

# Decodable, Steerable, but Used? Auditing a Latent Risk Direction in Financial Language Models

Anonymous Author(s)

## Abstract

Mechanistic interpretability is increasingly proposed as a way to inspect and steer what a financial language model “believes” about risk. A convincing risk-belief direction bundles four claims: the direction represents the risk regime, intervening on it changes the output, the effect is specific to the intended concept, and the model uses the direction when it responds to riskier inputs. Standard pipelines can produce strong numbers for all four while establishing at most one. We audit this bundle for a difference-of-means volatility direction in models reading a market context and an earnings-call disclosure, from GPT-2 and Qwen2.5-1.5B to Qwen3-32B, a finance-tuned model, and OLMo-3 after each post-training stage. Seven checks map to the four claims: matched and surface nulls, leave-one-firm-out estimation, multiplicity correction, output-based readouts, random and competing-concept controls, an input-leakage control that replaces the stated volatility with raw returns, and causal mediation and erasure at every block. The naive pipeline’s headline results, a steering effect of Cohen’s  $d \approx 118$  and a perfect regime probe, re-read the steered axis and the stated number. Representation survives: computed from raw returns, synthetic or real, the regime separates beyond a matched null in every model, as an order-invariant aggregate of return magnitudes. Intervention survives only with qualifications: steering moves an explicit High/Low risk judgement, but at the selected block only by editing the positions where the answer is read, at 20–30 times the natural shift in the small models, and a direction fit on non-financial “intensity” sentences steers as well. Use is not supported. Across 16 runs with a natural effect, the direction carries at most about 30% of the model’s response to riskier inputs at any block (Qwen2.5-1.5B: 0.004 of 0.043), and erasing the regime leaves at least 64% of the judgement in place, even where the reasoning steps as a whole carry the effect. A representation can be decodable and steerable without being what the model uses; we translate this into concrete requirements for validating representation-level explanations of financial models.

## CCS Concepts

• **Computing methodologies** → **Natural language processing; Neural networks; Learning latent representations; Causal reasoning and diagnostics; Model verification and validation**; • **Mathematics of computing** → **Hypothesis testing and confidence interval computation**; • **Applied computing** → **Economics**; • **Social and professional topics** → **Governmental regulations**.

## Keywords

mechanistic interpretability; activation steering; steering vectors; linear representation hypothesis; causal mediation analysis; activation patching; interpretability illusions; financial large language models; faithful evaluation of explanations; model risk management

## 1 Introduction

Language models now read filings, summarise earnings calls and draft risk commentary in financial institutions [50, 53]. Classical financial XAI explains such systems from the outside, attributing an output to input features [8, 30]. Mechanistic interpretability offers a stronger kind of explanation: locate an internal variable, say a direction in the residual stream that tracks expected volatility, and show by intervention that it drives the output [42, 47, 55]. To a model validator this looks like evidence of conceptual soundness in the sense of SR 11-7 [6], and like the kind of interpretability of outputs that the EU AI Act asks of high-risk systems [14]. Neither text prescribes such techniques; the question is what such evidence would actually establish if it were offered.

We argue that a convincing “risk-belief direction” bundles four claims that need separate evidence: **(R)** the direction *represents* the risk regime, beyond the surface edits that come with it; **(I)** *intervening* along it changes the model’s output; **(S)** the effect is *specific* to the intended concept, here risk at a given horizon; **(U)** the model *uses* the direction when it responds to riskier inputs. A standard pipeline (contrast pairs, difference of means, steering) can produce striking numbers for all four while establishing at most one. A first version of our own pipeline reported a steering effect of Cohen’s  $d \approx 118$  and a regime probe at 100% accuracy. The first number re-read the steered axis. The second read a number stated in the prompt.

*Question.* We test one falsifiable hypothesis: when a language model reads a market context and an earnings-call disclosure, the difference-of-means direction between high- and low-volatility contexts satisfies R, I, S and U. We test it on GPT-2 and Qwen2.5-1.5B-Instruct, with the volatility either stated as a number or left to be inferred from raw returns, and check the conclusions on larger general (Qwen3-8B, -14B and -32B, Llama-3.1-8B-Instruct), finance-tuned (Fin-R1) and open-pipeline models (OLMo-3-7B after SFT, DPO and full post-training), on further seeds, and on real return series.

*Contributions.* (1) An audit that maps each check to the claim it can support (Table 1), including two checks that matter in finance: a *matched null* that changes as many numbers as the treatment without changing the regime, and an input-leakage control. The pipeline is validated on a model with a planted direction. (2) Evidence that the naive results are artifacts of where and how they are read: re-projecting on the steered axis, comparing against random directions at the final block, and steering at strengths up to 30 times the natural shift. (3) A dissociation that holds across models. With raw returns the regime is represented beyond the matched null at every block, as an order-invariant aggregate of return magnitudes (R). Steering along the direction moves the output (I), but at the selected block only by editing the readout position, and at strengths that are meaningful only in units of the natural shift.

Mediation at every block, in models from 124M to 32B parameters, shows that the direction carries at most about 30% of the model’s natural response to riskier inputs, usually much less, even where the reasoning-step positions as a whole carry all of it, and that erasing the regime there leaves most of the judgement in place (U). Horizon specificity (S) cannot be tested because the models’ judgements do not depend on the horizon even when the ground truth does. (4) Guidance for validators on which claims a representation-level explanation can support (Section 6).

## 2 Related Work

*What decodability does not show.* Probes report what is linearly decodable, not what the model uses; control tasks and selectivity were introduced for exactly this gap [3, 21], and high probe accuracy need not mean the property is relevant to the task [41]. Linear structure has been reported for sentiment [46], truth [33], and space and time [18]; each such finding is a claim of type R. Single directions can look interpretable on one dataset and not another [7]. Amnesic probing and concept erasure ask instead whether removing the information changes behaviour [5, 13, 40]. Our leakage and matched-null checks are the financial analogue of control tasks: market prompts usually state the target quantity, so decodability is expected and uninformative.

*What steering does not show.* Adding a difference-of-means vector steers behaviour in many settings [2, 27, 42, 47, 55], often motivated by linear-representation arguments [33, 36]. Steering effects can be unreliable across inputs [43], and subspace interventions can succeed through pathways the model does not use naturally [31]. Representation finetuning trains such interventions directly [51], which makes them more effective, not more faithful. An intervention effect is therefore evidence for I, not for U.

*Testing natural use.* Activation patching and causal mediation ask whether a variable carries the effect of a natural input change [15, 20, 34, 48], and their conclusions depend on metric, patch design and layer [54]; where an effect localises need not be where an edit works [19]. Distributed alignment search learns the subspace that carries a causal variable [16]; we use the cheaper bracket of a rank-1 patch, a rank- $k$  patch and the full residual at the same positions. Circuit analyses trace how an effect is computed from input to output [11, 49], and sparse autoencoders [9, 12, 45] and feature circuits [32] offer finer-grained candidates; we leave these for future work. Whether an explanation reflects how the model actually reaches its output is the faithfulness question [23, 24]; U is a causal test of faithfulness for a representation-level explanation.

*Financial XAL.* Attribution [8, 30] and concept tests [26] dominate deployed explanations; financial language models [1, 50, 52, 53] are rarely examined internally, even as general models are asked to read returns and statements directly [25, 29]. What is missing is not another check but an account of how the known failure modes combine on a financial claim, and which of R, I, S, U survives.

## 3 Setting and Estimands

*Data.* 16 synthetic firms in six sectors, two reporting periods each, one per volatility regime (32 records). Daily log-returns follow a GARCH-like process with daily  $\sigma$  of 0.8% (low regime) or

2.6% (high regime), with stronger clustering in the high regime. The regime is constant over the 60-day look-back and the 60-day horizon, so *near- and medium-term risk are identical by construction*; we return to this in Section 5. Disclosures combine boilerplate with guidance sentences graded by regime and three uncertainty levels. The market context is rendered either as *stated* (cumulative return, annualised volatility as a number, trend) or *raw* (the last 20 daily returns, no summary statistic).

*Prompts and activations.* A prompt contains a task (downside risk, transient shock vs. regime shift, or scenario comparison), a horizon (near or medium term), the context, the disclosure, and eight reasoning-step markers. Let  $\tilde{h}_L(x) \in \mathbb{R}^d$  be the residual stream at the output of block  $L$  (index 0 is the embedding output), averaged over the marker tokens.

*Direction.* A history contrast pairs the same firm’s low- and high-regime contexts, with the low-period disclosure held fixed, for two task types per firm ( $i = 1, \dots, 32$ ). With per-pair differences

$$\delta_i^L = \tilde{h}_L(x_i^{\text{hi}}) - \tilde{h}_L(x_i^{\text{lo}}), \quad \bar{\delta}^L = \frac{1}{n} \sum_i \delta_i^L, \quad (1)$$

the belief direction is the normalised difference of means at the block where it is largest,

$$b = \frac{\bar{\delta}^{L^*}}{\|\bar{\delta}^{L^*}\|}, \quad L^* = \arg \max_L \|\bar{\delta}^L\|, \quad (2)$$

which gives block 12 of 12 in GPT-2 and 27 of 28 in Qwen.

*Readouts.* Both readouts are computed from the next-token distribution at the final position. The lexical readout contrasts risk words  $R = \{\text{risk, loss, decline, downside, volatile, weak}\}$  with calm words  $C = \{\text{stable, steady, strong, gain, growth, safe}\}$ :

$$y_{\text{lex}}(x) = \frac{1}{|R|} \sum_{w \in R} \log p(w | x) - \frac{1}{|C|} \sum_{w \in C} \log p(w | x). \quad (3)$$

It measures a lean in the next token, not a risk estimate. The decision readout appends “Final answer. Overall downside risk (High or Low):” to the prompt  $x$ , giving  $x^q$ , and takes

$$y_{\text{dec}}(x) = \log p(\text{High} | x^q) - \log p(\text{Low} | x^q), \quad (4)$$

which is closer to a decision but is still a single-token judgement.

*Intervention.* Steering edits the residual stream at block  $L$  during the forward pass, so all later blocks recompute:

$$h_{L,t} \leftarrow h_{L,t} + \alpha b, \quad \alpha \in [-8, 8] \text{ (raw residual units)}. \quad (5)$$

The edit is applied to every token  $t$  from the first reasoning-step marker to the end of the prompt. This span includes the final (readout) position and, for  $y_{\text{dec}}$ , the appended question; Section 5.5 also reports edits that stop before the question.  $h_L$  is written by a forward hook on block  $L$ , in the space it is read in: for the last block, whose state Hugging Face returns after the final norm, the hook sits on the final norm’s output. The steering effect is  $\Delta y(\alpha) = y(x; \alpha b) - y(x)$ , compared against norm-matched random and orthogonal directions. To make  $\alpha$  interpretable we also express it in units of the *natural gap*  $g_L = \frac{1}{n} \sum_i \langle \delta_i^L, b_L \rangle$ , the distance the regime itself moves the residual along  $b_L$ .

*Mediation.* Let  $P_b = bb^\top$  project onto  $b$ , and let  $y(x; P_b h \leftarrow v)$  denote the readout when the  $b$ -component of the block- $L$  residual is set to that of another run  $v$  while the orthogonal part keeps its own value. Patches act on the same token span as steering: each reasoning step's mean residual is moved to its target by adding one shift to every token of that step, so patching a run with its own values is the identity. Following [48],

$$TE = \mathbb{E}[y(x^{\text{hi}}) - y(x^{\text{lo}})], \quad (6)$$

$$NIE = \mathbb{E}[y(x^{\text{lo}}; P_b h \leftarrow P_b h(x^{\text{hi}})) - y(x^{\text{lo}})], \quad (7)$$

$$NDE = \mathbb{E}[y(x^{\text{hi}}; P_b h \leftarrow P_b h(x^{\text{lo}})) - y(x^{\text{lo}})]. \quad (8)$$

NIE is the part of the natural regime effect carried by the coordinate along  $b$ ; NDE is what remains when that coordinate is held at its low-regime value. Replacing  $P_b$  by the identity gives  $NIE_{\text{full}}$ , what the whole residual at these positions and block carries, which bounds what any direction there can carry; a rank- $k$  projection onto the top principal directions of the per-pair contrasts sits in between. We report all of these in absolute terms, because the ratio  $NIE/TE$  is unstable when  $TE$  is near zero.

*Models and inference.* The main audit uses GPT-2 (124M) [39] and the instruction-tuned Qwen2.5-1.5B-Instruct [37] in float32, with two further data seeds for Qwen. Section 5.6 repeats the representation, steering and mediation checks on larger models, in float32 up to 14B and with a float32 unembedding at 32B, since bfloat16 logits are too coarse for effects of order 0.01. The statistical unit is the firm: all  $p$ -values are firm-level cluster permutation tests, and all intervals are firm-level cluster bootstraps [10]. With 16 firms, the effective sample size of the contrast tests is about 16 (design-effect corrected), so we report raw effects in log-probability units alongside any standardised effect. Code, configurations and every result file will be released with the camera-ready version.

## 4 The Audit

The checks are ordered by the claim they can support. Each one closes a specific way of manufacturing evidence.

*Representation (R).* **A1 Nulls.** A *surface* null changes only form (whitespace, a paraphrased disclosure, reformatted headers). It is weak, because any edit that changes more tokens beats it. A *matched* null swaps in the context of another firm *in the same regime* while keeping the firm label and disclosure. It changes as many numbers as the treatment but not the regime. **A2 Cross-fitting.** Fitting  $b$  and testing it on the same pairs aligns it with them by construction. We hold out each firm  $f$ , fit the direction on the other firms at every block, and score only the held-out firm's pairs:

$$s_i^L = |\langle \delta_i^L, b_L^{(-f(i))} \rangle|, \quad b_L^{(-f)} = \frac{\sum_{j: f(j) \neq f} \delta_j^L}{\|\sum_{j: f(j) \neq f} \delta_j^L\|}. \quad (9)$$

The same score is computed for held-out null pairs, and the two are compared by Cohen's  $d$  and a firm-level permutation test. **A3 Multiplicity.** Per-block tests are Holm-corrected over blocks [22].  $L^*$  itself is chosen on all firms, so per-block results at  $L^*$  are descriptive rather than confirmatory. **A6 Leakage.** A regime probe peaking at block 1 means the regime is read off the input. We therefore rerun with raw returns, and test whether the representation aggregates the window, reacts to single large returns, or depends on order (C1).

*Intervention (I).* **A4 Independent readout.** Re-projecting steered activations on  $b$  measures the intervention itself. We score steering only on  $y_{\text{lex}}$  and  $y_{\text{dec}}$ , computed from the logits, at  $\alpha=8$  and at the natural gap  $g$ , and with edits that stop before the appended question, so that an effect cannot come from writing at the readout position. **A5a Controls.**  $b$  is compared with 8 norm-matched random directions and 8 random directions with  $b$  projected out (orthogonal), on 16 prompts. A signed average over random directions cancels, so on  $y_{\text{dec}}$  we compare  $b$  with their mean absolute effect. Because  $b$ 's effect is nearly constant across prompts, standardised effects are huge (Cohen's  $d$  of 48 on Qwen) and say little; we report raw effects.

*Specificity (S).* **A5b** We check that the model's behaviour depends on the horizon where the ground truth does (C2);  $b$  is compared with the other horizon's direction, with the leading principal components of the per-pair contrasts, with competing directions fit on non-financial text (intensity, sentiment), with the same windows in basis points, and through a logit lens [4, 35]: its cosine with the readout's own unembedding axis  $W_U(\bar{r} - \bar{c})$  after the final-norm gain, and the tokens it promotes.

*Natural use (U).* **A7 Mediation and erasure.** We report  $TE$ ,  $NIE$  and  $NDE$  in absolute terms with intervals, on both readouts (C3), at  $L^*$  and at every block with  $b_L$  refit there, bracketed by the full-residual and rank-8 patches. We also erase the regime at each block with LEACE [5], fit on the reasoning-step states of both regimes, and report the  $TE$  that remains, an amnesic counterfactual [13]. The proportion mediated  $NIE/TE$  is unstable when  $TE$  is near zero, so we do not lead with it.

*Positive control.* On a synthetic transformer with a direction planted at layer 13, the pipeline recovers the planted structure: held-out separation peaks at layer 13 with sign consistency 0.99, steering along  $b$  moves the readout while random directions do not, the difference-of-means direction is the most effective candidate (alignment 1.0), and 98% of the effect is mediated (95% CI 0.96–1.01). This shows the code can find a planted, used direction. It does not show the checks are exhaustive.

## 5 Results

### 5.1 The naive evidence is an artifact of the evaluation

On GPT-2 (stated), steering moves the projection on  $b$  by exactly  $\alpha$  (8.00 at  $\alpha=8$ ); this re-reading of the steered axis is what produced  $d \approx 118$ .<sup>1</sup> Readouts computed from the logits tell a different story. At  $L^*$ , GPT-2's last block, an edit reaches the output only through the final norm and the unembedding, so norm-matched random directions, which write directly into the logits, move  $y_{\text{lex}}$  by +0.236 against  $-0.007$  for  $b$ . This comparison is uninformative by construction. Away from the last block it is not:  $b$  moves  $y_{\text{lex}}$  two to four times more than random directions at blocks 3, 9 and 11, but  $y_{\text{dec}}$  only at block 3. The localisation evidence also dissolves. Against the surface null, 11 of 12 blocks separate in-sample after Holm correction.

<sup>1</sup>The submitted version wrote layer 12 before GPT-2's final LayerNorm while reading it after, which shrank the shift to 4.55 and made the GPT-2 mediation non-additive. Writing in the space that is read removes both (Section 3); the old placement reproduces the old numbers exactly.

**Table 1: Evidence ledger, grouped by the claim each check bears on: R = representation, I = intervention, S = specificity, U = natural use; *Test* names the check in Section 4 that produced the row. *stated*: volatility given as a number; *raw*: 20 daily returns only. Per-block entries are at the steering block  $L^*$  chosen by the pipeline (second row). *d*: Cohen’s  $d$  with held-out (leave-one-firm-out) directions. *p*: firm-level cluster permutation test. [CI]: 95% firm-bootstrap interval.  $y_{\text{lex}}$ : risk–calm next-token log-probability;  $y_{\text{dec}}$ :  $\log p(\text{High}) - \log p(\text{Low})$  after an explicit risk question; effects in log-probability units. Random: 8 norm-matched directions, a signed mean on  $y_{\text{lex}}$  and a mean absolute effect on  $y_{\text{dec}}$ . *g*: natural shift of the residual along  $b$  between regimes. Candidate effects are mean  $|\Delta y_{\text{lex}}|$  over eight prompts at  $\alpha=8$ . GPT-2’s  $L^*$  is its last block, where an edit acts only through the unembedding. “–”: not measured in that run.**

| Test     | Quantity                                                     | GPT-2                    |                          | Qwen2.5-1.5B             |                          |
|----------|--------------------------------------------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
|          |                                                              | stated                   | raw                      | stated                   | raw                      |
| R A6     | Regime probe acc. (best block)                               | 1.00 (L1)                | 1.00 (L1)                | 1.00 (L1)                | 1.00 (L1)                |
| R A1–2   | Held-out $d$ vs. surface null ( $L^*$ )                      | 0.37 (L12)               | 2.55 (L12)               | 0.22 (L27)               | 0.74 (L27)               |
| R A1–2   | Held-out $d$ vs. matched null ( $L^*$ )                      | 0.04, $p=0.45$           | 0.46, $p=0.065$          | -0.12, $p=0.67$          | 1.12, $p<.001$           |
| R A3     | blocks sig. vs. matched null (Holm)                          | 2 / 12                   | 11 / 12                  | 13 / 28                  | 28 / 28                  |
| I A4     | Shift along $b$ at $\alpha=8$ (circular)                     | 8.00                     | 8.00                     | 8.00                     | 8.00                     |
| I A4–5a  | $\Delta y_{\text{lex}}$ , $\alpha=8$ : $b$ / random          | -0.007 / +0.236          | -0.017 / +0.236          | +0.244 / -0.019          | +0.145 / -0.019          |
| I A5a    | $b >$ random; $b >$ orthogonal ( $p$ )                       | 1.00; 1.00               | 1.00; 1.00               | <.001; <.001             | <.001; <.001             |
| I A4–5a  | $\Delta y_{\text{dec}}$ , $\alpha=8$ : $b$ / mean  random    | +0.027 / 0.772           | +0.018 / 0.772           | +0.250 / 0.048           | +0.098 / 0.050           |
| I A7     | Natural gap $g$ along $b$ ; $8/g$                            | 1.01; 8                  | 7.23; 1                  | 0.26; 31                 | 0.38; 21                 |
| I A4–5a  | $\Delta y_{\text{dec}}$ , $\alpha=g$ : $b$ / mean  random    | +0.003 / 0.097           | +0.017 / 0.697           | +0.008 / 0.002           | +0.005 / 0.002           |
| I A4     | edit stops before question: $b$ /  random                    | –                        | +0.000 / 0.000           | +0.004 / 0.002           | -0.001 / 0.002           |
| S A5b    | $b >$ other-horizon direction ( $p$ )                        | <.001                    | 1.00                     | 1.00                     | <.001                    |
| S A5b    | cos(near-, medium-term direction)                            | 0.998                    | 1.000                    | 0.986                    | 0.990                    |
| S C2     | Horizon interaction on $y_{\text{dec}}$ [CI]                 | –                        | +0.003 [-0.004, +0.010]  | –                        | -0.004 [-0.009, +0.001]  |
| S A5b    | cos (b, readout unembedding)                                 | 0.001                    | 0.001                    | 0.076                    | 0.052                    |
| S A5b    | Best candidate / $b$ , mean $ \Delta y_{\text{lex}} $ ;  cos | pc4: 0.24 / 0.01; 0.02   | pc3: 0.26 / 0.02; 0.01   | pc1: 0.63 / 0.23; 0.01   | pc2: 0.28 / 0.18; 0.32   |
| U A7     | $y_{\text{lex}}$ : TE / NIE / NDE                            | +0.001 / -0.001 / +0.001 | -0.007 / -0.007 / +0.000 | +0.011 / +0.005 / +0.006 | +0.010 / +0.006 / +0.004 |
| U A7, C3 | $y_{\text{dec}}$ : TE [CI]                                   | -0.054 [-0.122, +0.016]  | -0.037 [-0.084, +0.006]  | +0.105 [+0.025, +0.190]  | +0.043 [+0.023, +0.063]  |
| U A7, C3 | $y_{\text{dec}}$ : NIE through $b$ [CI]                      | +0.001 [-0.001, +0.003]  | +0.003 [+0.000, +0.006]  | +0.010 [+0.002, +0.018]  | +0.004 [+0.002, +0.006]  |
| U A7     | NIE, full residual at $L^*$ [CI]                             | +0.005 [-0.003, +0.012]  | +0.006 [-0.001, +0.013]  | +0.028 [+0.005, +0.051]  | +0.014 [+0.009, +0.019]  |
| U A7     | largest NIE through $b_L$ , any block                        | +0.001 (L11)             | +0.007 (L7)              | +0.015 (L28)             | +0.006 (L28)             |
| U A7     | TE after LEACE erasure at $L^*$                              | -0.055                   | -0.044                   | +0.093                   | +0.037                   |

Against the matched null with held-out directions, only blocks 1–2 do, where the digits of the stated number differ (Figure 1); at  $L^*$  the held-out effect is  $d=0.04$  ( $p=0.45$ ). The same correction removes Qwen’s localisation claim. In-sample against the surface null, six deep blocks (18, 21, 22, 25, 27, 28) survive Holm correction. Held out against the matched null, the stated-format signal sits in blocks 3–16 (13 of 28 significant) and is absent at the steering block ( $d=-0.12$  at block 27). The regime probe is perfect at block 1 for every model and format. In the stated format it reads the number, and in the raw format it can read the size of individual returns.

## 5.2 What is represented: copied or aggregated?

With raw returns the regime must be computed from twenty numbers, and every model still represents it beyond the matched null. In Qwen held-out separation is significant at all 28 blocks ( $d$  from 1.1 to 2.9), and  $d=1.12$  at  $L^*$ ; the same holds in two further seeds, with the returns written in basis points ( $d$  1.2–3.0), and with real daily returns of 16 large US stocks from a calm (October 2017) and a turbulent (March 2020) window shown under the synthetic firm labels ( $d$  1.7–5.6). GPT-2 separates at 11 of 12 blocks but only weakly at  $L^*$  ( $d=0.46$ ,  $p=0.065$ ). Separation beyond the matched

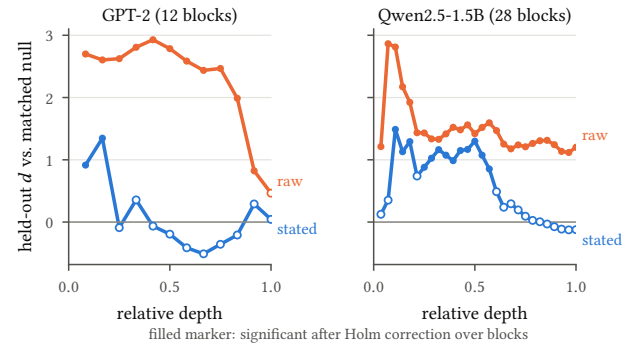


**Figure 1: Where the regime is represented beyond surface variation. Held-out Cohen’s  $d$  of the history contrast against the matched within-regime null, by depth, with directions fit with the firm held out; filled markers are significant after Holm correction over blocks. Raw: the regime is given only as 20 daily returns; stated: as a volatility number. Against the weaker surface null every curve sits higher (Table 1).**

null shows that the regime, not merely which digits appear, is represented. This alone does not show temporal inference: volatility depends on return magnitudes, so a magnitude aggregate would also pass. Two controls separate these readings (C1; Figure 2b). First, replacing the  $k$  most recent of a low-regime firm's 20 returns with its high-regime returns moves the held-out projection gradually: in Qwen at block 27, by 18%, 30%, 44%, 59% and 100% of the full shift for  $k=1, 3, 5, 10, 20$ , and similarly at block 2 and in the other seeds. A single large return does not trigger the signal. Second, shuffling the order of the high-regime returns, which keeps every token and destroys volatility clustering, leaves the projection unchanged (block 2:  $-0.001$ , 95% CI  $[-0.002, 0.001]$ ; block 27:  $+0.04$ ,  $[-0.05, 0.13]$ , against a full shift of 0.31). The represented quantity is therefore an order-invariant aggregate of return magnitudes over the window. That is consistent with a volatility estimate, and it is more than copying a stated number. But it is not evidence of a model of temporal dependence, and we do not claim one. Nor is the direction format-free: in basis points the regime is carried by a direction with cosine 0.21 to  $b$ , and moves the projection on  $b$  only 47% as far.

### 5.3 Intervention survives only as an edit at the readout

Steering along  $b$  at block 27 moves Qwen's  $y_{\text{lex}}$  by  $+0.24$  at  $\alpha=8$  and  $-0.25$  at  $\alpha=-8$  in the stated format, and by  $+0.145$  and  $-0.149$  in the raw format, against at most 0.02 for random and orthogonal directions of the same norm ( $p < .001$  against both). On the decision readout the gap is smaller:  $+0.098$  for  $b$  against a mean absolute effect of 0.050 for random directions (raw; stated:  $+0.250$  vs.  $0.048$ ). Two further checks change how these numbers should be read.

*Strength.*  $\alpha=8$  is not a natural magnitude. The regime itself moves the residual along  $b$  by  $g=0.38$  at block 27 (raw) and 0.26 (stated), so  $\alpha=8$  pushes 21 and 31 times further than the regime does. At  $\alpha=g$ , steering moves  $y_{\text{dec}}$  by  $+0.005$  and  $+0.008$ , about as much as the indirect effect through  $b$  ( $+0.004$ ,  $+0.010$ ; Section 5.5). Per unit of displacement, steering and natural use agree; the regime simply moves  $b$  very little. A comparison of steering at  $\alpha=8$  with the natural effect reflects the choice of  $\alpha$ , not the mechanism.

*Positions.* The edit covers every token from the first marker to the end of the prompt, including the appended question and the readout position. When it stops before the question, steering along  $b$  at block 27 no longer moves  $y_{\text{dec}}$  (raw:  $-0.001$ , random 0.002; stated:  $+0.004$ , random 0.002), and the same holds at every later block (from block 22 raw, 25 stated). At  $L^*$  the intervention therefore works by writing at the positions where the answer is read, not through what later tokens read from the reasoning steps. Middle blocks do propagate in the stated format: without touching the question, steering at blocks 15–24 moves  $y_{\text{dec}}$  two to four times more than random directions (block 19: 0.101 vs. 0.025); in the raw format it does not exceed random directions consistently at any block. Read through the unembedding, the stated-format  $b$  promotes movement tokens in both directions ("down," "downgrade," "Ups," "upwards") and the raw-format  $b$  tokens of alarm and intensity ("intensely," "excessively"); its cosine with the readout's own unembedding axis is inside or just above the range of random directions (0.052 raw, 0.076 stated; 99th percentile 0.068).

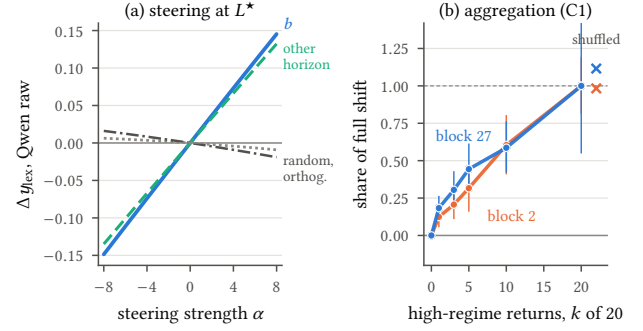


**Figure 2: (a) Steering Qwen on raw returns at  $L^*$ : change in  $y_{\text{lex}}$  against  $\alpha$  for  $b$ , the other horizon's direction, and orthogonal and random directions of equal norm (means over 16 prompts). (b) Aggregation control C1: held-out projection on  $b$  when the  $k$  most recent of 20 low-regime returns are replaced by high-regime returns, as a share of the shift at  $k=20$ , with 95% firm-bootstrap intervals; crosses: the high-regime window with its returns shuffled.**

### 5.4 Specificity is untested by the original design, and fails where it is tested

The near- and medium-term directions are almost identical (cosine 0.99 in Qwen raw, 1.00 in GPT-2), and the other horizon's direction steers about as well as  $b$ . This is not evidence against horizon-specific beliefs: the generator holds the regime constant over both horizons, so there is nothing horizon-specific to represent. We therefore built a case with horizon-specific ground truth (C2): a *transient* window (15 calm returns followed by 5 turbulent ones) against a *persistent* one (20 turbulent, same last five). Medium-term risk should differ more between them than near-term risk. Qwen rates the persistent window as riskier on  $y_{\text{dec}}$ , by  $+0.032$  ([0.015, 0.048]) for the near-term question and  $+0.028$  ([0.014, 0.041]) for the medium-term one, an interaction of  $-0.004$  ( $[-0.009, 0.001]$ ); GPT-2, the other seeds and the real windows show none either. The models do not condition their risk judgement on the horizon even when the ground truth does, so no direction-level test of horizon specificity can succeed on them.  $S$  is untestable here, rather than refuted. Concept specificity is informative. A direction fit on short non-financial sentences contrasting calm with intense scenes (storms, crowds, engines), cosine 0.09 with  $b$ , moves Qwen's  $y_{\text{dec}}$  by  $+0.136$  at block 27, more than  $b$  itself ( $+0.098$ ); and a principal component of the per-pair contrasts moves  $y_{\text{lex}}$  more than  $b$  in both formats (mean  $|\Delta y_{\text{lex}}|$  over eight prompts: 0.63 vs. 0.23 stated, 0.28 vs. 0.18 raw). Steering effectiveness does not single out the volatility concept.

### 5.5 The direction carries a small part of the natural effect, at every block

Switching the context from low to high volatility moves Qwen's  $y_{\text{lex}}$  by only  $+0.011$  (stated) and  $+0.010$  (raw); with total effects this small the proportion mediated is uninterpretable, and the decision readout is the informative one.

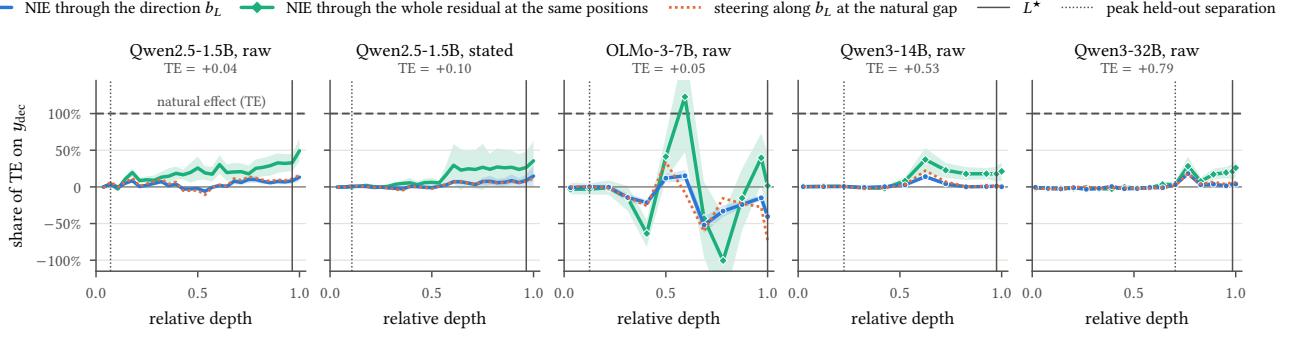


**Figure 3: Natural use at every block, as a share of the natural effect on  $y_{dec}$  (dashed line: 100%). Blue: indirect effect through  $b_L$ , the difference-of-means direction refit at each block. Green: indirect effect through the whole residual at the same reasoning-step positions, which bounds what any direction there can carry. Orange: steering along  $b_L$  at the natural gap  $g_L$ . Bands: 95% firm-bootstrap intervals. Solid vertical line:  $L^*$ ; dotted: block of peak held-out separation. Larger models are swept every third (OLMo) or fourth (Qwen3) block.**

Its natural effect is clearly non-zero,  $TE = +0.043$  ([0.023, 0.063], raw) and  $+0.105$  ([0.025, 0.190], stated). At  $L^*$ , patching only the  $b$ -component carries NIE =  $+0.004$  and  $+0.010$ , about a tenth.  $L^*$  is chosen by raw norm, which favours late blocks where residual norms are large, and a patch confined to the reasoning-step positions of a late block may carry little by construction. We therefore repeated the decomposition at every block, refitting  $b_L$  there, and bracketed it by the full residual at the same positions (Figure 3). Three results follow. First, no block rescues  $b$ : the largest NIE through  $b_L$  anywhere is  $+0.006$  (raw) and  $+0.015$  (stated), at most 15% of TE, including the blocks of peak held-out separation. Second, the positions carry more than  $b$  but far from everything: the full residual carries at most  $+0.021$  and  $+0.037$ , about a third to a half of TE. The rest reaches the answer from the context and question tokens directly; with the question excluded from the patch, even the full residual carries at most a tenth of TE at any block. Third, removing all linearly available regime information at  $L^*$  with LEACE [5] leaves 86% (raw) and 89% (stated) of TE in place. In the stated format the pattern replicates in two further seeds, where TE is larger ( $+0.27$ ,  $+0.19$ ):  $b$  carries at most a tenth at any block and is slightly negative at  $L^*$ , and LEACE leaves the whole effect. In the raw format the natural decision effect is itself seed-dependent: it is absent in seeds 1 and 2 ( $-0.013$ ,  $+0.011$ , both intervals including 0), while it is present on real returns ( $+0.031$ , [0.003, 0.058]), where  $b$  again carries little ( $\approx 0$  at  $L^*$ , at most  $+0.008$  at any block) and LEACE leaves 94% of TE. For GPT-2 the decomposition is now additive but there is no natural decision effect to decompose ( $-0.04$ ,  $[-0.08, 0.01]$ ). The rank-1 patch sets one coordinate and leaves the rest of the residual at its own-run value, so these are effects of  $b$  as a linear coordinate; a learned subspace [16] could carry more, but the full-residual bracket bounds how much more is available at these positions.

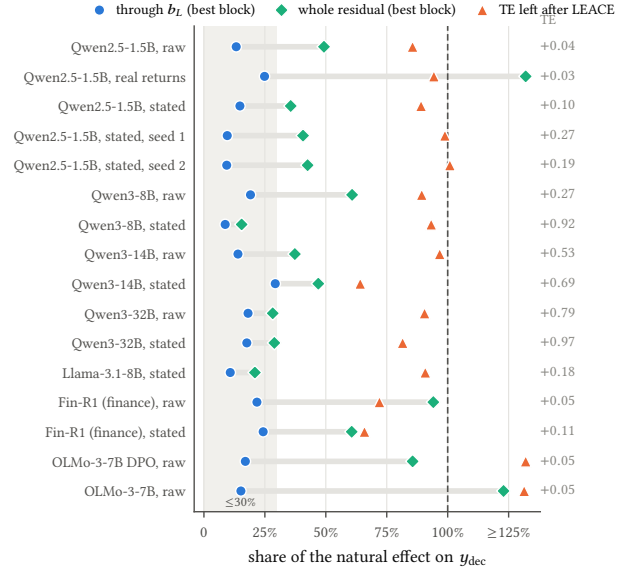


**Figure 4: Every run with a natural effect on  $y_{dec}$  (TE at right). Circles: the largest indirect effect through  $b_L$  at any swept block; diamonds: through the whole residual at the reasoning-step positions; triangles: TE remaining after LEACE erasure of the regime at  $L^*$ . All as a share of TE; values above 125% are drawn at the edge. In no run does the direction carry more than about 30% (shaded).**

## 5.6 Larger, finance-tuned and post-trained models

Appendix Table 2 repeats the representation, steering and mediation checks on Qwen3-8B, -14B and -32B [38], Llama-3.1-8B-Instruct [17], the finance-tuned Fin-R1 [28], and OLMo-3-7B after SFT, DPO and full post-training [44].

Representation holds in every model with raw returns: the regime separates beyond the matched null at every block in all but Qwen3-8B (21 of 36 blocks). The natural effect on  $y_{\text{dec}}$  is much larger than in Qwen-1.5B where it is present (Qwen3-8B +0.27, -14B +0.53, -32B +0.79), and absent in Llama-3.1-8B with raw returns. The pattern for use is the same at every scale. At  $L^*$  the direction carries 1–13% of the natural effect in the Qwen3 models (raw), and is negative in OLMo; at the best block anywhere it carries 14–19%. The full residual at the reasoning-step positions carries at most 28–61% in Qwen3, and LEACE erasure at  $L^*$  leaves 89–97% of the effect. With the stated number the effects are larger (Qwen3-32B: +0.97), and  $b$  carries 12% at  $L^*$  and 18% at best, with 82% left after erasure. In OLMo-3-7B and Fin-R1 the positions carry nearly all of the effect (86–123% of TE at the best block), yet  $b$  carries 15–22%: there the regime information the judgement uses passes through the reasoning steps, but not along the difference-of-means direction. The finance-tuned model is the strongest case for use, at 22% (raw) and 24% (stated) through  $b$  with 66–72% left after erasure, as is Qwen3-14B with the stated number (29%, 64% left); no run exceeds about 30%. The strength the pipeline uses is not comparable across scales:  $\alpha=8$  is 20–30 times the natural gap in Qwen-1.5B but below it in Qwen3 with raw returns (0.2–0.6), so steering strength must be reported in natural units. Post-training changes the natural effect, not the mechanism. In OLMo-3-7B the representation is present after SFT ( $d=1.70$  at  $L^*$ , all blocks), but the natural effect appears only after preference tuning (SFT +0.016, [−0.002, 0.036]; DPO +0.054; final +0.049), and at no stage does  $b$  carry more than 17% of it.

## 6 Discussion and Limitations

*What a validator can claim.* For Qwen with raw returns, a validator can claim that a magnitude-based volatility signal is linearly represented beyond matched surface variation (R), and the same holds with real returns, in basis points, across seeds and in every larger model we tested (at most blocks in Qwen3-8B, all blocks elsewhere). They can claim that pushing along the difference-of-means direction shifts the model’s risk language and its High/Low judgement (I), but only with the strength and positions stated: at the selected block the effect comes from editing the positions where the answer is read, at strengths 20–30 times the natural shift in the small models, and a direction fit on non-financial “intensity” sentences does as well. They cannot claim that this direction is how the model reaches its judgement (U): at no block, in any model, does it carry more than about 30% of the natural effect (in most models under a fifth), even in models where the reasoning-step positions as a whole carry the whole effect, and erasing the regime there leaves at least 64% of the judgement in place. Nor can they claim a horizon-specific belief, since the models’ behaviour is not horizon-specific (S). With the volatility stated as a number, the steering block carries no held-out regime signal at all, so the direction is a control knob rather than an explanation. A report built on the naive numbers would have claimed all four. For a model used to draft risk commentary, the practical consequence is that steering along such a direction is not a faithful way to explain or correct the model’s risk judgements: most of the judgement would be untouched by it.

*Recommendations.* A representation-level explanation offered in model validation should come with: (1) an output-based readout,

never a re-projection on the steered axis, with effects in output units; (2) directions fit out of sample, grouped by issuer, with the layer search corrected and reported at every block rather than at one selected by norm; (3) matched nulls that change as much input as the treatment; (4) a leakage control whenever the prompt states the target quantity; (5) comparisons against competing concept directions, including ones fit outside finance, and against the absolute effect of random directions; (6) steering strength in units of the natural shift, and edits that exclude the readout positions; (7) absolute natural direct and indirect effects with intervals, bracketed by what the whole residual at the same positions carries and by erasure; (8) the natural effect itself shown to be present, across data draws, before it is decomposed; and (9) a planted-ground-truth run of the same pipeline. These lessons are known separately [19, 21, 31, 43, 54]; our point is that on a realistic financial claim each one changes the conclusion.

*Limitations.* The firms and disclosures are synthetic; real return windows replicate the representation result and show a natural decision effect, but real transcripts may behave differently. We use 16 firms per run. Both readouts are single-token proxies, not risk estimates or portfolio decisions. The mediator is the reasoning-step positions: the regime may also be carried, and used, at the context tokens, which have different lengths across regimes and cannot be patched one-to-one. A learned alignment [16] or non-linear encoding could carry more than the linear directions we test, within the full-residual bound. Horizon specificity needs models whose judgements depend on the horizon.

## References

- [1] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063* (2019).
- [2] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*.
- [3] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48, 1 (2022), 207–219.
- [4] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112* (2023).
- [5] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect Linear Concept Erasure in Closed Form. In *Advances in Neural Information Processing Systems*.
- [6] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. 2011. Supervisory Guidance on Model Risk Management (SR Letter 11-7).
- [7] Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. 2021. An Interpretability Illusion for BERT. *arXiv preprint arXiv:2104.07143* (2021).
- [8] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. 2019. *Machine Learning Explainability in Finance: An Application to Default Risk Analysis*. Staff Working Paper 816. Bank of England.
- [9] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, et al. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread* (2023).
- [10] A. Colin Cameron and Douglas L. Miller. 2015. A Practitioner’s Guide to Cluster-Robust Inference. *Journal of Human Resources* 50, 2 (2015), 317–372.
- [11] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems*.
- [12] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *International Conference on Learning Representations*.
- [13] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics* 9 (2021), 160–175.

- [14] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series.
- [15] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*.
- [16] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*.
- [17] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).
- [18] Wes Gurnee and Max Tegmark. 2024. Language Models Represent Space and Time. In *International Conference on Learning Representations*.
- [19] Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. In *Advances in Neural Information Processing Systems*.
- [20] Stefan Heimersheim and Neel Nanda. 2024. How to Use and Interpret Activation Patching. *arXiv preprint arXiv:2404.15255* (2024).
- [21] John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of EMNLP-IJCNLP*.
- [22] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.
- [23] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of ACL*.
- [24] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of NAACL-HLT*.
- [25] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. 2024. Financial Statement Analysis with Large Language Models. *arXiv preprint arXiv:2407.17866* (2024).
- [26] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*.
- [27] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*.
- [28] Zhaowei Liu et al. 2025. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning. *arXiv preprint arXiv:2503.16252* (2025).
- [29] Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *arXiv preprint arXiv:2304.07619* (2023).
- [30] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*.
- [31] Aleksandar Makelov, Georg Lange, and Neel Nanda. 2024. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. In *International Conference on Learning Representations*.
- [32] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In *International Conference on Learning Representations*.
- [33] Samuel Marks and Max Tegmark. 2023. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv preprint arXiv:2310.06824* (2023).
- [34] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*.
- [35] nostalgebraist. 2020. Interpreting GPT: The Logit Lens. LessWrong.
- [36] Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *International Conference on Machine Learning*.
- [37] Qwen Team. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).
- [38] Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).
- [39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners*. Technical Report. OpenAI.
- [40] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of ACL*.
- [41] Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance?. In *Proceedings of EACL*.
- [42] Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of ACL*.
- [43] Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adria Garriga-Alonso, and Robert Kirk. 2024. Analysing the Generalisation and Reliability of Steering Vectors. In *Advances in Neural Information Processing Systems*.
- [44] Team OLMo. 2025. OLMo 3. Allen Institute for AI, <https://allenai.org/olmo>.
- [45] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, et al. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024).
- [46] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear Representations of Sentiment in Large Language Models. *arXiv preprint arXiv:2310.15154* (2023).
- [47] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering Language Models With Activation Engineering. *arXiv preprint arXiv:2308.10248* (2023).
- [48] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*.
- [49] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *International Conference on Learning Representations*.
- [50] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv preprint arXiv:2303.17564* (2023).
- [51] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024. ReFT: Representation Finetuning for Language Models. In *Advances in Neural Information Processing Systems*.
- [52] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.
- [53] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *arXiv preprint arXiv:2306.06031* (2023).
- [54] Fred Zhang and Neel Nanda. 2024. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*.
- [55] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405* (2023).

## A Robustness across models, seeds and data

Table 2 repeats the representation, steering and mediation checks on further data seeds, on real return windows, and on larger, finance-tuned and post-trained models. The 1.5B and GPT-2 runs on the main seed sweep every block; all other runs sweep every third block (every fourth at 14B and 32B) plus  $L^*$  and the last block, with four random directions.

Figure 5 shows the same decomposition for the runs not drawn in Figure 3. Three regularities hold across all of them. First, the direction’s share stays small at every block: the blue curve never rises above about 30% of TE, and it peaks in the second half of the network. Second, the gap between the blue and green curves depends on the model. In Fin-R1 with raw returns and in Qwen-1.5B on real returns the reasoning-step positions carry about the whole effect, yet the difference-of-means direction carries a fifth to a quarter of it at most. In the Qwen3 models with the stated number the whole residual carries at most about half, and the rest reaches the answer from the context and question tokens directly. Third, in most panels steering at the natural gap tracks the blue curve closely: an edit of natural size along  $b_L$  does about what the direction does naturally, and no more.

Two runs need care in reading. In OLMo-3-7B after DPO the curves change sign across blocks: the direction carries a negative share at  $L^*$ , and erasing the regime there increases TE (139%).

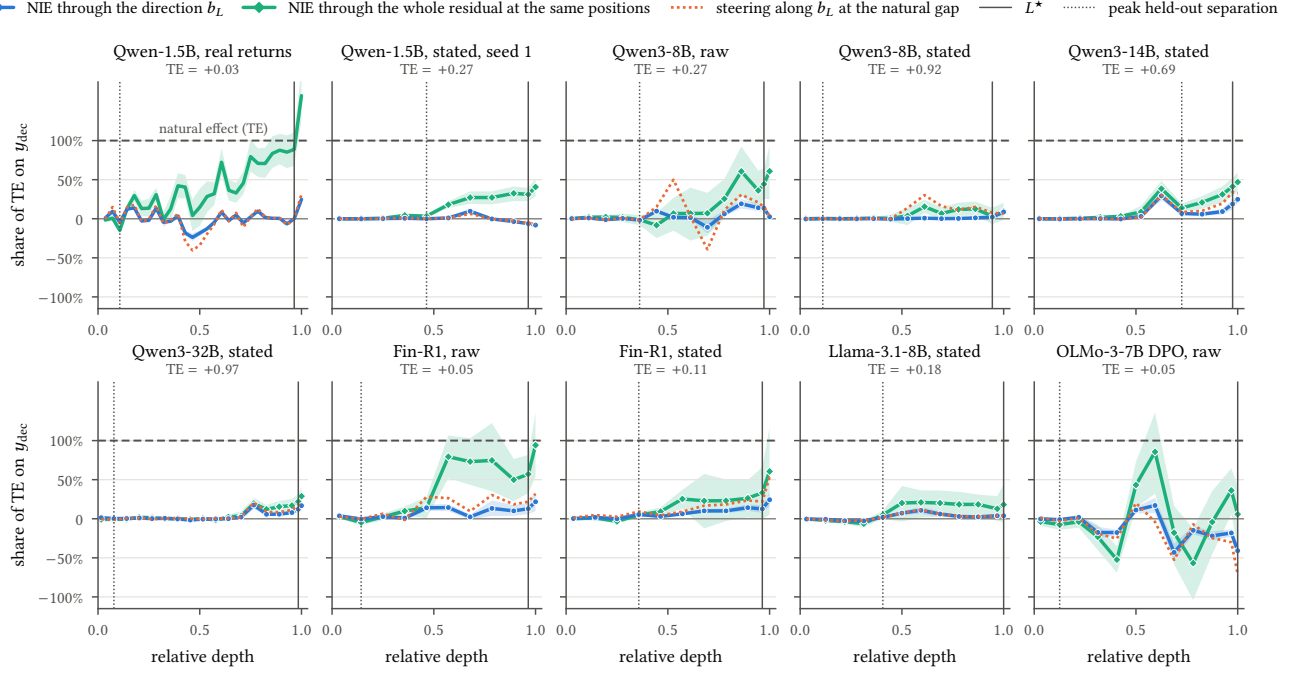


**Figure 5: Natural use at every swept block for the other runs with a natural effect on  $y_{dec}$ , drawn as in Figure 3: indirect effect through  $b_L$  (blue) and through the whole residual at the same positions (green), as a share of TE (dashed line: 100%), with steering along  $b_L$  at the natural gap (orange). In no panel does the blue curve rise above about 30% of TE, although the green curve reaches about the whole effect in Fin-R1 (raw) and on real returns.**

In Qwen3-8B with the stated number TE is large (+0.92) but its interval is wide ([0.24, 1.64]), so the shares are noisy, although the direction carries at most 9% at any block.

## B Implementation details

*Hook placement.* Steering adds  $\alpha b_L$  to the output of block  $L$ . Hugging Face returns the last hidden state after the final normalisation, so for  $L=N$  the vector is added after the final norm, and the projection of the steered state on  $b_N$  then moves by exactly  $\alpha$ , which we check for every model. The submitted GPT-2 results added the vector before  $\ln_f$ . That placement reproduces the submitted numbers exactly (a shift of 4.57 instead of 8.00), and all GPT-2 results in this version use the corrected hook.

*Edited positions.* By default the edit covers every token from the first reasoning-step marker to the end of the prompt, which includes the appended question and the readout position. The *pre-question* control stops the edit at the last token before the question. Mediation patches act on the reasoning-step positions only, pooled per step.

*Precision.* Bf16 logits are quantised in steps of  $2^{-6}$ , comparable to the indirect effects we measure. All models up to 14B therefore run in float32 with TF32 matrix multiplication; Qwen3-32B runs in bf16 with the output head in float32.

*Estimation.* Intervals are percentile intervals from 4,000 bootstrap resamples of firms. LEACE is fit in closed form at each block on the step-pooled activations of both regimes, with a small ridge on the covariance. The rank-8 subspace is spanned by the top singular directions of the per-pair regime differences. Random directions are drawn uniformly on the unit sphere and applied with the same  $\alpha$  as the unit-norm  $b_L$ .

## C Competing directions across models

Section 5.4 tests specificity against two competing directions in Qwen2.5-1.5B; Table 3 repeats the test at  $L^*$  in every model. The result generalises. A direction fit on non-financial calm-vs-intense sentences moves the High/Low judgement more than  $b$  in five of the eleven runs, as much in Qwen3-32B, and by 85% of  $b$ 's effect in Llama-3.1-8B; only with the stated number in Qwen2.5-1.5B does  $b$  clearly do more. In OLMo-3-7B no direction is specific: random directions move  $y_{dec}$  by about 0.19 on average, as much as  $b$ , which moves it the wrong way. The regime direction is also not format-free. With the same returns in basis points the regime is carried by a direction with cosine between  $-0.14$  and  $0.47$  to  $b$ , and moves the projection on  $b$  by at most 31% of the percent-format gap in every model except Qwen2.5-1.5B. A validator who finds a "risk" direction in one prompt format should therefore not expect it to transfer to another.

**Table 2: Robustness of the audit across seeds, real returns, and larger and finance-tuned models. All effects on  $y_{\text{dec}}$ , in log-probability units, with 95% firm-bootstrap intervals for TE.  $L^*$ : steering block (largest mean contrast) / number of blocks.  $d$ : held-out Cohen’s  $d$  against the matched null at  $L^*$  (A1–A2); sig.: blocks significant against it after Holm correction (A3). TE,  $\text{NIE}_b$ ,  $\text{NIE}_{\text{full}}$ : natural effect, and the part carried at  $L^*$  by the coordinate along  $b$  and by the whole residual at the same positions (A7). Best  $\text{NIE}_b$ : largest NIE through  $b_L$  over swept blocks (every block for the 1.5B and GPT-2 runs on seed 0, every third block otherwise). LEACE: TE with the regime erased at  $L^*$ .  $\alpha=8$  /  $g$ : steering strength in units of the natural gap.  $^\circ$ : TE interval includes 0, so there is no natural effect to decompose.**

| Model                 | ctx.   | data              | $L^*$ | $d$   | sig.  | TE [CI]                          | $\text{NIE}_b$ | $\text{NIE}_{\text{full}}$ | best $\text{NIE}_b$ | LEACE  | $\alpha/g$ |
|-----------------------|--------|-------------------|-------|-------|-------|----------------------------------|----------------|----------------------------|---------------------|--------|------------|
| Qwen2.5-1.5B          | raw    | synthetic         | 27/28 | 1.12  | 28/28 | +0.043 [+0.022, +0.064]          | +0.004         | +0.014                     | +0.006 (L28)        | +0.037 | 21         |
|                       | raw    | synthetic, seed 1 | 27/28 | 1.68  | 28/28 | -0.013 [-0.033, +0.004] $^\circ$ | +0.002         | +0.010                     | +0.006 (L28)        | -0.017 | 27         |
|                       | raw    | synthetic, seed 2 | 27/28 | 1.51  | 28/28 | +0.011 [-0.011, +0.034] $^\circ$ | +0.003         | +0.012                     | +0.007 (L28)        | +0.006 | 26         |
|                       | raw    | real returns      | 27/28 | 2.54  | 28/28 | +0.031 [+0.003, +0.058]          | -0.000         | +0.027                     | +0.008 (L28)        | +0.029 | 13         |
|                       | stated | synthetic         | 27/28 | -0.12 | 13/28 | +0.105 [+0.025, +0.190]          | +0.009         | +0.028                     | +0.015 (L28)        | +0.093 | 31         |
|                       | stated | synthetic, seed 1 | 27/28 | 1.26  | 28/28 | +0.274 [+0.185, +0.360]          | -0.017         | +0.085                     | +0.026 (L19)        | +0.271 | 9          |
| GPT-2                 | stated | synthetic, seed 2 | 27/28 | 0.57  | 6/28  | +0.187 [+0.119, +0.272]          | -0.017         | +0.061                     | +0.017 (L19)        | +0.189 | 11         |
|                       | raw    | synthetic         | 12/12 | 0.46  | 11/12 | -0.037 [-0.084, +0.006] $^\circ$ | +0.003         | +0.006                     | +0.007 (L7)         | -0.044 | 1          |
|                       | raw    | real returns      | 12/12 | 2.13  | 12/12 | +0.020 [-0.028, +0.062] $^\circ$ | +0.003         | +0.012                     | +0.008 (L10)        | +0.014 | 0          |
| Qwen3-8B              | raw    | synthetic         | 35/36 | 1.55  | 21/36 | +0.270 [+0.101, +0.440]          | +0.036         | +0.120                     | +0.052 (L31)        | +0.241 | 1          |
|                       | stated | synthetic         | 34/36 | 0.38  | 17/36 | +0.924 [+0.237, +1.640]          | +0.019         | +0.031                     | +0.081 (L36)        | +0.862 | 1          |
| Qwen3-14B             | raw    | synthetic         | 39/40 | 1.53  | 40/40 | +0.534 [+0.328, +0.767]          | +0.005         | +0.095                     | +0.075 (L25)        | +0.516 | 1          |
|                       | stated | synthetic         | 39/40 | 1.74  | 37/40 | +0.685 [+0.199, +1.201]          | +0.129         | +0.282                     | +0.200 (L25)        | +0.439 | 1          |
| Qwen3-32B             | raw    | synthetic         | 63/64 | 2.43  | 64/64 | +0.787 [+0.433, +1.141]          | +0.025         | +0.167                     | +0.142 (L49)        | +0.712 | 0          |
|                       | stated | synthetic         | 63/64 | 1.09  | 36/64 | +0.973 [+0.415, +1.559]          | +0.120         | +0.214                     | +0.171 (L49)        | +0.793 | 0          |
| Llama-3.1-8B-Instruct | raw    | synthetic         | 32/32 | 1.32  | 32/32 | +0.034 [-0.008, +0.075] $^\circ$ | +0.025         | +0.068                     | +0.025 (L32)        | -0.011 | 7          |
|                       | stated | synthetic         | 32/32 | 0.52  | 0/32  | +0.179 [+0.087, +0.273]          | +0.007         | +0.032                     | +0.019 (L19)        | +0.162 | 15         |
| Fin-R1 (7B, finance)  | raw    | synthetic         | 27/28 | 1.26  | 28/28 | +0.052 [+0.022, +0.083]          | +0.007         | +0.030                     | +0.011 (L28)        | +0.037 | 4          |
|                       | stated | synthetic         | 27/28 | 2.26  | 26/28 | +0.105 [+0.041, +0.169]          | +0.013         | +0.035                     | +0.026 (L28)        | +0.069 | 5          |
| OLMo-3-7B SFT         | raw    | synthetic         | 32/32 | 1.70  | 32/32 | +0.016 [-0.002, +0.036] $^\circ$ | -0.010         | +0.018                     | +0.002 (L19)        | +0.025 | 8          |
| OLMo-3-7B DPO         | raw    | synthetic         | 32/32 | 1.99  | 32/32 | +0.054 [+0.023, +0.087]          | -0.022         | +0.003                     | +0.009 (L19)        | +0.075 | 6          |
| OLMo-3-7B Instruct    | raw    | synthetic         | 32/32 | 1.73  | 32/32 | +0.049 [+0.014, +0.084]          | -0.020         | +0.001                     | +0.007 (L19)        | +0.065 | 6          |
|                       | stated | synthetic         | 32/32 | -0.35 | 0/32  | -0.006 [-0.178, +0.170] $^\circ$ | +0.011         | -0.006                     | +0.011 (L32)        | -0.020 | 26         |

**Table 3: Competing directions at  $L^*$  (raw returns unless stated). cos: cosine with  $b$  of the direction fit on non-financial calm-vs-intense sentences (int.) and of the regime direction with returns in basis points (bps). bps/%: regime gap along  $b$  in basis points relative to percent. Right: change in  $y_{\text{dec}}$  from steering each unit direction at  $\alpha=8$ ; |rnd| is the mean absolute effect of random directions.**

| Model         | cos with $b$ |       |       | $\Delta y_{\text{dec}}, \alpha=8$ |       |       |        |
|---------------|--------------|-------|-------|-----------------------------------|-------|-------|--------|
|               | int.         | bps   | bps/% | $b$                               | rnd   | int.  | sent.  |
| Qwen2.5-1.5B  | 0.09         | 0.21  | 0.47  | 0.098                             | 0.050 | 0.136 | -0.080 |
| real returns  | 0.05         | 0.31  | 0.69  | -0.001                            | 0.049 | 0.136 | -0.079 |
| stated        | 0.20         | -     | -     | 0.250                             | 0.048 | 0.144 | -0.088 |
| Qwen3-8B      | 0.31         | -0.14 | -0.09 | 0.025                             | 0.016 | 0.042 | -0.045 |
| Qwen3-14B     | 0.49         | -0.02 | -0.02 | 0.004                             | 0.007 | 0.020 | -0.013 |
| Qwen3-32B     | 0.47         | 0.47  | 0.31  | 0.007                             | 0.008 | 0.007 | -0.013 |
| Llama-3.1-8B  | 0.06         | 0.09  | 0.07  | 0.275                             | 0.029 | 0.234 | -0.082 |
| Fin-R1        | 0.06         | -0.04 | -0.05 | 0.047                             | 0.020 | 0.063 | -0.055 |
| OLMo-3-7B SFT | -0.04        | 0.13  | 0.10  | -0.132                            | 0.194 | 0.136 | -0.231 |
| OLMo-3-7B DPO | -0.02        | 0.22  | 0.15  | -0.241                            | 0.194 | 0.109 | -0.209 |
| OLMo-3-7B     | -0.03        | 0.25  | 0.17  | -0.223                            | 0.194 | 0.115 | -0.216 |

## D Prompt format

Every prompt has the same five parts: a task instruction, the market context, the earnings-call disclosure, a format line, and eight reasoning-step markers. The example below is a high-volatility prompt with raw returns. In a matched pair the low-volatility prompt is identical except for the twenty returns, so the two prompts differ only in the regime. The question line is appended for the decision readout  $y_{\text{dec}}$ , the difference between the next-token

log-probabilities of the first tokens of “High” and “Low” (with a leading space).

```

Assess the downside risk over the near term (roughly the next few trading
sessions). Reason step by step about how the historical market behavior and
the disclosure jointly shape risk.

[MARKET CONTEXT]
Historical market context for TEC00 (tech): daily returns over the last 20
sessions, oldest first: -1.9%, -0.7%, -1.9%, -1.7%, +4.0%, +4.5%, +0.4%, +3.3%,
+1.4%, -1.5%, -2.0%, +4.4%, -6.2%, +0.6%, +2.4%, +0.2%, -0.9%, -4.0%, +5.7%,
-2.1%.

[EARNINGS CALL DISCLOSURE]
Thank you all for joining today's call. The environment is choppy and we are
widening our guidance range. Near-term visibility is limited given recent
market swings. Thank you, and we will now take your questions.

Provide your reasoning as free-form prose.

<step0> reasoning step 1 . . . <step7> reasoning step 8

Final answer. Overall downside risk (High or Low):

```

In the stated format the market context is a single sentence that gives the number directly, for example: “Historical market context for TEC00 (tech) over the trailing 60 sessions: cumulative return +15.8%, annualized volatility 39.4%, trend up.” In the basis-point control the same twenty returns are written in basis points. In the real-return runs they are replaced by the last 20 daily log returns of one of 16 large US stocks in six sectors: AAPL, MSFT, INTC (technology), XOM, CVX, COP (energy), JPM, BAC, WFC (financials), JNJ, PFE, MRK (healthcare), PG, KO (consumer) and HON, CAT (industrials). The calm window ends on 31 October 2017 and the turbulent window on 31 March 2020; the median annualised volatility is 16% and 111% respectively, and every stock is at least 2.5 times as volatile in the second window. Each stock appears under a synthetic firm label from its sector.