Distillation by Prompt: Transferring Frontier Calibration to On-Premise LLM Deployments

Tom Pawelek
PATENT LAB
Mississippi State University
tompaw@tompaw.net

Abstract*—Frontier cloud language models deliver strong performance on specialized reasoning tasks but introduce cost, dependency, latency and data-residency constraints that are prohibitive in privacy-sensitive domains such as consumer lending. On-premises open-weight models remove those constraints but commonly underperform on the same task given the same instructions. We present *distillation-by-prompt*, a training-free transfer method: run a frontier teacher and an on-premises student over a shared evaluation set, diagnose where the student diverges, and encode the divergences as an explicit verification checklist appended to a student-specific rubric. Measured by agreement with the teacher, the method is highly effective—on a bank-statement screening task for installment lending it doubles exact-score agreement, raises agreement within ± 2 points to 91%, and eliminates a $+1.36$-point optimistic bias. We then re-evaluate the same outputs against *outcome-linked* labels—the terminal decisions of human underwriters—using the area under the ROC curve, and reach a different conclusion about what the method does. Discrimination does not improve ($0.743 \rightarrow 0.708$; paired 95% CI $[-0.102, +0.033]$), and the *unsupplemented* student already matched the teacher (0.743 vs. 0.742): there was no capability gap to close. The method transfers *calibration*—the mapping from latent judgment onto the output scale—which is what makes a model substitutable into a pipeline whose thresholds were tuned for another, but is not the capability gain that agreement metrics appear to show. We give the estimator and its properties, separate two evaluation regimes whose conflation produces this error, and report that across three model families the dominant error source was the rubric rather than any model. We also list other applicable domains, which - due to its similar input structure - could benefit from our solution. Those fields vary from medical, financial to HR and admissions.

***Index Terms*—large language models, prompt engineering, knowledge distillation, on-premises inference, model evaluation, ROC analysis, calibration**

I. Introduction

Deploying large language models (LLMs) for specialized decision support presents a recurring tension. Frontier cloud-hosted models deliver strong qualitative reasoning, but their use requires transmitting inputs to a third-party API, which is problematic when those inputs contain personally identifiable information or regulated financial data. Open-weight on-premises models eliminate the data-egress problem and reduce marginal cost to zero, but commonly produce weaker outputs on the same input with the same prompt.

A natural response is knowledge distillation [1]: train a smaller student to imitate a larger teacher. Classical distillation requires model internals, a training pipeline and a labeled corpus. For teams operating a single specialized task with a few dozen to a few hundred examples, that overhead is disproportionate.

We describe a lighter alternative, *distillation-by-prompt*, in which the frontier model generates not training signal for the student's weights but *targeted instructional content* for its prompt. Section III gives the procedure and Section V shows that by conventional agreement metrics it works well.

The larger part of this paper concerns a question those metrics cannot answer: *did the student get better at the task, or merely better at agreeing with the teacher?* The distinction is not academic. A screening model's operational value lies in ranking applicants, not in reproducing another model's number. Two models can agree closely and rank badly, or disagree systematically and rank identically. Score agreement conflates these because it is computed at fixed score values; the natural instrument for separating them is a rank statistic.

Our contributions are:

1) The distillation-by-prompt procedure and its agreement results (Sections III, V).
2) A formal treatment of the evaluation metric (AUC as a Mann–Whitney statistic) with its invariance properties, the discretization ceiling imposed by a discrete output scale, and a paired-bootstrap procedure for comparing models on shared inputs (Section IV).
3) Evidence that the method transfers **calibration rather than capability**, the student's discrimination being at teacher parity before any supplement (Section V).
4) A cross-vendor finding that the dominant error source is the *rubric*, not the model: statistical findings written as unconditional rules are followed literally by newer models and silently overridden by older ones (Section VI).
5) Two methodological cautions from live measurement—a selection-induced regression artifact and a demonstration of the winner's curse—and a separation of imitation and estimation regimes that explains why specification defects are invisible to teacher-referenced evaluation (Sections VII, VIII-A).

II. Application Context

A. Task

The task is risk screening of consumer bank statements for a short-term installment loan product (average principal

TABLE I
CORRELATION OF CANDIDATE NUMERIC SIGNALS WITH UNPROFITABILITY

Signal	Correlation	Verdict
Income transaction count	-0.226	Strongest
Account history length	-0.157	Strong
Payday loan count (90d)	-0.148	Reversed
Income variance	-0.083	Reversed
Negative balance %	-0.019	No signal
NSF / returned checks	-0.016	No signal

$\approx$ \$650, $\approx$ 11 biweekly payments). Each input is an Instant Bank Verification (IBV) pull containing 50–750 transactions spanning 30–180 days. The screening model must assess (a) whether deposit activity represents genuine employment income or loan proceeds from other lenders, (b) whether cash flow is adequate to support repayment, and (c) whether qualitative risk patterns such as gambling, concurrent lender stacking or pass-through account behavior are present. The output is an integer confidence in $[0, 10]$ with a list of identified signals and supporting evidence.

B. Data Signature

In general, our method requires the following data characteristics:

1) **Longitudinal event stream:** Here we have hundreds of independent bank transactions over 180 days with freeform text description.
2) **Graded score collapsed to binary:** Threshold for auto-decisions, training signals from manual processing.
3) **Weak numeric correlations:** Our aggregates aren't sufficient for decision-making.
4) **Training labels are human decisions:** We use a portfolio of tens of thousands of loans to run our analysis.
5) **Negative class is censored:** Denied leads do not produce loans - lack of signals to observe.

C. Why Qualitative LLM Analysis

Analysis of 45,983 historical loans (filtered to accounts at least 180 days mature, from an extraction of 53,903) found that traditional underwriting signals have minimal predictive lift in this population. Table I gives Pearson correlations between candidate signals and a binary $is\_profitable$ label. In a universally credit-stressed population, classical red flags such as non-sufficient-funds (NSF) fees and negative-balance episodes are the baseline condition rather than differentiators, and some signals invert. The remaining lift sits in qualitative transaction-description analysis, which aggregate numeric features cannot capture. That same analysis supplies the rubric's calibration content and, as Section VI shows, is the source of its principal defect.

D. Two-Phase Architecture

Each lead passes through a deterministic Phase 1 rule engine that extracts roughly 25 numeric features (balance metrics, income regularity, obligation indicators, behavioral ratios) and packages them, with a PII-stripped transaction list, into a structured JSON summary. Phase 2 passes that summary to an LLM together with a scoring rubric encoding the historical findings and the output schema. Phase 1 is a signal-enrichment layer rather than a gatekeeper. All work described here operates within Phase 2.

E. Model Selection

We evaluated eight candidate Phase 2 models on a held-out stress case: a checking account with 756 transactions over 134 days exhibiting extreme loan cycling across seven or more concurrent short-term lenders. The frontier baseline was Claude Sonnet 4.6. On-premises candidates included Gemma 3 27B, DeepSeek-R1 32B, Qwen 2.5 32B, Llama 3.3 70B, Qwen3 30B-A3B, Qwen3 235B-A22B and Llama 3.1 8B, served through Ollama on a workstation with 256 GB system RAM and an RTX 4090, at 4-bit quantization. Only the Gemma family produced an assessment substantively matching the frontier output, naming the cycling pattern explicitly; it was also the fastest tested and fit entirely in VRAM. We adopt it as the on-premises student for the remainder of the study.

III. METHOD: DISTILLATION-BY-PROMPT

A. Overview

The method assumes (i) a production rubric R_0 already tuned for a frontier teacher M_T, (ii) an on-premises student M_S, and (iii) an unlabelled evaluation set $\mathcal{E}$ of task inputs. It produces a student-specific rubric $R_S = R_0 \oplus \Delta_S$, where Δ_S targets the failure modes of M_S. Fig. 1 summarizes the procedure.

B. Procedure

1) **Dual inference.** Run M_T and M_S against every $x \in \mathcal{E}$ under the shared rubric R_0, recording confidence, signals and free-text reasoning.
2) **Divergence filtering.** Identify $\mathcal{D} = \{x \in \mathcal{E} : |s_T(x) - s_S(x)| > \tau\}$, where s is the confidence score and τ a divergence threshold (we used $\tau = 2$).
3) **Blind-spot extraction.** For each $x \in \mathcal{D}$, note which patterns appear in the teacher's output but not the student's, and cluster recurring misses into categories.
4) **Supplement authoring.** Encode each category as a mandatory verification step with concrete lexical anchors (lender names, gambling brands, quantitative thresholds) and a directional scoring constraint. Append the block Δ_S to R_0.
5) **Validation.** Re-run M_S under R_S, recompute the target metric, and accept R_S if the threshold is reached.

C. The Calibration Checklist

For our student, divergence analysis clustered into five blind spots, encoded as checks the model must perform before assigning any score above 5: *income legitimacy* (are credits from lending companies rather than employers?); *salary plausibility* (declared salary against observed deposit amounts and

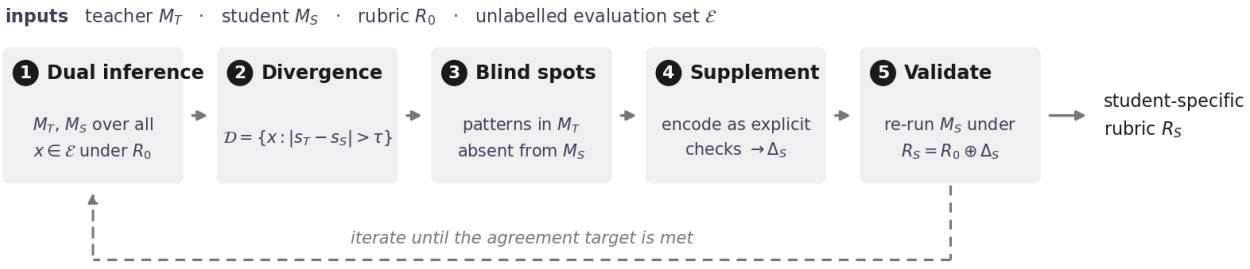

Fig. 1. The distillation-by-prompt procedure. The method consumes a teacher M_T, a student M_S, the production rubric R_0 and an unlabelled evaluation set $\mathcal{E}$, and produces a student-specific rubric $R_S = R_0 \oplus \Delta_S$ in which Δ_S is a calibration block targeting the student's divergences. No weights are modified and no labels are required.

cadence); *gambling detection* (named betting platforms and generic wagering descriptors, treated as distinct from baseline financial stress); *concurrent lender count* (5+ distinct lenders forces a score of 3 or below); and *pass-through detection* (high transaction volume with near-zero average balance). The production rubric R_0 used with the teacher is unchanged; only R_S carries the supplement.

IV. Evaluation Methodology

A. Why a Rank Statistic

Let $s(x) \in \{0, 1, \ldots, 10\}$ be the model's confidence for lead x. Score-agreement metrics compare $s_S(x)$ to $s_T(x)$ at fixed values. Any monotone shift in one model's scale (the exact effect a recalibrating supplement is designed to produce) registers as a large change in agreement while leaving the induced ranking untouched. Conversely, a model may match the teacher's scores closely while ordering applicants poorly.

We therefore adopt as the primary metric the probability that the model ranks a positive above a negative.

B. Definition and Estimator

Let $Y \in \{0, 1\}$ denote the binary outcome label, with the positive class defined in Section IV-F. Write $X^+ \sim (X \mid Y = 1)$ and $X^- \sim (X \mid Y = 0)$. Then

$$\mathrm{AUC} = \Pr\big[s(X^+) > s(X^-)\big] + \tfrac{1}{2}\Pr\big[s(X^+) = s(X^-)\big]. \tag{1}$$

This is the area under the curve traced by $(\mathrm{FPR}(t), \mathrm{TPR}(t))$ as the decision threshold t sweeps the score range, where

$$\mathrm{TPR}(t) = \Pr[s(X) \geq t \mid Y = 1], \tag{2}$$

$$\mathrm{FPR}(t) = \Pr[s(X) \geq t \mid Y = 0]. \tag{3}$$

Both rates are computed *within* a class, which is the source of the prevalence invariance noted below. The equivalence of (1) to the ROC area is due to Bamber [3]; the equivalence to the Mann–Whitney U statistic [2] gives the estimator used throughout,

$$\hat{A} = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \left[\mathbf{1}\{s_i^+ > s_j^-\} + \tfrac{1}{2}\mathbf{1}\{s_i^+ = s_j^-\} \right], \tag{4}$$

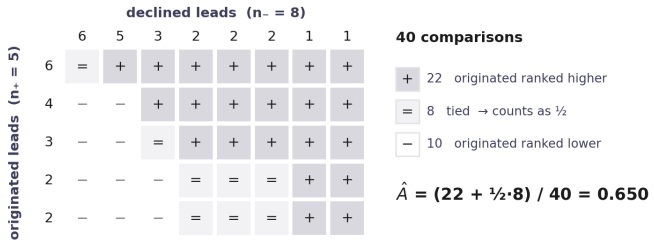

Fig. 2. The estimator of (4) on a 5×8 subset of the cohort. Every originated lead is compared with every declined lead; $\hat{A}$ is the share of those 40 comparisons the model orders correctly, with ties counted as one half. The subset is drawn at random subject to exhibiting all three outcomes and to its own $\hat{A}$ (0.650) tracking the full cohort's (0.645), so that it illustrates the estimator rather than flatters it. Note that ties arise purely from the discrete score: eight of forty comparisons here carry no information by construction.

computable in $O(n \log n)$ via midranks as $\hat{A} = (R_+ - n_+(n_+ + 1)/2)\,/(n_+ n_-)$, with R_+ the sum of midranks of the positive class.

Fig. 2 shows (4) evaluated by hand on a small subset.

With a discrete score, ties are a large share of all pairs; assigning them $\frac{1}{2}$ keeps $\hat{A}$ unbiased under the null, whereas discarding them inflates $\hat{A}$ whenever the two classes tie asymmetrically.

C. Invariance Properties

$\hat{A}$ is invariant under any *strictly monotone* reparameterization of the score. Table II verifies this on our production cohort and marks the boundary of the property: transformations that *merge* distinct values are monotone but not strictly so, and they change $\hat{A}$ by destroying ordering information.

$\hat{A}$ is additionally invariant to class prevalence, unlike accuracy or precision. This has a practical consequence we exploit in Section VII: because precision is governed by the smaller class, deliberately over-sampling the minority class costs nothing statistically and materially tightens the interval.

D. The Discretization Ceiling

A rubric that emits an integer in $[0, 10]$ admits at most nine distinct thresholds, so the empirical ROC curve is a nine-vertex

TABLE II
INVARIANCE OF $\hat{A}$ UNDER SCORE TRANSFORMATIONS ($n = 12{,}884$)

Transformation of s	$\hat{A}$	**Invariant**
none (raw integers)	0.6449	—
$s + 2$ (additive shift)	0.6449	yes
$10s + 7$ (affine rescale)	0.6449	yes
s^2 (monotone on domain)	0.6449	yes
$\lfloor s/4 \rfloor$ (3 buckets)	0.6009	**no**
$\mathbf{1}\{s \geq 5\}$ (binarized)	0.5710	**no**

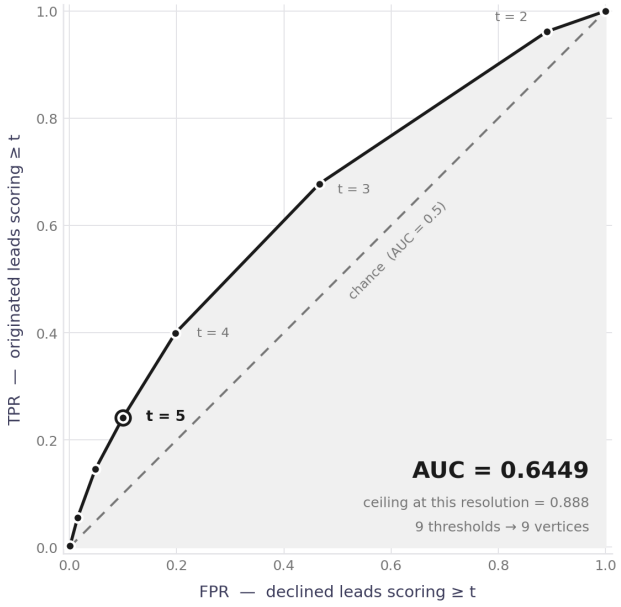

Fig. 3. Empirical ROC curve for the production screening model ($n = 12{,}884$; 3,885 originated, 8,999 declined). Each vertex is one decision threshold t; the ringed point is the approve/reject boundary used downstream, at which the model captures 24.2% of originated leads while passing 10.0% of declined ones. Because the rubric emits an integer, nine thresholds yield nine vertices: the polygon is the discretization ceiling of Section IV-D made visible. The vertical gap between the curve and the chance diagonal is the discriminative power.

polygon rather than a smooth curve, as Fig. 3 shows. On our cohort, ties constitute 22.4% of all $34{,}961{,}115$ pairs. Even if every non-tied pair were ordered correctly, $\hat{A}$ could not exceed

$$\hat{A}_{\max} = \frac{n_{\text{win}} + n_{\text{loss}} + \frac{1}{2} n_{\text{tie}}}{n_+ n_-} = 0.888. \tag{5}$$

Observed values should be read against this ceiling, not against 1.0. Requesting a continuous score in place of an integer is the cheapest available improvement to measurement resolution, since it changes the output schema rather than the model.

E. Inference and Model Comparison

For a single $\hat{A}$ at large n we use the Hanley–McNeil standard error [4],

$$\begin{aligned} \mathrm{SE}^2 = \big[& A(1-A) + (n_+ - 1)(Q_1 - A^2) \\ & + (n_- - 1)(Q_2 - A^2) \big] / (n_+ n_-), \end{aligned} \tag{6}$$

with $Q_1 = A/(2-A)$ and $Q_2 = 2A^2/(1+A)$. At the sample sizes typical of a per-model evaluation set we prefer the bootstrap [7], which assumes no particular sampling distribution.

Comparing two models on the same inputs requires a *paired* procedure. Account difficulty is a large shared variance component; resampling the same accounts for both models cancels it, in the spirit of DeLong et al. [5]:

for $b = 1$ **to** B **do**
  resample account indices $\mathcal{I}_b$ with replacement
  $\delta_b \leftarrow \hat{A}_{M_1}(\mathcal{I}_b) - \hat{A}_{M_2}(\mathcal{I}_b)$
end for
report the 2.5th and 97.5th percentiles of $\{\delta_b\}$

Comparing two independently computed intervals (or, worse, a sample estimate against a population constant) measures the difficulty mix of the two samples as much as the models. Section VII reports what happens when this is neglected.

F. Labels and Their Limits

The positive class is *originated*: the lead passed human underwriting review, the applicant accepted, and funding completed. The negative class is terminal non-origination. Leads still in flight are excluded. Four caveats follow the paper throughout.

First, this measures agreement with a human policy, not correctness: underwriters use signals the model never sees, including third-party credit scores and employment verification. Second, the negative class has no counterfactual: declined applications never become loans, so profitability is unobservable for roughly 70% of the population, and no reanalysis of existing data can recover it. Third, label noise is asymmetric: the negative class conflates credit rejection, operational rejection and customer dropout. Fourth, precision is bounded by the minority class, which at typical evaluation-set sizes yields intervals of order ± 0.1.

V. RESULTS

A. Setup

We evaluate 140 production leads for which (i) the teacher (Claude Sonnet 4.6) had scored the lead in the live system, (ii) the student had been re-scored under both the base rubric R_0 and the supplemented rubric R_S, and (iii) a terminal underwriting decision exists. Of these, 37 were originated. Teacher and student scores are compared to each other by conventional agreement metrics, and each is compared to the outcome label by $\hat{A}$.

B. Result

Table III is the central result of this paper. The supplement reproduces the previously reported agreement gains almost exactly: exact agreement doubles, ± 2 agreement rises to 91%, and the optimistic bias inverts to a small conservative lean. On the same leads, discrimination does not improve. It declines by 0.034, an amount indistinguishable from zero at this sample size but certainly not an improvement.

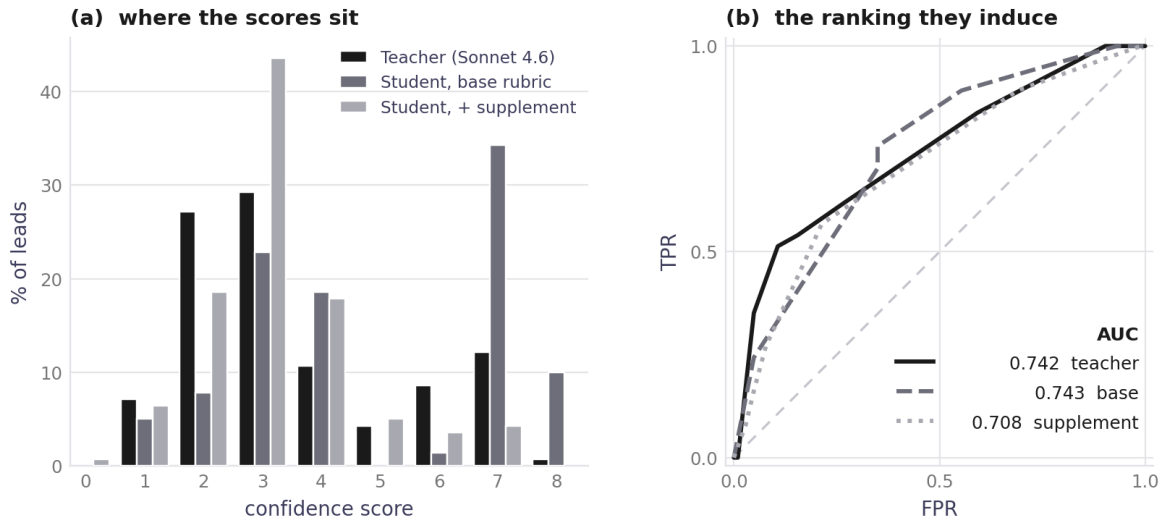

Fig. 4. Calibration moves; discrimination does not ($n = 140$, 37 originated). (a) The base student concentrates its scores at 7–8 where the teacher concentrates at 2–3; the supplement relocates the student onto the teacher's part of the scale, which is what the agreement metrics in Table III register. (b) The rankings those scores induce are nearly identical, and the base student's is indistinguishable from the teacher's (0.743 vs. 0.742). The supplement changed where judgments are placed, not their order.

TABLE III
AGREEMENT WITH TEACHER VS. DISCRIMINATION AGAINST OUTCOME
($n = 140$, 37 ORIGINATED)

	Exact	± 2	**Bias**	$\hat{A}$
Student, base rubric R_0	25%	74%	$+1.36$	**0.743**
Student, supplement R_S	50%	91%	-0.34	**0.708**
Teacher (Sonnet 4.6)	—	—	—	**0.742**

Fig. 4 shows the same result directly: panel (a) is what agreement measures, panel (b) is what it cannot. The decisive number is in the last column of the first and third rows. **The unsupplemented student already discriminated as well as the teacher** (0.743 vs. 0.742). There was no capability gap to close. What differed was where the student placed its judgments on the output scale: its base score distribution was concentrated at 7–8 where the teacher's was concentrated at 2–3, producing the $+1.36$ bias and the low agreement, while inducing essentially the same ordering.

Paired bootstrap confirms that neither the base nor the supplemented student differs from the teacher in discrimination: supplement $-$ base is -0.034 (95% CI $[-0.102, +0.033]$) and supplement $-$ teacher is -0.033 (95% CI $[-0.092, +0.024]$).

C. Interpretation

The method works, but not by the mechanism previously claimed. What the checklist supplies is not missing analytic capability (the student's ranking was already teacher-equivalent) but instead a shared *scale*: the checklist tells the student which patterns warrant which score band, aligning its calibration to the teacher's.

This is operationally valuable and should not be dismissed. A model whose scores are on a known scale can be substituted into a pipeline whose thresholds, alerts and automation rules were tuned for another model, without re-tuning any of them. That is precisely the deployment problem the method set out to solve. In this context, we gain interoperability (calibration match) between the frontier and on-premises models, not a direct capability match, which is a commercially sufficient outcome for our application.

It also suggests our original diagnosis was partly mistaken. Divergence analysis attributed the gap to the student failing to *look for* patterns the teacher found. An alternative account consistent with these data is that the student found the same patterns and weighted them onto a different scale. Distinguishing these requires signal-level rather than score-level comparison, which we leave to future work.

VI. THE RUBRIC AS THE DOMINANT ERROR SOURCE

A. A Defect Class

The rubric encoded the historical analysis as a list of factors that are "NOT predictive (do not penalize)": non-sufficient-funds events, negative-balance frequency, income variance; together with a reversed-signal claim that declining balance trend correlates with better outcomes. Each of these was measured *one variable at a time*. Written as unconditional rules, they were read as licence to dismiss those factors *in combination*. The rubric additionally instructed the model to centre its confidence at 6–7 on the grounds that $\approx 60\%$ of loans are profitable—a statistic computed on originated loans, applied to a pre-screening population of which only 29.7% is ever originated.

B. Model-Dependent Consequences

Table IV shows the effect of narrowing these statements to their single-variable scope. Two newer models from different

TABLE IV
EFFECT OF SCOPING SINGLE-VARIABLE FINDINGS, BY MODEL FAMILY

Model / rubric	Reversals	Justified	Severe flags
Sonnet 5, original	24	25%	10
Sonnet 5, scoped	9	**44%**	26
GLM-5.2, original	21	24%	17
GLM-5.2, scoped	4	**50%**	27
population base rate	—	*24%*	—

vendors follow the original instructions literally, producing decision reversals at exactly the population base rate (that is, carrying no information) and almost entirely suppressing the rubric's highest severity tier. Narrowing the exemptions restores both, independently, in each family.

The incumbent model, Sonnet 4.6, was unaffected by the same correction ($\hat{A}$ $0.619 \rightarrow 0.643$, paired 95% CI $[-0.051, +0.096]$; severe flags $123 \rightarrow 124$; zero decision reversals). Inspection of its outputs explains why: it had been violating both instructions all along, flagging near-zero balances at the highest severity despite the prohibition and centring its confidence at 2.95 against an instructed 6–7.

C. Implication

An instruction-following improvement is not automatically a capability improvement, and can present as a regression when the instructions are wrong. Models that silently override a defective rubric will appear to outperform models that comply with it, and the apparent difference will be attributed to the model. Rubric defects should therefore be treated as a confound in any model comparison, and rubric text should be audited for scope: particularly wherever a statistical finding has been transcribed as a rule.

VII. TWO MEASUREMENT PITFALLS

A. Selection-Induced Regression

A prior engine release raised the LLM output-token cap after it was found that a substantial fraction of reviews were truncating mid-response and being recorded as parse failures. Measured $\hat{A}$ fell from 0.726 to 0.644 across that release, which was initially read as a regression caused by an intervening rubric change.

Table V stratifies the post-fix cohort by input length. Discrimination degrades monotonically with statement size, and the extreme quartiles do not overlap. Truncation had been concentrated in long statements; the pre-fix cohort was therefore implicitly filtered to short, easy accounts, and its $\hat{A}$ of 0.726 is almost exactly the post-fix Q_1 value of 0.728. The fix recovered the hard quartiles and the measurement became honest: nothing regressed.

Input length is thus a covariate of the metric, with a spread (0.114) larger than any model difference we measured. Samples drawn for model comparison must be paired on identical inputs or matched on length.

TABLE V
$\hat{A}$ BY INPUT-LENGTH QUARTILE (POST-FIX COHORT)

Quartile	Input tokens	n	$\hat{A}$ (95% CI)
Q_1	2,086–14,462	3,221	0.728 [0.704, 0.751]
Q_2	14,462–22,674	3,221	0.661 [0.640, 0.682]
Q_3	22,674–32,486	3,221	0.627 [0.607, 0.648]
Q_4	32,491–44,269	3,221	0.614 [0.594, 0.635]

TABLE VI
CANDIDATE MODELS, PAIRED AGAINST THE INCUMBENT

Configuration	$\hat{A}$	δ	95% CI
Incumbent + scoped rubric	0.643	$+0.025$	$[-0.051, +0.096]$
Incumbent + adaptive think.	0.672	$+0.053$	$[-0.042, +0.159]$
GLM-5.2 + scoped rubric	0.669	$+0.002$	$[-0.113, +0.131]$
Sonnet 5 (selection sample)	0.748	$+0.127$	$[-0.018, +0.290]$
Sonnet 5 (fresh sample)	0.737	$+0.027$	$[-0.062, +0.114]$

B. The Winner's Curse

We evaluated three rubric variants on one 50-lead sample and selected the best, which measured $\delta = +0.127$ against the incumbent ($p = 0.047$). On a fresh, class-balanced 100-lead sample the same configuration measured $\delta = +0.027$ ($p = 0.26$). Its own $\hat{A}$ replicated ($0.748 \rightarrow 0.737$); what failed to replicate was the *difference*, because the incumbent scored 0.621 on the first sample and 0.710 on the second: a swing driven by the length covariate above.

Selection and evaluation on the same sample are not independent, and the resulting optimism is of the same order as the effects being measured. Table VI reports all candidates on the paired basis. No candidate demonstrably outperforms the incumbent.

VIII. DISCUSSION

A. Two Evaluation Regimes

The results above are easiest to read once two distinct problems are separated, because they call for different metrics and admit different conclusions.

Imitation. A student is made to reproduce a teacher's behavior on a fixed task specification. The reference is another model's output: available on demand for any input, in unlimited quantity, with no missing values and no censoring. Agreement with the reference *is* the objective, so agreement metrics are correct here. The teacher bounds the metric by construction. This is the regime in which the method was originally developed and validated.

Estimation. A configuration is chosen to best predict an external quantity. The reference is an outcome or decision: scarce, delayed, noisy, and (in our setting and many others) censored, since declined applications never produce an outcome. No model is the target, so a rank metric is appropriate, no candidate is bounded by any other, and the specification is itself a variable. This is the regime of the present paper.

Three consequences follow.

First, a metric from one regime cannot support a claim belonging to the other. The agreement figures of Section V are correct measurements of imitation; the difficulty is that "closing the frontier gap" is an estimation claim, and imitation metrics cannot evidence it. Table III shows the two dissociating cleanly on the same leads.

Second, and more consequentially: **the imitation regime is structurally blind to defects in the task specification.** Teacher and student read the same rubric, and the metric compares them only to one another. A specification error is inherited by both and therefore invisible: agreement is unaffected by it. The exemption defect of Section VI persisted undetected across months of production operation precisely because every evaluation was referenced to a model that was silently overriding it. Only external labels exposed it.

Third, the regimes fail differently, so they need different safeguards. Imitation risks a student that agrees for the wrong reasons: the recalibration-as-capability confusion documented here. Estimation risks selection effects: the population shift and winner's curse of Section VII. A pipeline that runs both, as most deployed systems eventually do, needs both sets of controls.

B. What Transfers, and What Does Not

Taken together, Sections V and VI suggest that in rubric-driven screening tasks the binding constraint is frequently neither the student model's reasoning nor the teacher's superiority, but the specification. The student already ranked as well as the teacher; the supplement aligned its scale; and the largest single behavioral effect we measured came from correcting two sentences of rubric text that had been silently misapplying single-variable statistics.

This reframes prompt-level distillation as a *scale-alignment* technique. As the open-weight models evolve, their reasoning skills will naturally improve. At the same time, the best performing on-premise MoE models at the time of our experiment (GLM 5.2 and K3) do require aggressive quantization and optimized expert selection in order to be usable at a business-grade infrastructure (96GB VRAM Blackwell chipsets, which we use in our lab). Distillation helps with those optimizations by providing clear signals for an expert router.

The resource-oriented distillation of these next-generation open-weight models targeting Blackwell GPUs is a subject of our follow-up research.

C. Practical Guidance

Three recommendations follow. Evaluate rank, not agreement, whenever outcome-linked labels exist; agreement alone cannot distinguish a recalibration from an improvement. Pair every comparison on identical inputs, and hold out a fresh sample for confirmation after any selection step. Audit rubric text for statements that convert a measured correlation into an unconditional rule, since newer models comply with such statements literally.

D. General Applications

Based on the input data signature described in previous sections, we could apply our distillation protocol to various real life scenarios, such as:

1) **Clinical deterioration:** early-warning scoring with escalation decision, addressing the common treatment paradox.
2) **AML warning:** transaction monitoring scoring with SAR submission as a binary decision.
3) **Insurance claim triage:** claim history, adjuster notes, timed repair invoices, decide whether to pay without investigation.
4) **VC, grant and litigation funding:** mirrors our winner's curse and no signals on unfunded cases.

IX. Limitations

The labels are human decisions rather than loan outcomes, with the four caveats of Section IV-F; the negative class in particular admits no counterfactual. The re-evaluation in Section V uses 140 leads from a single product and geography, of which 37 are positive, and the reported AUC differences are not individually significant: the claim that discrimination did not *improve* is well supported, but the apparent 0.034 decline is not established. The outcome-linked evaluation uses a student generation adjacent to, rather than identical with, the one on which the checklist of Section III was authored; the agreement figures reproduce closely (exact agreement $22\% \rightarrow 46\%$ on the original 68-lead development set against $25\% \rightarrow 50\%$ here), but the correspondence is not exact. Finally, the discretization ceiling of 0.888 applies to all figures reported here, and comparisons across studies using continuous scores are not directly commensurable.

X. Conclusion

Distillation-by-prompt is a lightweight, training-free technique for aligning an on-premises open-weight model to a frontier model's specialized screening behavior. On a production bank-statement screening task it doubles exact-score agreement with the frontier reference, raises within-± 2 agreement to 91%, and eliminates a 1.36-point optimistic bias, using only a differential analysis of paired outputs and a five-item calibration checklist.

Evaluated against outcome-linked labels, shows that discrimination alone is not the target of our distillation. With carefully prepared baseline rubrics, unsupplemented student are close to matching the teacher's reasoning. What the checklist supplies is a shared *scale*, not missing analytic capability. A model whose scores are on a known scale can be substituted into a pipeline whose thresholds were tuned for another.

Two further findings point the same way. Across three model families the dominant error source was the rubric rather than any model, and correcting two sentences of it produced a larger behavioral change than any model substitution we tested. And an evaluation referenced only to another model cannot detect errors in the specification that both models share,

which is why that defect survived months of production operation. In rubric-driven specialized tasks, the binding constraint is frequently neither the student's reasoning nor the teacher's superiority, but the specification: and finding that out requires labels that come from outside the models being compared.

REFERENCES

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[2] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[3] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, vol. 12, no. 4, pp. 387–415, 1975.

[4] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[5] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[6] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[7] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall, 1993.

[8] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24,824–24,837.

[9] Google DeepMind, "Gemma 3 technical report," Technical report, 2025.

[10] Ollama, "Ollama: Run large language models locally," https://ollama.com, accessed 2026.