

The Word, Not the Number: Counterfactual Sweeps Expose a Status Shortcut in Laya, a Fine-Tuned Typed-Decision Encoder

Mathew Varghese
mathewvarghesemanu@gmail.com
Independent Researcher
USA

Abstract

Encoders fine-tuned to make typed decisions over structured agent state offer a non-generative alternative to prompting a large language model. Many of these decisions should depend on the *size* of a number: how many days until an invoice is due, or how far an invoiced price deviates from the purchase order. We use counterfactual sweeps and dependency ablations to examine numerical behaviour that aggregate held-out accuracy alone does not explain. We study Laya, a ModernBERT-large decision encoder, before and after fine-tuning on the typed-decisions benchmark, where fine-tuning raises agreement with teacher labels from 0.36 to 0.77. Across 147,136 counterfactual evaluation records, the fine-tuned model exhibits sharp, largely context-stable boundaries that are absent for base-model urgency and less consistent for base-model reconciliation. Urgency switches near `days_until_due = 0`, but the switch is carried by the co-varying word "overdue". When that word is held fixed, the sweep's effect on expected urgency drops from 0.71 to 0.11 levels, and the step at zero disappears ($0.60 \rightarrow -0.02$). An invoice labelled overdue stays urgent even when it is due in 75 days. Price reconciliation is predominantly equality-sensitive: all 68 states exhibit a decision change during the sweep, and severity calls nonzero discrepancies of at most 0.1% "Material" at 93% of evaluated points. Residual sensitivity is weakly associated with character edit distance. Status and the sign of `days_until_due` agree on all 400 invoices, preventing on-distribution evaluation from distinguishing those two cues, although teacher urgency labels are graded. The smallest nonzero price variance in the single-line invoices is approximately 3%, leaving smaller tolerances untested by the original benchmark. Two single-seed retrains retain the near-zero urgency step; a separate written-rule probe shows incomplete rule adherence. These findings support a status-dominated urgency shortcut and limited price-magnitude sensitivity, without establishing an exclusively string-based mechanism. We present the protocol as a template for auditing numerical decisions.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Natural language processing*; • **Software and its engineering** → Software verification and validation.

Keywords

shortcut learning, counterfactual evaluation, numeracy, decision boundaries, fine-tuning, underspecification, interpretability, Laya, ModernBERT

1 Introduction

Small encoders fine-tuned to answer typed questions (choices, scores, yes/no) about a JSON state are an attractive component of agent pipelines. They are fast, they are deterministic, and they return a full probability distribution rather than a sampled string. Specialised non-generative predictors have been proposed as faster alternatives to prompting an LLM in other settings too, such as real-time sales-conversion prediction [12]. Laya [1, 2] is a recent example of such an encoder. It is a ModernBERT-large encoder [18] with a decision head that scores each answer option, and fine-tuning on the typed-decisions benchmark [11] roughly doubles its agreement with teacher labels. Numbers pervade the states these models read: due dates, amounts, counts, error tallies. A reasonable user would take a jump in accuracy to mean the model has learned *how the number should move the decision*.

This paper shows, for one model and one workflow, that the inference need not follow. Numerical fields can co-vary with derived cues, such as a status string determined by the sign of a due-date offset. If a numerical decision rule depends only on that sign, an equivalent status-based rule gives the same predictions wherever the dependency holds. Held-out accuracy on that distribution cannot distinguish the two implementations. A graded urgency rule, however, need not agree with either rule. This distinction connects our case study to underspecification [3] and shortcut learning [6]. What makes the numerical case worth studying on its own is that the two hypotheses give different answers to a question practitioners actually ask: *where* does the decision switch, and *how much* does the number matter?

We answer that question with **counterfactual sweeps**. A number is intervened on, $\text{do}(x = x')$ [13], over a dense grid with documented dependent fields updated and other fields held fixed, and the full output distribution is recorded. We pair each sweep with a **dependency ablation**: the same sweep, but with the documented dependent field frozen at its original value. The consistent sweep measures the model's response to the joint numerical and dependent-field change. The ablation measures the model's sensitivity to the number at the original fixed cue value.

Contributions.

- A protocol for auditing numerical decisions in fine-tuned classifiers. It combines dense counterfactual sweeps that respect documented field dependencies, dependency ablations, adaptive refinement around transitions, and boundary metrics: counterfactual flip distance, flip point x^* , slope, entropy, margin, monotonicity violation and boundary consistency. Statistics are computed over base states as the sampling unit (§4).

- Evidence of sharp boundaries after fine-tuning (§5.1), a status-dominated urgency response (§5.2), and predominantly equality-sensitive price reconciliation with limited severity scaling (§5.3).
- An account of which candidate rules the observed data cannot distinguish, and why this does not imply equivalence to graded teacher urgency labels (§5.4).
- Exploratory checks of training-support removal (§5.6) and written numeric thresholds (§5.7), with their inferential limits made explicit.

2 Related Work

Shortcuts and underspecification. Models often succeed on i.i.d. test sets by exploiting features that are predictive in the data but are not the intended ones [6]. D’Amour et al. [3] show that pipelines with equivalent held-out performance can differ widely off-distribution. We examine a numerical instance where sign-based and status-based candidate rules are observationally equivalent, and quantify changes in model sensitivity when the dependency is broken.

Counterfactual and behavioural evaluation. Contrast sets [5], counterfactually augmented data [10] and CheckList [15] evaluate models on minimally edited inputs, usually written by hand and usually textual. Individual conditional expectation curves [7] trace a model’s response to one feature. Counterfactual explanations [16] search for the smallest change that flips a decision, which relates to our counterfactual flip distance. We bring these ideas to structured-state encoders. Two additions matter in this setting: the interventions keep documented field dependencies consistent, and the ablations deliberately break those dependencies to attribute the response.

Numeracy in language models. Token-based models represent numbers unevenly [17], and their numerical behaviour tracks surface frequency [14]. Our price results are consistent with surface-form sensitivity: price predictions are weakly associated with character edit distance. Separately, reformatting the days field from an integer to a float changes some decisions.

Distillation from a teacher. typed-decisions labels are a teacher LLM’s answer distribution, so fine-tuning is a form of distillation [9]. The learned boundaries can therefore be compared against the teacher’s own graded behaviour (§5.2).

3 Setting

Models. Laya encodes each (state, question) pair as [CLS] question [SEP] [MASK] $o_1 \dots$ [MASK] o_K [SEP] json(state) [SEP]. A two-layer head scores each [MASK] marker to produce one logit per option, and the logits are scaled by the checkpoint’s temperature settings [8]; this does not establish calibration on the counterfactual inputs. The state is serialised verbatim, so number formatting reaches the tokenizer. We compare the zero-shot **base** checkpoint [1] with the published **fine-tuned** checkpoint [2]. Both are pinned by revision. Inference runs in fp32, and our adapter reproduces the library’s own predict_batch to its four-decimal rounding.

Data. typed-decisions [11] has four synthetic workflows. Each has 300 training and 100 test states and five questions per state, and the gold labels are a teacher LLM’s answer distributions. We study invoice_processing, the workflow with the strongest numeric–decision associations. Its questions include urgency (score, four levels: *no time pressure* / *routine* / *elevated, same week* / *critical, same day*), matches_order (true/false: “line items, quantities and amounts reconcile”) and discrepancy_severity (score: *None* / *Trivial: rounding or cosmetic* / *Moderate* / *Material*). The other two questions, disposition and duplicate, are reported only in Table 1. Two numerical variables matter:

- **Days until due** x (integer field payment.days_until_due). Its dependent field payment.status is “overdue” iff $x < 0$ and “scheduled” otherwise. This holds on 400/400 invoice states. The field takes only 10 distinct values: $\{-21, -9, -3, 0, 2, 5, 11, 18, 27, 44\}$.
- **Unit-price variance** v , the invoice unit price relative to the purchase-order (PO) price, in %. Line and invoice totals are arithmetic functions of price and quantity (verified on all rows).

Sanity check. On the 2,000 original test decisions, accuracy (argmax agreement with the teacher) is 0.362 [0.337, 0.385] for the base model and 0.766 [0.746, 0.787] after fine-tuning. This reproduces the published figures. Table 1 gives the per-question numbers for the invoice workflow.

4 Method: Counterfactual Sweeps with Dependency Ablations

Interventions. Let s_0 be a held-out test state and x a numerical field with documented dependents $c = g(x)$. A *consistent* counterfactual sets $x \leftarrow x'$ and recomputes $c \leftarrow g(x')$. Everything else is unchanged, including key order and number formatting. This is a $\text{do}(x = x')$ intervention within the data-generating rules. A *dependency ablation* sets $x \leftarrow x'$ and freezes $c = c_0$. The resulting states can fall outside the generator’s support. This is intentional: the ablation is used to attribute the model’s response, not to predict how it should behave (Figure 1).

- **days_until_due:** integer grid $[-45, 75]$ at step 1 (121 values; 100 base states). The consistent sweep updates status; the ablation freezes it.
- **Price variance:** unit price set to $\text{PO} \times (1 + v/100)$ and rounded to cents, with line and invoice totals recomputed. We use a coarse 1% grid on $[-30, 30]$ and adaptive refinement to 0.1% and 0.01% around every argmax transition. The sweep uses the 68 single-line test invoices whose invoiced, ordered and received quantities agree, removing quantity mismatches. Other contextual fields remain unchanged. The realised variance after cent rounding is recorded to four decimal places and used in place of the nominal grid value.

Metrics. For a score question with levels $\ell \in \{0, \dots, 3\}$ we track $\mathbb{E}[\ell] = \sum_{\ell} \ell p_{\ell}$, normalised entropy and top-1/top-2 margin. Per base state we compute:

- the **counterfactual flip distance** $\text{CFD} = \min\{|x' - x_0| : \arg \max f(x') \neq \arg \max f(x_0)\}$, normalised by the observed training range;

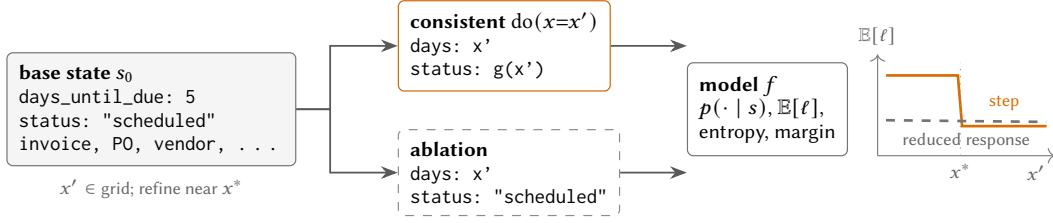


Figure 1: Protocol. Each held-out base state is swept over a grid of values of one number. In the *consistent* sweep, documented dependent fields are recomputed ($\text{status} = g(x')$). In the *ablation*, they are frozen. Loss of the step under ablation supports dependence on the cue, without excluding residual numerical sensitivity or interactions.

- every argmax transition, with the **flip point** x^* interpolated where the two exchanging classes’ probabilities cross, and the **dominant** (sharpest) boundary;
- the local slope of $p_{\text{to}} - p_{\text{from}}$ at x^* ;
- **monotonicity violation**, the share of $\mathbb{E}[\ell]$ movement against a semantically expected direction (urgency non-increasing in days_until_due), so lower values are more monotone;
- **boundary consistency**, the spread of x^* across contexts.

For the ablation, let R_i be the range of $\mathbb{E}[\ell]$ over the sweep for state i . We report the **relative range reduction**

$$\kappa = 1 - \frac{\frac{1}{n} \sum_i R_i^{\text{frozen}}}{\frac{1}{n} \sum_i R_i^{\text{consistent}}},$$

a descriptive reduction in mean response range when the cue is fixed. It is not a causal mediation fraction, can be negative, and is undefined if the denominator is zero.

Statistics. The base state is the sampling unit. We report means or medians as specified, IQRs, and 95% percentile bootstrap CIs (2,000 resamples) [4]. Base-versus-fine-tuned comparisons are paired over the same states: we give the mean difference and Cohen’s d_z , and Wilcoxon signed-rank p as a descriptive statistic.

Controls. We check four things (Table 3):

- determinism: repeats in differently composed batches;
- reversibility: x_1 to x_2 and back to x_1 ;
- irrelevance: an unrelated numeric feature ($\text{vendor_history.invoices_12m}$) swept against urgency;
- number formatting: 5, 5.0 and "5.00".

Each comparison uses outputs computed on the same device.

5 Results

5.1 Fine-tuning creates sharp, context-stable boundaries

Urgency vs. days_until_due. The base model has *no* numerical boundary. Across 100 states and 121 values, its urgency argmax never changes (0/100 states flip), and $\mathbb{E}[\ell]$ varies by only 0.09 levels over 120 days (Figure 2a, blue). The fine-tuned model flips in 100/100 states. Its dominant boundary sits at $x^* = -0.32$ (median; IQR 0.23), and 95/100 states place it within $[-3, 1.2]$. Overdue invoices are predicted *elevated* (84.5%) or *critical* (15.5%) across swept points. Invoices that are not overdue are predicted *routine* (95%). Away

Table 1: Invoice-workflow accuracy on the original test states (agreement with teacher argmax; 100 states per question). Accuracy rises sharply with fine-tuning. The right column summarises what the counterfactual sweeps reveal about the same decisions.

question	base	fine-tuned	sweep finding (fine-tuned)
urgency	0.32	0.52	step carried by "overdue"
matches_order	0.49	0.94	equality-sensitive
discrepancy_severity	0.44	0.84	small errors: "Material"
disposition	0.44	0.80	—
duplicate	0.11	0.92	—
invoice (all)	0.36	0.80	
all workflows	0.36	0.77	paired +0.40 [.37, .44]

from zero, both plateaus are approximately flat. The mean CFD is 11.9 [9.7, 14.2] days (median 9), and the normalised CFD is 0.18. The geometry changes along with accuracy. Paired against the base model, the fine-tuned model has +3.6 [2.8, 4.5] more transitions per sweep, and the share of movement against the expected direction falls from 0.50 to 0.30 ($d_z = -6.4$; Wilcoxon $p < 0.001$ for both).

Reconciliation vs. price. The base model produces nearly uniform probabilities (normalised entropy ≈ 1) and flips in only 14/68 states. The fine-tuned model flips in 68/68 states. The dominant boundary sits at $x^* = -0.002\%$ (IQR 0.008 percentage points), and the median CFD is 0.010 percentage points. Its dollar equivalent depends on the PO unit price. The mean slope at x^* is 1.5 per percentage point for the base model, over its 14 states with a boundary, and 127 for the fine-tuned model, over all 68 states. Paired over the 14 states with boundaries in both models, the difference is +128 [108, 151] ($d_z = 2.9$, Wilcoxon $p < 0.001$).

By the usual reading, fine-tuning has taught the model where urgency and reconciliation change, and to place those changes consistently across irrelevant context. The next two subsections test *what* the model is reading when it makes these decisions.

5.2 Urgency is dominated by the status cue

Freezing status largely removes the step at zero (Table 2, Figure 2). The mean range of $\mathbb{E}[\ell]$ across the full 120-day sweep falls from 0.71 to 0.11 levels, a relative range reduction of $\kappa = 0.85$. The step between $x = -1$ and $x = 0$ goes from 0.60 to -0.02 , a small

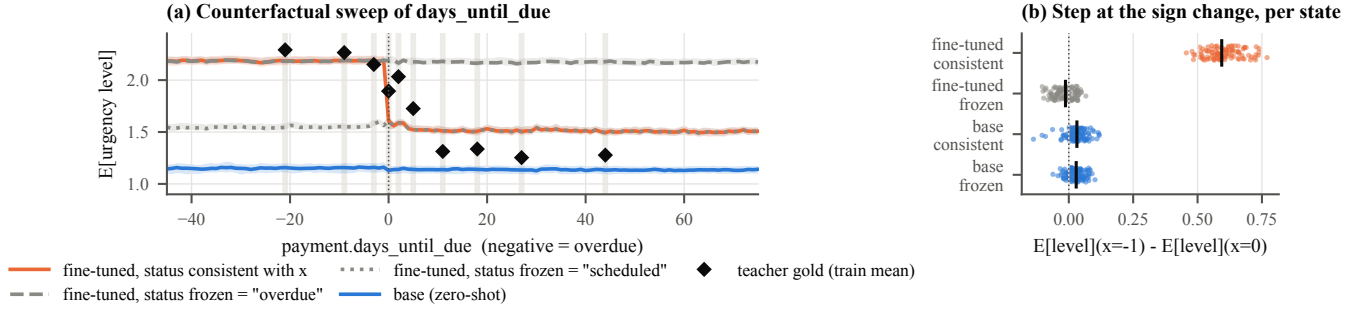


Figure 2: The main urgency step depends strongly on the status word. (a) Median $\mathbb{E}[\ell]$ over 100 held-out states, with the IQR shaded. With status kept consistent, the fine-tuned model (orange) steps at `days_until_due` = 0. With status frozen, its curve stays at the level of the frozen word across the entire range. Frozen "overdue" (dashed, 30 states) stays at ≈ 2.2 even 75 days before the due date. Frozen "scheduled" (dotted, 70 states) stays at ≈ 1.55 even 45 days overdue. The base model (blue) is flat. Diamonds show the teacher’s mean training label at each observed value, and vertical bands mark the 10 values present in the data. (b) Per-state step $\mathbb{E}[\ell](-1) - \mathbb{E}[\ell](0)$. Freezing the status word largely removes the main step.

Table 2: Dependency ablation for the fine-tuned model on urgency vs. `days_until_due` (100 states; bootstrap CIs). With status frozen, the mean response range falls substantially ($\kappa = 0.85$). Entries with CIs are means.

	status consistent	status frozen
range of $\mathbb{E}[\ell]$ over the sweep	0.713 [.701, .725]	0.109 [.101, .116]
step $\mathbb{E}[\ell](-1) - \mathbb{E}[\ell](0)$	0.603 [.591, .617]	-0.016 [-.024, -.009]
states with any argmax flip	100/100	32/100
slope at nearest x^* (per day)	0.093	0.007
base model: range / step	0.093 / 0.030	0.084 / 0.027

movement *against* the expected direction. Figure 2a makes the dissociation concrete:

- With the word frozen at "overdue", invoices that are not yet due, up to 75 days ahead, receive *elevated* (92%) or *critical* (8%) urgency, and never *routine*.
- With the word frozen at "scheduled", invoices that are up to 45 days overdue are *routine* in 93% of cases.

Where the frozen and consistent states coincide, outputs agree exactly ($\max |\Delta \mathbb{E}[\ell]| = 0$), supporting pipeline consistency. At fixed status the number has a small residual effect: a bump of ≈ 0.05 levels for $x \in [0, 3]$ under "scheduled", and 32/100 states with near-tie flips.

Relation to accuracy. The teacher’s labels are associated with `days_until_due` (training Spearman -0.78). Mean training urgency is 2.29, 2.15, 1.89, 2.03, 1.73 and 1.31 at $x = -21, -3, 0, 2, 5, 11$: an overall decline with local reversals. The modal training label changes from 2 to 1 at day 0, back to 2 at day 2, and to 1 at day 11. These are comparisons across different states, not counterfactual teacher responses. They nevertheless show supervision varying within a status category. Fine-tuned urgency accuracy is 0.38 on test states with `days_until_due` $\in [0, 5]$ against 0.58 elsewhere. Across the 22 original test states at -21 or -9 days, the model predicts *critical* once, whereas 13 teacher argmax labels are *critical*.

Overall urgency accuracy is 0.52, the lowest of the five invoice questions (Table 1). This is an observed score, not a proven ceiling or an attribution of every error to the shortcut. The control retrain in §5.6 shows a similar step with urgency accuracy 0.65, so the step alone does not determine the score.

5.3 Reconciliation is predominantly equality-sensitive

A tolerance-based rule would accept sufficiently small nonzero differences; no numerical tolerance is specified for the original reconciliation question. The fine-tuned model assigns a median $P(\text{true}) \approx 0.84$ across states at equality, with a pooled mean of about 0.15 over nonzero swept variances. At nonzero variance "true" is the argmax in only 0.3% of points. The existing sweeps include exact one-cent unit-price deviations from the PO price in 29 of the 68 states; all 29 have at least one such point whose prediction differs from the equality prediction. The other 39 states lack this exact probe, so the cached outputs do not establish a one-cent effect for all 68. A unit-price change is distinct from an invoice-total change, which also depends on quantity.

`discrepancy_severity` shows limited scaling with magnitude (Figure 3b). Among 1,176 evaluated points with $0 < |v| \leq 0.1\%$, the fine-tuned model answers "Material" 93.1% of the time and "Trivial" 0%. The rubric includes rounding or cosmetic differences, but does not specify a percentage cutoff. Small relative differences need not be only a few cents in absolute terms. Mean $P(\text{Material})$ is about 0.49 in this small-variance group and 0.55 at 5–30%, indicating weak rather than absent scaling. These percentages describe pooled, adaptively sampled sweep points, not a population frequency.

Isolated spikes of $P(\text{true})$ also appear away from zero. In one invoice, $P(\text{true}) = 0.77$ at approximately $+0.20\%$ (unit \$398.83 vs. \$398.03; total \$99,707.50 vs. \$99,507.50), against ≈ 0.14 at neighbouring prices. At that point each invoice figure differs from its corresponding PO figure by one character. We measure this as a character edit distance summed over the unit price and the total, which is 2 here. Across all 12,516 nonzero fine-tuned points, mean $P(\text{true})$ generally declines with this summed edit distance, from

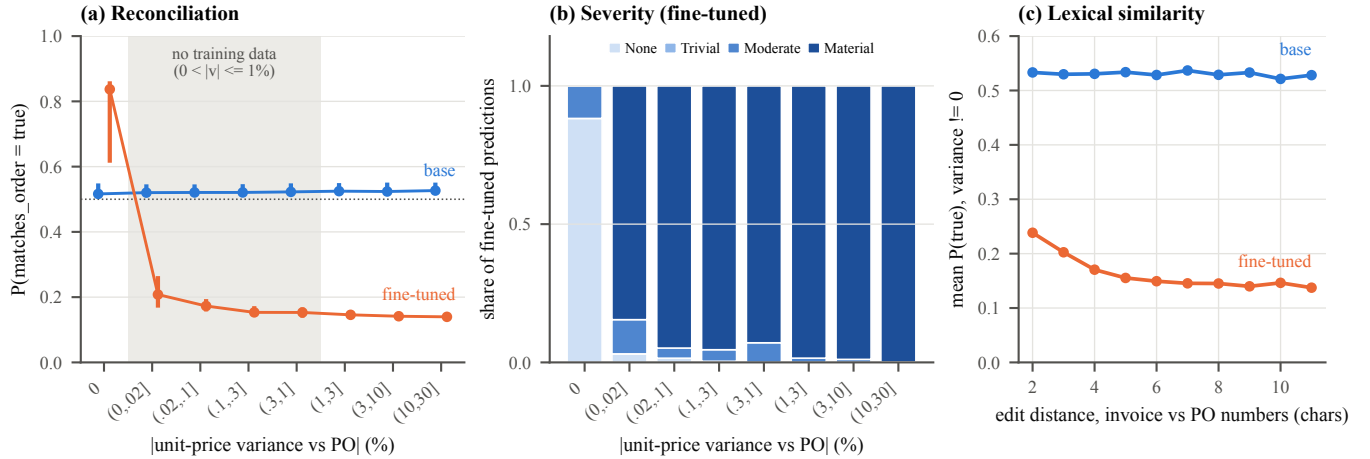


Figure 3: Price reconciliation is predominantly equality-sensitive, with limited severity scaling. (a) $P(\text{matches_order} = \text{true})$ by $|v|$ bin, as the median over 68 states with IQR bars. The fine-tuned model assigns 0.84 at exactly 0% and about 0.2 already at $0 < |v| \leq 0.02\%$. The shaded bins contain no training examples. (b) The fine-tuned discrepancy_severity prediction, pooled over evaluated points in each bin. Nonzero variance is mostly “Material”. “Trivial (rounding)” is almost never chosen. (c) At nonzero variance, the fine-tuned $P(\text{true})$ decreases with the character edit distance between the invoice and PO figures, summed over unit price and total. After controlling for $|v|$ the partial Spearman is $-0.12 [-0.16, -0.07]$.

0.24 at distance 2 to ≈ 0.14 at ≥ 8 (Figure 3c). Edit distance grows with $|v|$, so we control for $|v|$. The per-state partial Spearman given $|v|$ is $-0.12 [-0.16, -0.07]$ for the fine-tuned model and $+0.08$ for the base model. This is a weak, post-hoc association consistent with sensitivity to surface form, not proof of an exclusively string-based mechanism. The separate days-formatting control changes 6.3% of fine-tuned decisions when integers are represented as floats (Table 3); it does not directly test price formatting.

5.4 What the observed data can and cannot distinguish

For fixed remaining context z , suppose a numerical rule factors through a cue: there exists a function $\tilde{h}$ such that $h_x(x, z) = \tilde{h}(g(x), z)$ on the observed support. If $c = g(x)$ there, the cue rule $h_c(c, z) = \tilde{h}(c, z)$ makes identical predictions. This requires h_x to be constant across values of x sharing a cue at fixed z ; a deterministic cue alone is insufficient. No inverse of the generally many-to-one map g is required. Two specific ambiguities occur in the observed benchmark:

- **Status.** $\text{status} = \text{"overdue"} \Leftrightarrow \text{days_until_due} < 0$ on 400/400 invoice states (300 train, 100 test). “Urgent iff overdue” and “urgent iff $\text{days_until_due} < 0$ ” are indistinguishable on every example. Neither rule is equivalent to the teacher’s graded urgency labels.
- **Price.** Of 275 single-line training invoices, 244 have $v = 0$ exactly. The smallest nonzero $|v|$ is approximately 2.998074%. The test split has 83 of 97 at zero, with a smallest nonzero $|v|$ of approximately 2.999481%. Three training and one test variance fall just below 3% after cent rounding. Let $v_{\min}$ be the exact minimum nonzero $|v|$ across these splits. The predicates $v = 0$ and $|v| \leq \tau$ agree on these single-line examples for $0 \leq \tau < v_{\min}$, holding all other decision

Table 3: Controls (10 states each; repeats cover 4 values of days_until_due per state, giving 40 inputs, and formatting covers 8 values, giving 80 inputs per comparison). Repeated predictions are stable, the irrelevant field causes no argmax changes, and number formatting moves near-tie fine-tuned decisions.

control	result
repeats, rotated batches	$\max \Delta \text{logit} = 6 \times 10^{-8}$; 0/40 argmax changes
reversibility $x_1 \rightarrow x_2 \rightarrow x_1$	$ \Delta \text{logit} = 0$
irrelevant invoices_12m $\in [0, 60]$	0/10 flips; fine-tuned $\mathbb{E}[\ell]$ range 0.04 (vs. 0.71 for days_until_due)
5 vs. 5.0 vs. "5.00"	fine-tuned argmax changes 6.3% / 8.8%; base 0% / 0%

criteria fixed. This is a statement about the observed rows, not unseen generator outputs. The “Trivial” severity level has 3 training examples out of 275.

The urgency labels do provide graded supervision, and accuracy can expose errors against those labels. What on-distribution accuracy cannot identify is whether a sign-aligned response uses the number’s sign or its status cue. For price, the observed support leaves small-tolerance behaviour unconstrained by examples in that interval. Thus 0.94 reconciliation accuracy can coexist with rejection of very small price changes. Counterfactual probes reveal response patterns that aggregate accuracy does not explain.

5.5 Controls

Table 3 supports repeatability of the sweeps. Varying the irrelevant field causes no argmax changes, although probabilities move

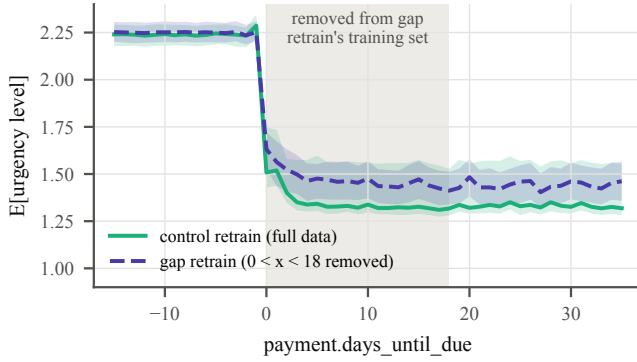


Figure 4: Training-support gap. Two retrains from the base checkpoint use the same recipe, with and without training cases where $0 < \text{days_until_due} < 18$. Both learn the step at 0. Across the withheld interval, the gap model largely extends the "scheduled" plateau. Median $\mathbb{E}[\ell]$ over 100 test states, IQR shaded.

slightly. Representing days as a float rather than an integer flips 6.3% of fine-tuned decisions (5 of 80) and no base decisions. The quoted-string variant also changes the JSON type, so it is not a pure formatting control. Further analyses of the saved outputs show:

- Non-dominant transitions are low-margin near-ties: the median margin is about 0.1 on the overdue plateau. These account for the mean of 3.6 transitions per sweep.
- Entropy is no higher at unobserved integer values of `days_until_due`, whether inside the observed range (0.794) or outside it (0.792), than at the 10 observed values (0.799).
- On a 2D grid over both variables, `matches_order` forms a single band at $v = 0$, independent of `days_until_due`, while urgency mostly follows the status step, with state-specific interactions with price.

5.6 The plateau extends across a withheld training interval

We examine whether withholding a numerical interval changes behaviour there. We retrained Laya twice with the reference fine-tuning recipe (4 epochs, effective batch 64, one seed each, from the base checkpoint). The *control* run used full data; it is a retraining arm, unrelated to the controls of §5.5. The *gap* run removed the 76 invoice training cases with `days_until_due` $\in \{2, 5, 11\}$. The edge values 0 and 18 and all negative values remain in training, so the step at zero keeps its training support; only the interior of the interval is withheld. Inside (0, 18) the word is always "scheduled", as it is at `days_until_due` ≥ 18 , so both status categories remain represented after removal. The results are in Table 4 and Figure 4:

- Both retrains exhibit a near-zero step resembling that of the published checkpoint. This replicates the geometry; status reliance was not separately ablated in the retrains.
- The gap model creates no new argmax boundary inside the withheld interval in 90% of states. Its argmax inside the interval matches its own prediction at the day-0 edge at 90% of points and at the day-18 edge at 99.6%.

Table 4: Training-support gap (one seed per arm; 100 test states).

	control	gap
invoice accuracy (5 questions)	0.81 [.77, .84]	0.77 [.74, .81]
urgency accuracy, all states	0.65	0.50
urgency accuracy, states inside gap	0.62 [.44, .77]	0.38 [.24, .53]
states with transition inside gap	2%	10%
entropy inside gap / at edges	0.758 / 0.755	0.758 / 0.748

- Mean normalised entropy inside the interval equals the control's (0.758). It is slightly above the gap model's own edge value (0.748, against 0.755 for the control).
- Its offset from the control is similar at several plateau locations: $P(\text{critical})$ is about 0.12 vs. 0.085 at days 5, 10 and 20.

The geometry is consistent with the status-word account, without uniquely identifying it: persistence of the plateau could also reflect numerical generalisation. Urgency accuracy falls from 0.65 to 0.50 (paired -0.15 [$-0.25, -0.06$]), and falls most on test states whose values were withheld (0.62 to 0.38). These runs associate support removal with lower accuracy and persistence of the main step; they are not proof of an unchanged internal rule. The models also differ outside the gap ($|\Delta \mathbb{E}[\ell]| = 0.111$ vs. 0.154 inside). This outside-gap difference cannot estimate retraining variance: removing data can affect both regions. With one seed per arm, the accuracy difference is provisional; state-level bootstrap CIs do not capture variability across training runs.

5.7 Written numerical rules are followed incompletely

Laya's choice questions accept free-text option criteria, so a numerical rule can be *written* into the input without retraining. We posed a new question, "Which payment window does this invoice fall into, based on `days_until_due`?", in two versions. The *complete* version has the options soon (< 5), mid (5–14) and later (≥ 15). The *gap* version has only soon (< 5) and later (≥ 15); here "gap" means the unspecified range 5–14, not the withheld training interval of §5.6. Both were evaluated over $x \in [-10, 30]$ on 20 states with status consistent (Figure 5).

- The base model has near-maximal mean normalised entropy (about 0.995–0.997), with low rule agreement. Its complete-condition predictions include all three options; they are not globally constant.
- With complete criteria, the fine-tuned model agrees with the written rule on only 0.48 [0.43, 0.53] of the 820 points. The most frequent per-state argmax switches at 0, aligned with status, rather than at 5. However, "later" is selected at 174 points (21.2%). Mean probability changes near 5 and 15 are consistent with threshold sensitivity. Since status was not frozen in this experiment, attributing the misplaced boundary specifically to status remains an inference.
- With the gap criteria, agreement is 0.83 [0.76, 0.90] on the 620 specified points, and the soon $\rightarrow$ later switch lands at a mean per-state crossing point of 16.8 (the within-state

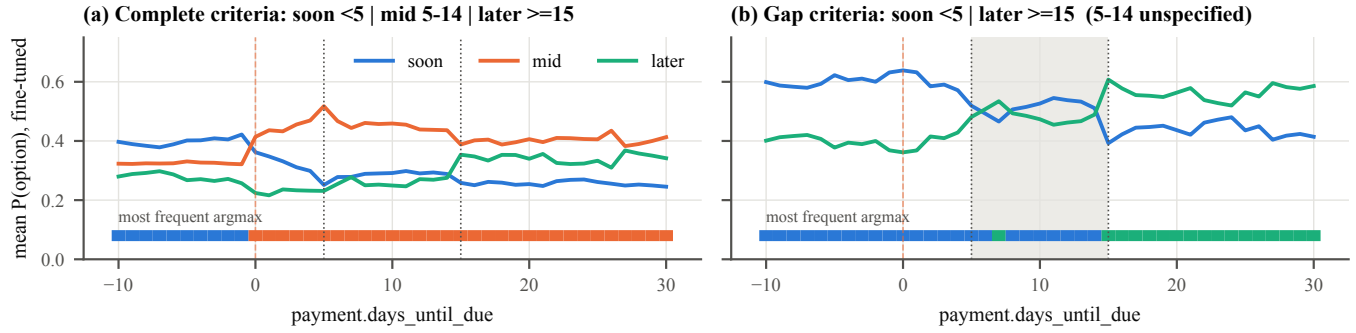


Figure 5: Specification gap: a new question whose options state numeric thresholds on `days_until_due` (fine-tuned model; mean over 20 states). The bar at the bottom shows the most frequent per-state argmax. (a) With complete criteria, the most frequent per-state argmax switches from “soon” to “mid” at 0, where status changes (dashed), rather than at 5. “later” is never the most frequent answer at a grid value, but is chosen at 174/820 individual points. Mean probabilities also change near 5 and 15. (b) With 5–14 unspecified, agreement is 0.83 on specified points only; the mean per-state soon-to-later crossing point is 16.8.

median where a state has multiple crossings). Inside the unspecified interval the model leans towards the earlier option (65% “soon”) with only slightly raised entropy (0.974 vs. 0.954).

The two agreement scores use different option sets and evaluation domains, so they are not a controlled measure of improvement from omitting criteria. Fine-tuned mean normalised entropy is about 0.96 in both conditions, but individual values fall to 0.83 and 0.75 respectively. High entropy describes diffuse output probabilities; it does not by itself establish calibrated uncertainty about this unseen question.

6 Discussion

What the findings have in common. The fine-tuned model’s responses are dominated by coarse distinctions: overdue versus scheduled for urgency, and equal versus unequal prices for reconciliation. The urgency ablation directly supports dependence on the status cue. The price evidence supports equality sensitivity and limited magnitude scaling, with string similarity as a tentative contributing factor. Neither result establishes that the model lacks numerical representations. Derived categorical fields are a useful target for similar audits; the prevalence of these behaviours in other models and datasets remains unknown.

Implications for evaluation. Aggregate held-out accuracy measures agreement but does not identify the input features responsible for it. For candidate rules that agree on the observed support, off-support probes can reveal differences. Our protocol combines three complementary measurements:

- Consistent sweeps measure the joint response to a number and its documented dependent fields.
- Dependency ablations measure numerical sensitivity at fixed cue values, including off-support states.
- Boundary metrics complement accuracy: flip point, slope, boundary consistency and relative range reduction.

Sweep evaluations can be batched and cached by input hash.

Implications for training. Both retrains reproduced the main step, withholding the interval in the gap retrain did not create a new boundary in most states, and written thresholds were followed incompletely in the new question. These exploratory results do not establish a general failure of either approach or identify the same mechanism in each model and task. Candidate remedies include counterfactual augmentation varying `days_until_due` at fixed status with explicit graded labels [10], and small price perturbations labelled under a specified tolerance. Their suitability and effectiveness remain to be evaluated.

Limitations. Our evidence comes from one model family, one fine-tuned checkpoint plus two retrains, and one synthetic workflow with two numerical variables. This is a case study, not evidence of how frequently encoders learn such shortcuts. Gold labels are a single teacher LLM’s distributions, so “accuracy” is agreement with that teacher. Expected directions are semantic readings of the question text, not documented rules. The status dependency ablation creates states the generator never emits: it measures sensitivity to controlled input edits, not how the model ought to behave on contradictory states. The edit-distance analysis is post hoc; controlling for magnitude does not eliminate all possible confounding. Teacher trends across original states do not show how the teacher would answer our counterfactuals. Adaptive price grids also give unequal weight to different regions and states in pooled point summaries. The retraining experiment uses one seed per arm, and the published checkpoint’s configuration differs from the documented recipe, which is why we compare two local retrains. Controls, the 2D surface and the specification experiment ran on CPU fp32, and the remainder on Apple MPS fp32. Each comparison uses outputs computed on the same device. Integer-day crossings are interpolated between sampled values; cent rounding and grid resolution limit price-boundary precision.

7 Conclusion

Fine-tuning Laya raises agreement with teacher labels from 0.36 to 0.77 while producing a status-dominated urgency boundary and predominantly equality-sensitive price decisions. Freezing status

reduces the mean urgency response range by about 85%, while small nonzero price differences are usually labelled material. These findings reveal behaviours that aggregate accuracy alone does not explain. They do not establish complete absence of numerical reasoning, an exclusively string-based mechanism, or equivalence between the status shortcut and graded teacher labels. Consistent sweeps paired with dependency ablations provide a practical complement to accuracy when auditing decisions expected to vary with numerical magnitude.

Acknowledgments

We thank the authors of Laya and of the typed-decisions dataset for releasing models, code and data with pinned revisions, which made this audit possible.

Reproducibility

The analyses reuse cached model outputs and the pinned dataset. The main sweep file contains 147,136 evaluation records across models and questions; this is not the number of unique states or batched forward passes. The reported numbers were checked against the saved predictions and the dataset; these checks used no additional model inference or retraining. Model inputs are cached by exact serialised input. Run metadata records package versions, device, model and dataset revisions, and seed 20260928.

References

- [1] ConvAI Innovations. 2026. Laya: a ModernBERT-based typed decision model. Hugging Face model `convaiinnovations/Laya`, revision 55cf4c4; code release 0.3.21 (commit 9d95567). Accessed September 2026.
- [2] ConvAI Innovations. 2026. Laya fine-tuned on typed decisions. Hugging Face model `convaiinnovations/Laya-typed-decisions`, revision 1a793eb. Accessed September 2026.
- [3] Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. 2022. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research* 23, 226 (2022), 1–61.
- [4] Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC.
- [5] Matt Gardner, Yoav Artzi, Victoria Basmov, et al. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1307–1323. doi:10.18653/v1/2020.findings-emnlp.117
- [6] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673. doi:10.1038/s42256-020-00257-z
- [7] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics* 24, 1 (2015), 44–65. doi:10.1080/10618600.2014.907095
- [8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 1321–1330.
- [9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).
- [10] Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. In *International Conference on Learning Representations (ICLR)*.
- [11] LocalLLaMA. 2026. typed-decisions: synthetic agent-workflow decisions with teacher label distributions. Hugging Face dataset `LocalLLaMA/typed-decisions`, revision f7a2487. Accessed September 2026.
- [12] Nandakishor M. 2025. SalesRLAgent: A Reinforcement Learning Approach for Real-Time Sales Conversion Prediction and Optimization. *arXiv preprint arXiv:2503.23303* (2025).
- [13] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2 ed.). Cambridge University Press.
- [14] Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 840–854. doi:10.18653/v1/2022.findings-emnlp.59
- [15] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 4902–4912. doi:10.18653/v1/2020.acl-main.442
- [16] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 841–887.
- [17] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5307–5315. doi:10.18653/v1/D19-1534
- [18] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, et al. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint arXiv:2412.13663* (2024).