: has the insight been turned into a field?

表 1 / Exhibit 1 Time granularity of reviews in the public text. Counting rule: a line in which 「复盘」 and the granularity term both occur, with identical lines deduplicated, counts once. The denominator is the 65 deduplicated lines containing 「复盘」 that are not third-party quotations. The count is made here.

| Granularity | Occurrences | Lines |
|---|---|---|
| Every meeting | 5 | L912, L3228, L5882, L9260, L9265 |
| Same day / same evening | 5 | L912, L3228, L5882, L9260, L9265 |
| Weekly | 10 | L912, L3228, L3877, L3889, L5256, L5414, L5882, L8218, L9260, L9265 |
| Monthly | 1 | L3228 |
| Quarterly | 2 | L3228, L5882 |

表 2 / Exhibit 2 Review output fields named in the public text (continued from Table 1). Counting rule and denominator as in Table 1. The last row is not an omission; it is the result of this count. The count is made here.

| Output field | How the website puts it | Occurrences | Representative lines |
|---|---|---|---|
| Prohibition | 「踩过的坑写成禁令」 *write the hole you fell into into a prohibition* | 11 | L932, L3065, L4067 |
| Stop-doing list | 「Stop Doing List（只做减法的清单）」 | 8 | L932, L3065, L5950 |
| Method | 「蒸馏过的经验，才是方法」 *only distilled experience is method* | 5 | L1524, L5874, L5903 |
| A ruler | 「把『我感觉』还原成『指标显示』」 *turn "I feel" back into "the measure shows"* | 3 | L3142, L5907, L8326 |
| Memo | 「机构反馈逐条落成备忘录」 *investor feedback written line by line into a memo* | 2 | L7974, L8218 |
| Gap | 「评估上一周的判断质量、找出缺口」 *assess last week's judgements, find the gaps* | 1 | L3889 |
| Contact list | 「能在几天内触达决策人的名录」 *a list that reaches decision-makers within days* | 1 | L8118 |
| The option ranked second at the time | no corresponding expression anywhere in the text | 0 | — |

## 2.4 Three counts

The three counts below are made here. The rules are stated in the writing note §6.1, and anyone can recompute them against the same plain text.

**Count one: time granularity.** Rule: the number of deduplicated lines in which 「复盘」 (review) occurs together with a term for a time granularity.

**Count two: output fields.** Same rule, counting how many kinds of thing a review is said to record.

**Count three: an absence check.** The scope here is not those 65 lines but all 10,079 — **because proving an absence inside a range you drew yourself proves nothing.**

The last row is the only near-hit in the check, and deserves to be set out verbatim:

> 「一次否决往往比十次赞美更有信息量。」
>
> *One rejection usually carries more information than ten compliments.* — 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。」
>
> *When you are turned down, do not explain yourself first — write it down first: where the no came from and why. That is a ruler somebody else measured for us.* — L8622 [1]

**These two sentences are almost what this paper is asking for, but the direction is reversed.** They say: record it when others turn us down. This paper says:

表 3 / Exhibit 3 The absence check (continued from Table 2). The search covers all 10,079 lines of the plain text, not only the lines containing 「复盘」. Search terms are listed individually; hits are raw line hits, not deduplicated. The count is made here.

| Search term | Hits | Note |
|---|---|---|
| 次优 *second best* | 0 | |
| 第二好 *second choice* | 0 | |
| 没选／未选择／没有选 *not chosen* | 0 | |
| 替代方案 *alternative* | 0 | |
| 当时还有 *what else there was* | 0 | |
| 为什么不选／为什么没选 *why it was not chosen* | 0 | |
| 被排除的选项／排除的理由 *the option ruled out, the reason* | 0 | |
| 差在哪一档／差多少 *by how much it fell short* | 0 | |
| 备选 *fallback* | 1 | L5838, referring to a founder's own fallback options, unrelated to a review field |
| 否在哪里，为什么否 *where the no came from and why* | 1 | L8622, see below |

**record it when we ourselves rule an option out.** In the first the teacher is external; in the second the teacher is inside. And the second is far more frequent: a firm is turned down a limited number of times, while it rules its own options out many times more often.

### 2.5 What the three counts say together

**First, the cadence is specified finely and the content coarsely.** Five granularities, with the weekly one confirmed by ten separate statements. Against that, only seven output fields are ever named, and the two most frequent of them (writing into a prohibition, and the stop-doing list) refer to the same thing, together accounting for most of the named occurrences.

**Second, all seven named outputs are of the conclusion type.** Prohibition, method, ruler, memo, gap, contact list — all of them are what a review **arrives at**, not what it **saw at the time**. In the vocabulary of §3.1, they are all hard labels.

**Third, there is a gap between the insight and the instrument.** L5886 says the most valuable thing is not the conclusion; L5893 says review the signals, not just the conclusion. Yet none of the seven named fields carries a signal. **An insight that has been written down but has no field waiting for it will be back to being a nice sentence by next week.**

> Five cadences specified, seven output fields named, and all seven are conclusions. The insight is there; the instrument is not.

## 3. Mechanism: Hard Labels and Soft Labels

### 3.1 Dark knowledge: what a teacher really teaches is how wrong each wrong answer is

Hinton, Vinyals and Dean [2] train a strong teacher model, then train a small student to fit the teacher's **full probability distribution** rather than the correct answer. Their key observation is stated plainly: the relative probabilities the teacher assigns among the wrong classes carry information the correct label does not. An image may be classified as a BMW with the highest probability; but if its probability of being a garbage truck is far higher than its probability of being a carrot, the model has learned that BMWs and garbage trucks are both vehicles. The authors call this part dark knowledge.

The approach has a predecessor. Buciluǎ, Caruana and Niculescu-Mizil [3] proposed model compression: label data with a large ensemble, then train a small model on those labels, and the small model approaches the ensemble's performance. Together the two papers give a proposition: **the student learns faster not because the answers are more correct, but because the labels are softer.**

The correspondence to an organisation is direct:

- **Hard label** = this one closed / did not close; this judgement was right / wrong.
- **Soft label** = what was ranked second at the time; by how much it fell short; under what conditions it would have ranked first; which piece of information was missing when the call was made.

**Soft labels are valuable because they teach the shape of the decision space rather than one point in it.** A point is only usable in an identical situation. A shape extrapolates.

### 3.2 Self-distillation: what happens when the teacher is the student

Distillation presumes the teacher is stronger than the student. Reviews inside an organisation often fail this condition — the people assessing last week's judgements are the people who made them.

Machine learning has studied this case directly. Furlanello and colleagues [4] found that a model teaching a structurally identical new model — they call it a "born-again" network — can yield a student that outperforms its teacher. That looks like good news, but its cause needs care. Müller, Kornblith and Hinton [5] show that part of this class of gain comes from **softening the labels itself**, a regularising effect, rather than from the teacher being better. They also report a result in the other direction: a teacher trained with label smoothing has representations in which same-class examples are packed more tightly and the relative structure between classes is erased, so **it is a worse teacher for distillation** — softened too evenly, the dark knowledge is gone.

Together this gives an organisation a very concrete warning:

1. Assessing yourself is **not entirely useless**, because softening the output has value in itself.
2. But softening has to preserve structure. A review record that always says "we all share responsibility, let's be careful next time" has flattened the label, not softened it. **A real soft label has to be ordered: who came second, and by how much.**

The website's line 「复盘先复自己，再复别人」 (*review yourself before reviewing others*, L5899) [1] holds in this frame, subject to point 2.

### 3.3 The empirical record on reviews: they work, but it depends on the kind

After-event review is not a practice supported only by anecdotes.

Ellis and Davidi [6] ran a key study in which trainees reviewed not only failed experience but **successful** experience as well. Those who reviewed both built more complete mental models than those who reviewed failures alone, and performed better afterwards. This matters greatly here: **the value of a review lies not in fault-finding but in covering more of the possibility space.** Reviewing only failures samples one side of the distribution.

Ellis, Mendel and Nir [7] then compared different **kinds** of after-event review and found that the effect depends on what the review directs people to do, not on whether the meeting happened. Villado and Arthur [8] compared subjective reviews with reviews based on objective records on a complex task, and found the latter produced a greater improvement in team performance — **a review with a record in hand and a review from memory are not the same thing.**

Tannenbaum and Cerasoli [9] meta-analysed this literature and conclude that debriefs do improve performance, and that the effect holds for individuals and teams, in simulated and in real settings.

Together: that reviews work is established. What works is a review that has structure, has records, and covers both success and failure — **not the act of reviewing itself**. However high the frequency, a string of hard labels is still a string of hard labels.

### 3.4 Organisational learning: failure teaches more, but failure is undersampled

Madsen and Desai [10] studied an industry where failure is catastrophic and the record is complete — orbital launch vehicles. They found that organisations learn more from failure than from success, and that knowledge gained from failure depreciates more slowly. Haunschild and Sullivan [11] reach a complementary conclusion from airline accident and incident data: **accidents with heterogeneous causes produce more learning** than those with a single cause, because they force an organisation to examine more possible explanations.

There is a trap here, and Denrell [12] states it most clearly. When learning vicariously from others' practices, the sample you see has been filtered — organisations that

表 4 / Exhibit 4 The soft-label review record. Eleven fields, filled alongside the ordinary review record. Field 1 is the hard label; fields 2 to 7 are the soft labels; fields 8 to 11 are the four constraints that keep the sheet from degenerating. The table is this paper's own; the website contains no equivalent.

| No. | Field | What it asks | A good entry | A bad entry |
|---|---|---|---|---|
| 1 | The judgement | What was decided | One sentence naming the action and its object | Writing the outcome instead of the action |
| 2 | The option ranked second | What nearly got chosen | Specific enough to be executable | "There were other options" |
| 3 | By how much it fell short | Large or small, and on which dimension | Names the dimension and the direction | "On balance" |
| 4 | What would flip the ranking | What it would take to reverse the call | One observable condition | "Depends" |
| 5 | The option ruled out earliest | What was dismissed first, and why | States the reason for dismissing it | Left blank |
| 6 | The information missing then | What was not in hand but would have changed the call | Names the item and where it would come from | "Not enough information" |
| 7 | Which existing prohibition was used | Whether past experience was invoked | Cites the prohibition by number or wording | Left blank |
| 8 | When it can be verified | How long before this is settled | A specific date | "We'll know eventually" |
| 9 | If the result is the opposite, suspect what first | Names the attribution starting point in advance | Names one specific step | Filled in afterwards |
| 10 | Any new prohibition from this | Whether anything was deposited | Write it, or write "none" | Left blank |
| 11 | Recorded by, reviewed by | Who wrote it, who has seen it | The reviewer is not the person who judged | The same name twice |

failed have disappeared, are not reported, are not interviewed. The same risky strategy then looks effective when read off the survivors. He calls this **undersampling of failure**, and shows that many management "success principles" are its product.

Argote and Miron-Spektor [13] integrate the literature into a framework: organisational learning is the conversion of experience into knowledge, and the **context** in which that conversion happens determines whether it happens at all.

The hard corollary for this paper: **if a review records only the path that was taken, the organisation's learning sample is permanently filtered by its own choices.** Recording the option ranked second is the only counterfactual sample an organisation can manufacture for itself — it requires waiting neither for someone else to fail nor for itself to fail.

### 3.5 Boundary literature: why soft labels do not appear on their own

This subsection explains why a form is needed, rather than a reminder to "write a bit more".

**First, hindsight bias.** Fischhoff's [14] classic experiments show that once people know the outcome, they systematically overstate how much they expected it beforehand. The moment the result is revealed, the uncertainty that was genuinely there is erased. This means **soft labels have to be written before the outcome is known**; a second-best option reconstructed afterwards has already been contaminated by the result.

**Second, self-assessment is unreliable.** Kruger and Dunning [15] show that less competent performers systematically overrate themselves, partly because recognising one's own shortfall requires exactly the competence they lack. Dunning, Heath and Suls [16] extend this to work, health and education with the same conclusion: self-assessment correlates only weakly with objective performance. So "have everyone write down the quality of their own judgement" has a ceiling built into it.

**Third, saying it requires safety.** Edmondson [17] shows that **psychological safety** in a team determines whether members will voice errors and questions, and that learning behaviour mediates performance. Writing down "I nearly took the other route" is an exposure. Without safety, that field fills up with platitudes. The website's line 「错处得先摆上桌——看得见的错，才改得掉」(*the error has to be on the table first — only a visible error can be fixed*, L5911 – L5912) [1] says exactly this.

**Fourth, treating errors as learning material can be trained.** Keith and Frese [18] meta-analysed error management training — training that explicitly encourages errors during practice and guides people to extract information from them — against conventional training, and found the former more effective, especially for transfer to new tasks. This yields an optimistic corollary: soft labelling is not a talent. It is an action that can be arranged.

**The four together are the reason this paper asks for a form rather than issuing an exhortation:** hindsight bias requires it be written before the outcome; distorted self-assessment requires fixed fields rather than free writing; psychological safety requires it be a routine action rather than an ad hoc demand; and error management training shows it can be taught.

> Soft labels have to be written before the outcome. The moment the result lands, the uncertainty that was there is erased.

## 4. The Three Premises of Distillation, Taken One at a Time

A borrowed concept is usable only if its premises hold here. Knowledge distillation has three.

### 4.1 Premise one: the teacher is stronger than the student

**Often false in an organisation.** The people assessing last week's judgements in a review are usually the people who made them. This is not self-distillation either, because in self-distillation the teacher is at least a trained, fixed model; here teacher and student are the same people in two roles in the same week.

More seriously, the teacher's errors are passed to the student intact. A judgement that was wrong but went unnoticed will be written into the prohibition list as correct teacher output, and then read aloud by everyone. **Self-distillation amplifies existing bias rather than correcting it.**

**What can be done is limited but not nothing.** Villado and Arthur [8] in §3.3 show that objective records beat subjective impressions. Keeping a record of *what was in hand at the time* means the teacher is not purely oneself, but oneself plus a timestamp that does not lie.

### 4.2 Premise two: the teacher was trained independently, first

**Often false in an organisation.** In the original paper the teacher is trained beforehand on separate data. A review that uses only the material this week produced is not an independent teacher; it is the same model's echo.

This is exactly where the conclusion of GI-WP-2026-P12 lands. That paper argues that the price of end-to-end is that attribution becomes hard, and therefore that an outside assessor is required. **The technical meaning of "outside assessor" here is: part of the teacher's input has to come from someone who was not involved.**

The website's line 「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] is the one place found here that explicitly puts the teacher outside. Its scope is narrow — only the case of being turned down by others — but the direction is right.

### 4.3 Premise three: the teacher's output is soft

**This is the only one of the three an organisation can guarantee by rule.** The first two depend on people and on whether external input exists; the third depends only on whether the form has that column.

And by the three counts of §2.4, the column is not there. All seven named output fields are of the conclusion type, and the option ranked second at the time appears zero times in the whole text.

**So this paper's recommendation lands on the third premise, not the first two.** Not because the first two do not matter, but because the first two cannot be changed and the third can be changed tomorrow. One rule that can actually be changed beats two principles that are correct and unexecuted.

### 4.4 The soft-label proportion: a quantity that can be counted

Turn the third premise into a number:

> **Soft-label proportion = the number of review records in which the fields "option ranked second", "by how much it fell short" and "the information missing at the time" are all filled, divided by the number of records sampled.**

Three things have to be fixed in the rule or the number will drift on its own:

1. **The sample must be random.** Not the ones that were written up well.
2. **All three fields must be non-empty** to count in the numerator. One field is not enough — naming the runner-up without stating the margin gives the name but not the score, and the dark knowledge is still missing.
3. **"None" is a valid entry, but it has to be written.** If there genuinely was no second option, write "only this one was visible at the time" — which is itself a highly valuable record, because it says how narrow the field of view was.

Compute it quarterly. The trend matters more than the level. **The first figure will almost certainly be low; that is not a problem. It is the baseline.**

记忆点 · KEY POINT

Of the three premises, only "make the output soft" can be changed tomorrow. So change that one first.

## 5. The Tool: A Soft-Label Review Record

### 5.1 Eleven fields

Add one page to the review record. Fill it at the time the judgement is made, **before the outcome is known** (§3.5, first point).

The four constraints deserve a note.

**Fields 8 and 9 exist to counter hindsight bias [14].** Writing the verification date and the attribution starting point in advance removes the option of saying afterwards "I always thought that was off".

**The "none" in field 10 is a deliberate design.** Allowing "none" is what stops the field from being padded. And if a whole quarter is nothing but "none", that does not mean there were no lessons. It means nobody is looking.

**Field 11 exists to counter premise one (§4.1).** If the reviewer is not the person who judged, the teacher gains a little independence. The reviewer does not have to agree with the judgement; he only has to confirm that fields 2 to 6 are not platitudes.

### 5.2 What the sheet costs

One page, a few extra minutes each time. This paper does not claim it is free. Its real cost is not the time: it is that **fields 2, 5 and 6 require a person to admit that he saw another road and did not take it**, while field 11 requires a second person to see that admission. This is why it needs psychological safety [17], and why it has to be written as a routine action rather than an ad hoc request — a routine action is not aimed at anyone.

记忆点 · KEY POINT

Allowing "none" is what stops the field from being padded. A quarter of nothing but "none" does not mean no lessons — it means nobody is looking.

## 6. A Proposition That Can Be Refuted

**Main proposition.** If two organisations review at the same frequency, but one of them carries a higher soft-label proportion in its review output (computed by the rule in §4.4), then after some period the one with the higher proportion performs better in **new situations that differ from past ones**, while in situations closely resembling the past the two do not differ.

The shape of the proposition is deliberate: **the payoff from soft labels should show up in extrapolation, not in repetition.** An organisation that only does repetitive work is well served by hard labels — which is also §7, counter-case two.

**What evidence would refute it.** Among organisations reviewing at comparable frequency, split them by soft-label proportion and compare performance in new versus familiar situations. If the high-proportion group has no advantage in new situations either, the proposition fails — soft labels would then be a recording cost rather than learning material. If the two groups differ in **familiar** situations, the proposition also needs revision: the mechanism would then not be extrapolation but something else.

**Corollary one (gap test, run once here).** If a review practice grew up on its own and was never specified by a form, then its public statements should specify **how often** to review in more detail than **what to record**. Tables 1 and 2 show five granularities with the weekly cadence confirmed in ten statements, against seven named output fields of which the two most frequent refer to the same thing. **Corollary one is not refuted on this one sample.** Take another firm and find its statements specify content more finely than cadence, and the corollary is refuted — the asymmetry would then not be a general feature of self-grown practice.

**Corollary two (timing test, not run).** If soft labels have to be written before the outcome [14], then second-best options reconstructed afterwards should differ systematically from those recorded at the time: the reconstructed ones will sit closer to the eventual result. The test takes one batch of judgements in two groups, one recording at the time and one reconstructing afterwards, and compares the distribution of "second place". This paper does not run it, because it requires a controlled internal trial. **Left explicitly open; not filled with conjecture.**

**Corollary three (softening is not flattening).** §3.2, citing Müller and colleagues [5], notes that softening too evenly destroys dark knowledge. Testable form: if an organisation's fields 2 and 3 are chronically filled with undifferentiated boilerplate, then even a high soft-label proportion should not produce the extrapolation advantage the main proposition predicts. This yields a diagnostic: **a high proportion with no effect — check whether field 3 discriminates at all.**

记忆点 · KEY POINT

The payoff from soft labels should appear in new situations. A difference in familiar ones means something else is doing the work.

## 7. Boundaries and Counter-Cases

An analysis with only supporting cases does not deserve trust. Five places where this practice is known to fail.

**Counter-case one: the premise that the teacher is stronger often fails.** This is the important one; §4.1 stated it, and here is its full consequence. Self-distillation passes the teacher's bias down intact, and passes it down very evenly — because prohibitions are read aloud by everyone. A wrong prohibition is far more damaging than a wrong judgement, because it gets recited repeatedly. **So the sheet needs an exit mechanism alongside it: every prohibition must record which judgement produced it, so that it can be withdrawn when that judgement is later shown to have been wrong.** This is the second of the two recommendations of this paper, alongside the record sheet itself.

**Counter-case two: not all work needs soft labels.** Highly repetitive work under fixed rules is well served by hard labels, and better served by them — every extra field is pure cost. The payoff from soft labels comes from extrapolation, and repetitive work does not need to extrapolate. The test is simple: **ask whether the next situation will differ from this one.** If the answer is "basically the same", the sheet should shrink to fields 1, 8 and 10.

**Counter-case three: written down is not the same as read.** Of the seven output fields counted in §2.4, prohibitions and the stop-doing list account for most, and what the website repeatedly emphasises is that everyone re-reads them (L932, L3065, L4067 and others) [1]. That emphasis is right: **a record nobody re-reads differs from no record only in storage cost.** The sheet proposed here faces the same problem, and faces it more acutely — soft labels are longer and more specific than prohibitions, and therefore harder to re-read repeatedly. One possible remedy is to index by field 4 (what would flip the ranking): rather than reading everything, read the one entry when a similar situation arises. This paper has not tested that remedy.

**Counter-case four: this field will be gamed.** As soon as the soft-label proportion becomes a number someone watches, it will be padded. Villado and Arthur [8] remind us that what works is a review **grounded in objective records**, not a fully completed form. So the proportion is unsuited to being a performance metric and suited only to being a diagnostic — **watch the trend, publish no ranking, attach it to no one's evaluation.** This one is hard: the moment it is attached, fields 2 to 6 become creative writing.

**Counter-case five: a gap in the public text is not a gap in practice.** The absence check in Table 3 establishes that this set of public statements does not prescribe the field. It does not establish that the firm

never records second-best options. Internal records are not visible here, and should not be. **An absence in public text is a conclusion about wording, not a conclusion about behaviour.** This is the same point as §7, counter-case three of GI-WP-2026-P11 and §7, counter-case five of GI-WP-2026-P12: what is not written down does not enter the denominator.

> A wrong prohibition is worse than a wrong judgement. It gets recited by everyone, repeatedly.

## 8. Scope

**One, the object.** This paper analyses sixty-five public statements about review practice in one public text from one professional service firm. Its conclusions apply to the design of reviews in judgement-intensive knowledge work with long feedback cycles and situations that are never quite the same twice. They do not extend to highly repetitive operational work (§7, counter-case two), nor to settings with immediate objective scoring.

**Two, the nature of the evidence.** Tables 1, 2 and 3 are mechanical counts; the rules are in the writing note §6.1, and anyone recomputing them against the same plain text should get the same figures. **These three tables are counts, not evaluations** — they say how many times a term appears in a public text, not whether the firm does its job well. Table 4 is a tool constructed here, not a measurement.

**Three, a known weakness in the counting rule.** Tables 1 and 2 use "occurring on the same line", and line breaks in the plain text depend on the tag structure of the original page. A sentence spanning two lines has its halves counted separately or not at all. The rule was chosen because it is mechanical and recomputable; the price is that it handles long sentences poorly. **Under a different rule — by paragraph rather than by line — individual figures would change, but the three readings in §2.5 rest on orders of magnitude (ten against one, seven against zero), not on any single figure.**

**Four, the limits of the absence check.** Table 3 searches words, not meanings. An organisation could perfectly well express "the option ranked second at the time" in wording this paper did not think of. The full list of search terms is given so that it can be checked and extended, but under-detection cannot be ruled out. **That is a real limitation, stated here rather than hidden.**

**Five, no causal inference.** This paper does not claim that adding the sheet improves any outcome. The main proposition of Section 6 is a claim to be tested, not a conclusion already established.

**Six, an honest gap.** Corollary two — the difference between recording at the time and reconstructing afterwards — requires a controlled internal trial, which this paper cannot run and leaves explicitly open. Until that work is done, the rule that soft labels must be written before the outcome rests on Fischhoff's [14] experimental evidence plus an analogy, not on direct verification in this setting.

**Seven, what this paper does not do.** It does not evaluate any business outcome of the firm, does not constitute investment advice, and gives no information traceable to any specific transaction or company. The three counting tables count words in a public text and involve no operating data.

> Table 3 counts words, not meanings. The search terms are listed in full so that someone can add the ones I missed.

## 9. Conclusion

Back to the two questions of Section 1.

**"What does a review actually record?"** The answer should not be a table of frequencies. Frequency determines how fresh the information is; content determines how thick it is. The counts here show five granularities in the public statements with the weekly cadence confirmed ten times over, against seven named output fields, all of the conclusion type. **What needs adding is not cadence. It is fields.**

**"Why does every new colleague still step into the same hole?"** Because what he inherits is a string of conclusions, not a decision space. A prohibition tells him not to do this. It does not tell him what the other two routes were, or why this one did not work. Someone holding only conclusions, faced with a slightly different situation, has no choice but to try it again himself. **Conclusions can be inherited. Judgement cannot — unless what was visible at the time is kept along with them.**

This paper proposes two things, both small:

**First, add one page of eleven fields to the review record** (Section 5), filled at the time and not after the outcome. The proportion of records with three fields simultaneously non-empty is the soft-label proportion: compute it quarterly, watch the trend, publish no ranking, attach it to no evaluation.

**Second, record for every prohibition which judgement produced it** (§7, counter-case one), so that it can be withdrawn when that judgement turns out to have been wrong. Once a prohibition enters a list everyone re-reads, withdrawing it is far harder than writing it was.

Neither requires a new meeting, a new role or a new system. They require a form and a field.

Three sentences summarise the paper:

1. Recording only the outcome compresses a whole meeting into one word; soft labels are what got compressed away.
2. Of distillation's three premises, only "make the output soft" can be changed tomorrow — so change that one first.
3. Soft labels have to be written before the outcome; the moment the result lands, the uncertainty that was there is erased.

> Conclusions can be inherited. Judgement cannot — unless you keep the options that were visible at the time.

## References

Format: author (year). Title. Journal/publisher, volume(issue), pages. DOI or accessible link. Grouped by class of source: primary material is the firm's own public text; secondary material is peer-reviewed literature and public preprints. Each entry was checked by live request on 2026-09-22 using the check() function of _仓储/_build/refcheck.py; **all 18 are reachable**, and the title of each was compared word for word against the title returned by Crossref or the arXiv API. Details are in the writing note, § 4. No unverified entry is listed.

**Primary material (the firm's public text; every quoted sentence comes from [1], unchanged)**

[1] Glacier Institute (2026). Glacier Institute website: to founders and investors (L907–L912), the stop-doing list (L932, L3064–L3065), the cadence of review (L3227–L3228), Chapter 18 pair 4, "data quality and the teacher model / review practice" (L3886–L3889, L3919), the essay "A review is a metronome, not a fire brigade" (L5864–L5918), and the essay "Treat investors as upstream" (L8615–L8622). https://glacier.mba/institute.html (accessed 2026-09-22)

**Secondary (knowledge distillation and soft labels)**

[2] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. https://arxiv.org/abs/1503.02531 (accessed 2026-09-22)

[3] Buciluǎ, 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 (accessed 2026-09-22)

[5] Müller, R., Kornblith, S., & Hinton, G. (2019). When Does Label Smoothing Help? https://arxiv.org/abs/1906.02629 (accessed 2026-09-22)

**Secondary (the empirical record on after-event reviews)**

[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

**Secondary (organisational learning and the failure sample)**

[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

**Secondary (boundaries: hindsight, self-assessment, psychological safety, error management)**

[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

*Verification note: [1] HTTP 200; [2][4][5] arXiv pages 200, with titles confirmed word for word against the arXiv API ("Distilling the Knowledge in a Neural Network", "Born Again Neural Networks", "When Does Label Smoothing Help?"); [6][7][8][14][15][18] DOIs resolve to 200 with matching Crossref titles; [3][9][10][11][12][13][16][17] return 403 to scripts (anti-scraping, not absence) and were checked against Crossref metadata for title, volume, issue and pages, all matching.* **One correction is on record***: [12] Denrell 2003 was first written with the DOI 10.1287/orsc.14.3.227.15164, inferred from the volume and issue, which returns no metadata; a title search in Crossref gives the correct DOI 10.1287/orsc.14.2.227.15164 (title, volume 14, issue 3, pages 227–243 all matching), and it has been substituted.*

**Cross-references within this series (no DOI yet; cited by number)**

- GI-WP-2026-P9, *Calibration: aligning expectations before the question is asked*
- GI-WP-2026-P11, *Borrowed theories: how frontier-technology concepts explain an investment bank's daily work*
- GI-WP-2026-P12, *Loss at the handover: where the boundary of a full mandate is drawn*

## Sources of the quoted Glacier Institute text

All quotations come from [1] https://glacier.mba/institute.html (retrieved 2026-09-22; plain text obtained by stripping comments, scripts, styles and all tags — 10,079 lines). Line by line:

- Chapter 18 pair 4, theory sentence 「决定模型上限的不是数据量，是数据质量……就补哪一类场景。」— L3887
- Chapter 18 pair 4, phenomenon sentence 「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。」— L3889
- Chapter 18 pair 4, source line "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 (only the first half is quoted here; the second half is unrelated to this paper's argument)

- 「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
- 「信号会馊。」— 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
- All remaining line numbers covered by the counts of Tables 1, 2 and 3 are listed inside those tables.

One note on punctuation: the inner quotation marks at L3065, L5887 and L5907 are half-width double quotes or 「」 in the plain text; quoted inside an outer 「」 here, they are rendered 『』 per Chinese convention. **Only the shape of the quotation marks changed; not one character did.**

English renderings in italics are translations supplied for readers of this edition. The Chinese is the original and governs.

Redaction note: none of the sixteen quoted sentences names any client company, so no "a certain company" rewriting was required; rewrites = 0. The denominators of Tables 1 to 3 exclude five lines of client and third-party quotation, as recorded in the writing note §5.1; they are excluded because they are not the firm's own statements, not because they carry a company name. Nothing in this paper's own prose describes any feature traceable to a specific company. See the writing note, §5.