DIFFERENT LEARNING TRAJECTORIES, SAME CONCLUSION: AN INVESTIGATION OF LLMS VIA DIFFERENTIAL WARRANT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning traces are increasingly used as training supervision, yet reaching the same conclusion does not imply learning from the same inferential support. We ask whether changing the validity of a critical warrant, the inferential rule that licenses a targeted inference, changes a model's reasoning behavior and how the resulting learning difference is organized internally when the reasoning instance, target inference, and training exposure are held fixed. To answer this question, we introduce Differential Warrant Learning (DWL), a matched counterfactual training framework that isolates this contrast. Across experiments on four different LLMs, Healthy (valid-warrant) training consistently yields a higher reasoning-sensitive Terminal validity margin than matched Severe (invalid-warrant) training, although the effect size varies substantially across models. Bidirectional activation transplantation further identifies later, prediction-associated representations that causally carry the Healthy–Severe behavioral contrast across all four settings. How this learned difference matures through depth, however, is model-dependent: representational magnitude does not track causal influence on behavior, and the Healthy–Severe representational difference decreases even as its decision alignment and finite causal leverage increase in one model. Across three model settings, independently constructed validity-related directions—built without Healthy/Severe condition labels or the Terminal-margin gradient—recover the predicted model-specific depth ordering and causally steer Terminal-validity behavior. Finally, activation-level causal expression does not determine where Healthy parameters can be productively inserted into the Severe-trained model. Together, these results distinguish representational difference, causal expression, and parameter replaceability as related but non-equivalent properties of the same matched learning contrast.

1 INTRODUCTION

Reasoning traces are increasingly used as training supervision (Lightman et al., 2024) for LLMs, yet a correct final answer does not guarantee that the underlying reasoning is faithful or valid (Turpin et al., 2023; Lanham et al., 2023; Bhambri et al., 2026). In this paper, we focus on an important training-time question. Two traces may support the same target inference while differing in whether a critical *warrant*, the inferential rule that licenses that inference, is valid. Because the target inference is unchanged, the traces are indistinguishable at the answer level, but do not provide the same training signal. This motivates the following question:

> *If two matched training traces support the same target inference, but only one uses a valid warrant, do the resulting models exhibit the same reasoning behavior and causally use internal representations in the same way?*

To answer this question, we propose **Differential Warrant Learning (DWL)**, a matched counterfactual training framework. Specifically, for each reasoning instance, we identify a designated dependency

$$P + W_{\text{valid}} \rightarrow B, \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

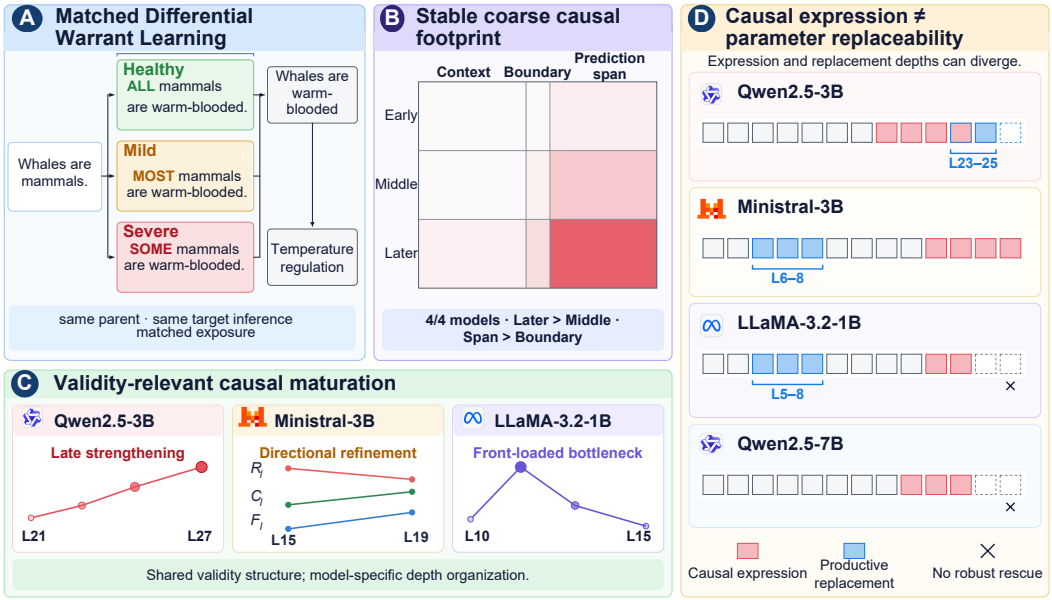

Figure 1: **Overview of Differential Warrant Learning and its causal geometry. A**: DWL degrades the validity of one targeted warrant to different levels (Healthy, Mild, and Severe) while keeping the underlying reasoning opportunity and training exposure fixed. **B**: The resulting Healthy–Severe difference exhibits a stable, coarse causal footprint concentrated in the later layers of the model and associated with its predictions. **C**: This footprint emerges through distinct, model-specific trajectories, i.e., the magnitude of a representation does not necessarily reflect its causal influence. **D**: Causal expression does not coincide with productive parameter replaceability.

where P contains the available premises or source facts, W_{valid} is the targeted warrant, and B is the inference licensed by that warrant. We then construct matched training trajectories that keep the parent reasoning instance and target inference fixed, match the source distribution, corruption family, and training exposure, and vary the validity of the targeted warrant. Our primary goal is to measure the difference between matched Healthy (valid-warrant) and Severe (invalid-warrant) training conditions, rather than the changes relative to the pretrained initialization.

To achieve this, DWL deliberately creates a narrow training contrast dataset as shown in Figure 1. We construct matched Healthy, Mild, and Severe variants by changing only the validity of a designated warrant while keeping the parent reasoning instance and target inference fixed. The goal is not to enlarge capability degradation, but to isolate whether warrant validity changes a reasoning-sensitive behavioral endpoint and the internal representations causally involved in that behavior. Across four different LLMs, the Healthy–Severe effect on Terminal validity varies substantially in magnitude. However, activation transplantation consistently identifies later, prediction-associated representations, defined relative to each model, where swapping Healthy and Severe activations shifts behavior toward the source training condition.

The shared causal effect in model-relative later regions, approximately the final third of each model, raises a deeper question: do different LLMs all induce this effect through the same depth-wise process? A simple explanation is that the Healthy–Severe representation difference grows with depth until it has a larger causal effect on Terminal validity. However, the experimental results rule out this explanation. We use *causal maturation* to describe how the Healthy–Severe representation difference changes through depth in decision alignment and causal effect. Most strikingly, representational magnitude can decrease even as decision alignment and finite causal leverage increase. Thus, these quantities do not evolve monotonically or in parallel across depth. Therefore, we can have the following statement:

Representational magnitude does not determine causal leverage. (2)

Because decision alignment is defined using the Terminal-margin gradient, its depth profile could partly reflect this readout rather than independently recoverable validity-related structure. We therefore construct validity-related directions without Healthy/Severe labels or the Terminal-margin gradient. We test whether these frozen directions separate valid from invalid candidates on held-out data, recover the model-specific depth ordering predicted by the earlier causal analysis, and causally change Terminal validity under steering. These tests recover the predicted depth organization, while stricter robustness analyses identify where it is less stable.

This leaves one final question. If a region causally expresses the learned Healthy–Severe difference, does that also identify where the corresponding parameters can be productively replaced? We find that activation-level causal expression and counterfactual parameter replaceability can occupy different geometries: a representation may be sufficient to carry the learned difference without the local parameters forming an independently replaceable module.

Together, these results show that the Healthy–Severe training contrast need not have a single mechanistic "location." It induces differences in model parameters ($\Delta\theta$), internal representations (Δh), and behavior (Δy), but the corresponding localization criteria need not coincide:

$$\boxed{\text{Representational difference} \not\equiv \text{Causal expression} \not\equiv \text{Parameter replaceability.}} \tag{3}$$

Our contributions are summarized as follows:

- We introduce Differential Warrant Learning (DWL), which varies the validity of a targeted inferential warrant while matching the underlying reasoning instance, target inference, and training exposure. This design isolates a controlled Healthy–Severe training contrast rather than deterioration relative to the pretrained model.
- Across four LLM settings, Healthy training produces a consistent advantage over matched Severe training on Terminal validity. Activation transplantation further identifies later, prediction-associated representations as causally involved in this behavioral difference, despite substantial variation in effect size across models.
- We show that the Healthy–Severe representational difference does not simply grow with depth. Its magnitude, decision alignment, and finite causal leverage follow different model-specific trajectories. Across three model settings, independently constructed validity-related directions recover the corresponding pre-specified depth orderings and causally steer Terminal validity.
- For the same matched Healthy–Severe contrast, regions where activation interventions causally affect behavior do not generally coincide with locations where Healthy parameters can be productively inserted into the Severe-trained model.

2 Related Work

Reasoning supervision and rationale validity. Process supervision uses intermediate reasoning as a training signal rather than supervising only final outcomes (Lightman et al., 2024; Jiang et al., 2025). Yet correct answers and plausible rationales do not guarantee faithful or valid reasoning (Turpin et al., 2023; Lanham et al., 2023; Kawabata & Sugawara, 2024; Paul et al., 2024; Bhambri et al., 2026; Tutek et al., 2025). Fine-tuning can itself alter chain-of-thought reasoning and faithfulness (Lobo et al., 2025), while other work studies learning from incorrect reasoning signals or corrupted supervision (Yao et al., 2025; Wang et al., 2026). Counterfactual Sensitivity Regularization (CSR) is especially close in spirit: it applies operator-level counterfactual perturbations during training and penalizes models whose outputs remain insensitive to logically invalid traces (Shihab et al., 2025). DWL differs in its estimand: rather than optimizing sensitivity to a perturbed trace, it compares matched training trajectories in which the reasoning instance, target inference, and exposure are held fixed while the validity of one designated warrant changes, and then traces the resulting learning difference into internal causal organization.

Validity-related representations and activation steering. Probing and steering methods show that behaviorally relevant properties can be decoded from and manipulated in activation space (Burns et al., 2023; Marks & Tegmark, 2023; Li et al., 2023; Rimsky et al., 2024; Arditi et al., 2024); such

interventions can also transfer across model families in some settings (Oozeer et al., 2025). More closely related, Wu et al. (2025) separate content and logic representations using valid–invalid examples, while Bertolazzi et al. (2026) show that logical validity and semantic plausibility are linearly represented, geometrically related, and causally manipulable under steering. These works establish the existence of validity-related structure, whereas DWL asks how a *training-induced* validity difference is organized through depth under a matched learning contrast. We further test whether the resulting depth ordering can be recovered by a direction constructed using candidate valid-versus-invalid identity, without Healthy/Severe condition labels or the Terminal-margin gradient.

Mechanisms induced by fine-tuning and model diffing. Mechanistic analyses compare related checkpoints to identify how adaptation changes internal computation. Fine-tuning can alter access to existing capabilities through lightweight "wrappers" (Jain et al., 2024), strengthen existing mechanisms (Prakash et al., 2024), reorganize task-relevant circuits (Wang et al., 2025), and leave readable activation differences across model families and scales (Minder et al., 2026). These studies largely compare pretrained and adapted models, or otherwise differently adapted checkpoints. DWL instead compares matched trajectories that differ in one targeted source of inferential support, isolating the learning difference due to warrant validity and tracing it from behavior to representation and causal computation.

Causal localization and parameter intervention. Activation patching, interchange interventions, and causal tracing identify internal states that causally affect model behavior, while methodological work shows that localization can depend on the intervention design and evaluation metric (Geiger et al., 2022; Zhang & Nanda, 2024; Geiger et al., 2025). At the parameter level, prior work localizes and edits factual knowledge (Dai et al., 2022; Meng et al., 2022). Causal localization, however, need not predict where parameter modification is most effective: Hase et al. (2023) find substantial divergence between causality-based localization and knowledge-editing locations, while Weight Patching transplants parameters between paired models to probe candidate source modules (Sun et al., 2026). DWL places these questions in the same controlled setting: *where is a matched learning difference causally expressed, and where do the corresponding parameters remain productively replaceable?* This allows causal-expression and parameter-replaceability geometries to be compared for the same Healthy–Severe training contrast.

3 Differential Warrant Learning

Differential Warrant Learning (DWL) defines a matched training counterfactual for asking how a model learns differently when the same reasoning opportunity is supported by a valid rather than invalid inferential warrant. We then trace this controlled learning difference from behavior into causal computation.

Matched warrant intervention. The intervention is deliberately local. Rather than replacing an entire reasoning trace, we vary the validity of one designated inferential dependency. For each reasoning instance, we identify

$$P + W_{\text{valid}} \to B \to D \to C, \tag{4}$$

where P contains the available premises, W is the targeted *warrant*, B is the inference licensed by that warrant, and D, C denote downstream reasoning that depends on B. Intuitively, W is the reasoning step that makes the inference from P to B justified. Figure 1 (A) gives a simple example. Given the following P and B

$$P:\ \text{Whales are mammals}, \qquad B:\ \text{Whales are warm-blooded}. \tag{5}$$

DWL keeps both P and B fixed while varying only the validity of the designated warrant:

$$\text{HEALTHY}:\ \text{ALL mammals are warm-blooded}, \tag{6}$$

$$\text{MILD}:\ \text{MOST mammals are warm-blooded}, \tag{7}$$

$$\text{SEVERE}:\ \text{SOME mammals are warm-blooded}. \tag{8}$$

The target inference is therefore unchanged, while its inferential justification varies across conditions. The full dataset applies the same principle across a predefined family of warrant corruptions.

Table 1: **DWL intervention design.** Warrant validity is the targeted intervention; other controllable aspects are matched or audited.

Role	Specification
Matched by design	Parent instance, target inference, source/domain distribution, training exposure, and Mild/Severe corruption family
Intervened	Warrant validity: Healthy (valid), Mild (invalid but locally plausible), Severe (stronger violation)
Preserved when contract-consistent	Non-target reasoning content (P, D, C)
Audited	Wording and length; measured rather than assumed matched

Table 2: **Frozen DWL corpus statistics.** Each condition contains the same matched parent set. Effective tokens are computed on the actual training text $P+W+B+D+C$ using `o200k_base`, without EOS or metadata. A frozen manual audit accepts 97.6% of sampled matched interventions.

Condition	Records	Effective tokens	Tokens / record
Healthy	24,111	3,188,713	132.25
Mild	24,111	3,179,133	131.85
Severe	24,111	3,127,397	129.71

To isolate the target dependency $P + W \rightarrow B$, we require the designated warrant to provide non-redundant support for B. Instances are excluded if B remains validly derivable through an independent route, since intervening on W would then leave the target inference justified despite corruption of the designated warrant.

Table 1 summarizes the DWL intervention design. DWL matches the reasoning opportunity and training exposure across conditions, with warrant validity as the targeted intervention. Because changing the warrant can also alter wording and sequence length, we audit and quantify these differences rather than assume the conditions are linguistically identical. Table 2 reports statistics for the final audited corpus.

The realized corpora are similar, though not identical, in length. These differences concern corpus realization rather than training exposure: all compared conditions follow the same pre-specified optimization schedule and processed-token budget. We therefore report residual differences in wording and length as properties of the intervention rather than assume exact linguistic matching.

Accordingly, we interpret DWL as a controlled *warrant-validity-related* contrast rather than a context-free "pure logic" variable. Our primary estimand is the matched Healthy–Severe contrast,

$$\tau_{\mathrm{DWL}} = Y_{\mathrm{Healthy}} - Y_{\mathrm{Severe}}, \tag{9}$$

where Y denotes the evaluation metric of interest, and Y_{Healthy} and Y_{Severe} denote the corresponding metric values for the matched Healthy- and Severe-trained models, respectively.

The pretrained checkpoint serves only as a descriptive reference and is not the causal control. Mild is an auxiliary construction-level severity condition and is not assumed to lie behaviorally between Healthy and Severe. Full construction rules, corruption operators, and audit details appear in Appendix A.

We evaluate this matched learning difference at two levels. First, whether it changes model behavior. Second, whether the corresponding internal difference is causally used to produce that behavior.

Behavioral endpoint. Given a context x and a candidate continuation $c = (c_1, \ldots, c_T)$, we score the candidate by its average token log-probability:

$$s(c \mid x) = \frac{1}{T} \sum_{t=1}^{T} \log p_\theta(c_t \mid x, c_{<t}), \tag{10}$$

where T is the number of tokens in c, c_t is its t-th token, $c_{<t}$ denotes the preceding candidate tokens, and $p_\theta(c_t \mid x, c_{<t})$ is the probability assigned to c_t by the model with parameters θ.

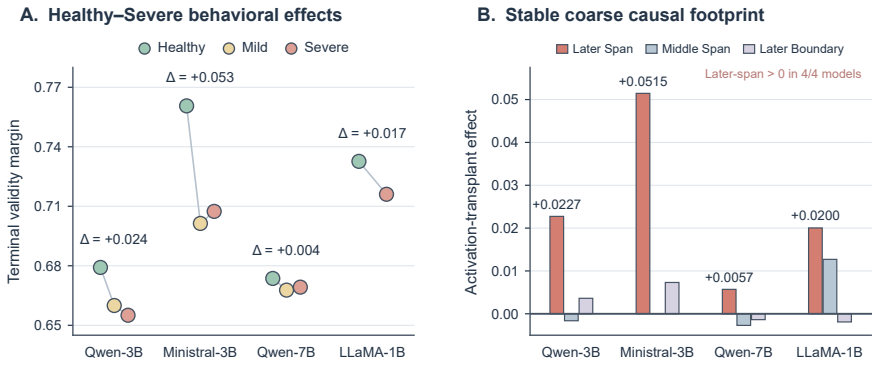

Figure 2: **Behavioral effects vary, but coarse causal expression is more stable.** Healthy training outperforms matched Severe training across all four model settings, while later Prediction-span activations show the most consistent causal transfer between the matched training conditions.

We then define the Terminal validity margin as

$$m_{\text{term}}(x) = s(c^{+}_{\text{term}} \mid x) - \max_{c^{-} \in \mathcal{C}^{-}_{\text{term}}} s(c^{-} \mid x), \tag{11}$$

where c^{+}_{term} denotes the valid terminal continuation and $\mathcal{C}^{-}_{\text{term}}$ denotes the set of matched invalid terminal alternatives used for evaluation. Thus, $m_{\text{term}}(x)$ measures the score advantage of the valid continuation over its highest-scoring matched invalid alternative.

Causal expression. Behavioral separation alone does not establish causal use. We therefore transplant residual activations bidirectionally between matched Healthy and Severe models. In plain terms, we copy selected internal states from one trained model into its matched counterpart and ask whether the target model's behavior moves toward the source training condition. For region R,

$$\Delta^{R}_{H \to S} = m^{R}_{S \leftarrow H} - m_S, \qquad \Delta^{R}_{S \to H} = m_H - m^{R}_{H \leftarrow S}. \tag{12}$$

Positive values indicate source-condition-aligned transfer in both directions.

We compare model-relative middle and later regions under *Prediction-span* and *Boundary* interventions. Prediction-span transplantation replaces the residual states used throughout candidate scoring, whereas Boundary transplantation replaces only the state immediately preceding the first candidate token. All exploratory and confirmatory analyses follow a frozen Discovery/Formal/Reserve protocol. Hypotheses are frozen before held-out evaluation, and failed or reversed predictions are retained as negative evidence rather than re-selected after the fact. Full construction, statistical, and implementation details are given in Appendices A and B.

4 Behavioral Effects Vary, but the Coarse Causal Footprint Is Stable

We evaluate DWL on Qwen2.5-3B, Qwen2.5-7B, Ministral-3B, and LLaMA-3.2-1B; the first three use Healthy/Mild/Severe training, while LLaMA-3.2-1B serves as a prospective Healthy/Severe replication.

We first ask whether warrant validity changes what the model learns. Across all four LLMs, Healthy training produces a reproducible Terminal-margin advantage over matched Severe training (Fig. 2A). The magnitude varies substantially across models, and Mild does not always lie behaviorally between Healthy and Severe. Thus the stable behavioral object is the matched Healthy–Severe contrast, not a universal three-level dose–response law.

The effect is selective rather than a broad capability collapse: Clean-NLL, the token-averaged next-token negative log-likelihood on a fixed clean-text evaluation set, shows no severity-dependent de-

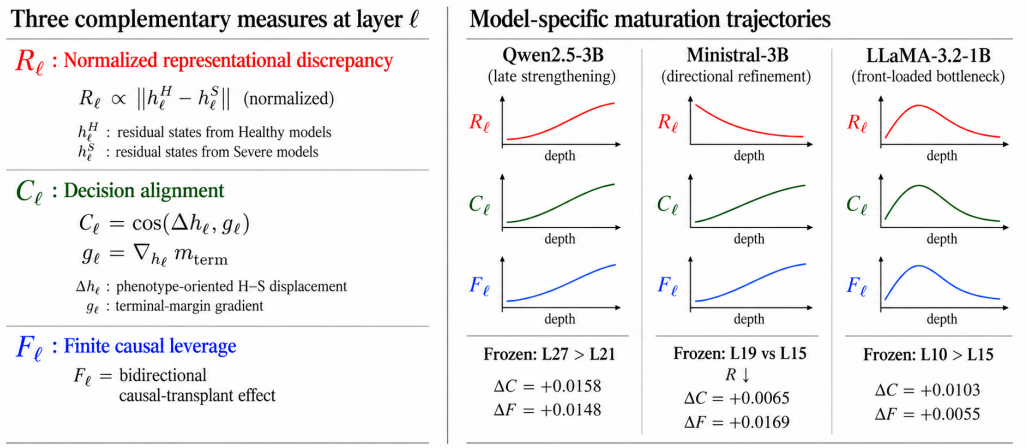

Figure 3: **The same DWL contrast develops differently through depth.** R_ℓ is normalized representational discrepancy, C_ℓ decision alignment, and F_ℓ finite causal leverage. Qwen2.5-3B shows late strengthening, Ministral-3B directional refinement, and LLaMA-3.2-1B an earlier bottleneck followed by dissipation.

terioration, while collateral reasoning benchmarks vary across models. Full behavioral controls are reported in Appendix C.

The internal result is more consistent. Bidirectional activation transplantation shows that later prediction-associated representations transfer the Healthy–Severe behavioral difference in all four settings (Fig. 2B), whereas Middle-span and single-Boundary interventions are weaker or less consistent. This pattern remains visible even in Qwen2.5-7B, the weakest behavioral-effect setting. We therefore use *coarse causal footprint* to represent a broad stage at which the learned DWL difference can causally affect behavior. This does not imply a shared exact layer, module, or microscopic circuit; fine-grained module allocation is less consistent. The contrast motivates the next question: *if different models reach a similar coarse causal footprint, do they get there in the same way?*

5 The Same Causal Footprint Can Arise in Different Ways

5.1 A Bigger Representation Difference Is Not Necessarily More Causal

A simple explanation for the coarse causal footprint in Section 4 is that the Healthy–Severe difference merely grows with depth until it becomes behaviorally important. We distinguish three properties of this learned difference: its normalized magnitude, decision alignment, and finite causal leverage. Let $h_{\ell,p}^H(x)$ and $h_{\ell,p}^S(x)$ denote the Healthy and Severe residual states at layer ℓ and prediction-associated position p, and let $\mathcal{P}_x$ be the Prediction-span positions for question x. We define normalized representational discrepancy as

$$R_\ell(x) = \frac{1}{|\mathcal{P}_x|} \sum_{p \in \mathcal{P}_x} \frac{\left\| h_{\ell,p}^H(x) - h_{\ell,p}^S(x) \right\|_2}{\frac{1}{2}\left(\left\| h_{\ell,p}^H(x) \right\|_2 + \left\| h_{\ell,p}^S(x) \right\|_2 \right) + \epsilon}, \qquad R_\ell = \mathbb{E}_x[R_\ell(x)]. \tag{13}$$

Let Δh_ℓ denote the Healthy–Severe-oriented residual displacement and let $g_\ell = \nabla_{h_\ell} m_{\text{term}}$ denote the local Terminal-margin gradient. We define

$$C_\ell = \cos(\Delta h_\ell, g_\ell), \qquad F_\ell = \frac{1}{2}\left(\mathbb{E}_x[\Delta_{H \to S}^\ell(x)] + \mathbb{E}_x[\Delta_{S \to H}^\ell(x)] \right). \tag{14}$$

Thus, R_ℓ measures the normalized magnitude of the Healthy–Severe difference, C_ℓ measures its alignment with the local decision direction, and F_ℓ measures its finite causal effect on behavior.

Qwen2.5-3B, Ministral-3B, and LLaMA-3.2-1B follow different causal-maturation trajectories (Fig. 3).

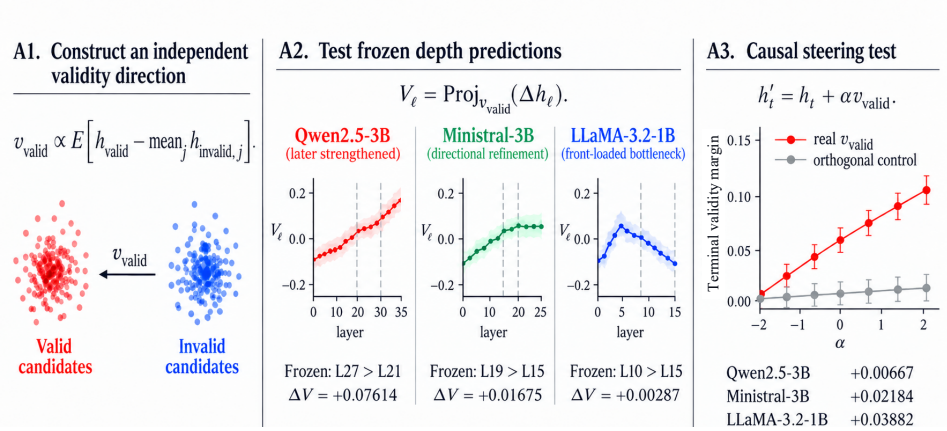

Figure 4: **An independent validity-related direction recovers the predicted depth organization.** A1 tests held-out readability, A2 the frozen depth prediction, and A3 causal steering against an orthogonal control.

Throughout this section, numerical subscripts denote layer indices.

Qwen2.5-3B most closely follows the simple amplification account:

$$C_{27} > C_{21}, \qquad F_{27} > F_{21}, \tag{15}$$

with both frozen contrasts replicating across all three training seeds. Ministral-3B provides the clearest counterexample:

$$R_{19} < R_{15}, \qquad C_{19} > C_{15}, \qquad F_{19} > F_{15}. \tag{16}$$

Its normalized Healthy–Severe difference becomes smaller while becoming more decision-aligned and more causally effective. A larger representational difference is therefore not required for greater behavioral control.

LLaMA-3.2-1B shows a third pattern:

$$C_{10} > C_{15}, \qquad F_{10} > F_{15}, \tag{17}$$

revealing an earlier high-leverage bottleneck followed by dissipation.

We use *causal maturation* for these depth-dependent changes in how a learned difference controls behavior. The central result is

$$\boxed{\text{representational magnitude does not determine causal leverage.}} \tag{18}$$

5.2 Is This Pattern Just an Artifact of Our Metric?

Using Discovery data only, at a pre-specified construction layer ℓ^*, we define

$$v_{\text{valid}} \propto \mathbb{E}_{\text{state},\,a}\left[\text{centroid}(h^{\ell^*}_{\text{valid}}) - \text{mean}_j\, \text{centroid}(h^{\ell^*}_{\text{invalid},j})\right]. \tag{19}$$

Here $a \in \{H, S\}$ indexes the two training conditions, which contribute symmetrically, and j indexes invalid candidates. Only the candidate valid-versus-invalid identity is used as supervision; neither Healthy/Severe condition identity nor the Terminal-margin gradient is used. The resulting direction is unit-normalized and frozen before evaluation.

The frozen direction separates valid from invalid candidates on held-out data in all three models. More importantly, Figure 4.A2 (frozen depth prediction) recovers the depth ordering predicted by the earlier causal-maturation analysis: Qwen2.5-3B and Ministral-3B show a stronger validity-related component at the later layer of their pre-specified depth pair, whereas LLaMA-3.2-1B shows the stronger component at its earlier bottleneck. The procedure therefore does not simply favor later layers.

Figure 4.A3 further shows that steering along the same frozen direction changes Terminal validity more strongly than a frozen orthogonal control. The recovered structure is therefore behaviorally relevant rather than merely decodable. A stricter crossed-anchor robustness test remains positive for Qwen2.5-3B and Ministral-3B but is mixed for LLaMA-3.2-1B (Appendix E.9).

Taken together, Sections 5.1 and 5.2 give a simple picture: DWL induces a learned difference whose coarse causal expression is comparatively similar across models, but the depth-wise route by which that difference becomes behaviorally controlling is model dependent.

This raises one final question. If we know where a learned difference is causally expressed, does that also tell us where the corresponding parameters can be productively replaced?

6 Causal Expression Does Not Prescribe Parameter Replaceability

Section 5 showed that representational difference and causal leverage need not peak at the same depth. We therefore ask whether causal expression also identifies where Healthy parameters can be productively inserted into the Severe model.

For a selected parameter set $\mathcal{P}$, we define the replacement effect for question x as

$$\Delta_{\text{replace}}(x) = m_{\text{term}}\left(x; \theta_{\mathcal{P}}^{S \leftarrow H}\right) - m_{\text{term}}\left(x; \theta^{S}\right), \tag{20}$$

where $\theta_{\mathcal{P}}^{S \leftarrow H}$ is the Severe checkpoint with parameters in $\mathcal{P}$ replaced by their matched Healthy values. Positive values indicate movement toward Healthy-trained behavior, and reported effects are averaged over evaluation questions.

Activation transplantation and parameter replacement probe distinct mechanistic objects: the former asks where an internal representation is sufficient to carry the Healthy–Severe difference, whereas the latter asks where Healthy parameters remain productive within the surrounding Severe model.

The two geometries do not generally coincide. Qwen2.5-3B shows partial co-location, although productive late replacement requires broad parameter coverage rather than a sharply localized subset. Ministral-3B shows a clearer dissociation, with productive replacement substantially upstream of its main causal-expression region. LLaMA-3.2-1B shows the same qualitative separation: its pre-specified late replacement fails, while an earlier region is strongly productive; because this upstream effect was not prospectively predicted, we retain it as secondary exploratory evidence. Qwen2.5-7B provides the limiting case: later causal expression remains detectable, yet none of the tested local replacement operators yields robust positive rescue.

Same-condition cross-seed swaps produce a smaller but non-zero compatibility background, providing a reference scale for generic checkpoint effects (Appendix G). These controls do not assume zero cross-checkpoint interaction; rather, they show that the principal replacement effects are not explained solely by generic cross-seed incompatibility. Thus, causal-expression and parameter-replaceability geometries do not coincide: a representation can causally carry a learned difference without the parameters at that location forming an independently replaceable module. Causal localization therefore identifies where a learned difference can affect behavior, but does not by itself prescribe where the corresponding parameters should be replaced.

7 Conclusion

We introduced Differential Warrant Learning (DWL), a matched counterfactual framework that isolates the effect of warrant validity while holding the reasoning instance, target inference, and training exposure fixed. Across four model settings, DWL yields model-dependent behavioral effects but a comparatively stable later, prediction-associated causal footprint. This footprint matures differently across models: representational magnitude, decision alignment, and finite causal leverage need not evolve together. In three settings, independently constructed validity-related directions recover the predicted depth ordering and causally steer Terminal validity. Parameter replacement reveals a further dissociation: causal expression need not coincide with productive parameter replaceability. Thus, representational difference, causal expression, and parameter replaceability are distinct properties of the same matched learning contrast.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and language editing, and with the creation and refinement of scientific figures. Generative model APIs were also used during dataset construction to instantiate candidate warrants under a fixed, pre-specified intervention contract; they did not select parent instances, target dependencies, corruption-operator families, or acceptance criteria. Generative AI tools were not used as a source of experimental results. All dataset construction procedures, scientific claims, results, interpretations, and final presentation were reviewed and verified by the authors, who take full responsibility for the content of this paper.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Leonardo Bertolazzi, Sandro Pezzelle, and Raffaella Bernardi. How language models conflate logical validity with plausibility: A representational analysis of content effects. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9316–9347, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.454. URL https://aclanthology.org/2026.findings-acl.454/.

Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. Interpretable traces, unexpected outcomes: Investigating the disconnect in trace-based knowledge distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36390–36408, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1686. URL https://aclanthology.org/2026.acl-long.1686/.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0774.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects

of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/29b5de0a8925cb4920d9c64542108335-Abstract-Conference.html.

Dongwei Jiang, Guoxuan Wang, Yining Lu, Andrew Wang, Jingyu Zhang, Chuyu Liu, Benjamin Van Durme, and Daniel Khashabi. RATIONALYST: Pre-training process-supervision for improving reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26547–26566, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1288. URL https://aclanthology.org/2025.acl-long.1288/.

Akira Kawabata and Saku Sugawara. Rationale-aware answer verification by pairwise self-evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16178–16196, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.905. URL https://aclanthology.org/2024.emnlp-main.905/.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi: 10.48550/arXiv.2307.13702. URL https://arxiv.org/abs/2307.13702.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11679–11698, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.584. URL https://aclanthology.org/2025.naacl-long.584/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow finetuning leaves clearly readable traces in activation differences. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3b939edfdf9c1211fb764a888078f13d-Abstract-Conference.html.

Narmeen Fatimah Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amir Abdullah. Activation space interventions can be transferred between large language

models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47212–47286. PMLR, 2025. URL https://proceedings.mlr.press/v267/oozeer25a.html.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.882. URL https://aclanthology.org/2024.findings-emnlp.882/.

Nikhil Prakash, Tamar Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2082273791021571c410f41d565d0b45-Abstract-Conference.html.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Counterfactual sensitivity for faithful reasoning in language models. *arXiv preprint arXiv:2509.01544*, 2025. doi: 10.48550/arXiv.2509.01544. URL https://arxiv.org/abs/2509.01544.

Chenghao Sun, Chengsheng Zhang, Guanzheng Qin, Rui Dai, and Xinmei Tian. Weight patching: Toward source-level mechanistic localization in LLMs. *arXiv preprint arXiv:2604.13694*, 2026. doi: 10.48550/arXiv.2604.13694. URL https://arxiv.org/abs/2604.13694.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract.html.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9935–9960, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.504. URL https://aclanthology.org/2025.emnlp-main.504/.

Fuxin Wang, Amr Alazali, and Yiqiao Zhong. How does unfaithful reasoning emerge from autoregressive training? a study of synthetic experiments. *arXiv preprint arXiv:2602.01017*, 2026. doi: 10.48550/arXiv.2602.01017. URL https://arxiv.org/abs/2602.01017.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63088–63112. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25ak.html.

Xin Wu, Yuqi Bu, Yifei Chen, and Yi Cai. Content-free logical modification of large language model by disentangling and modifying logic representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25507–25515, 2025. doi: 10.1609/aaai.v39i24.34740. URL https://doi.org/10.1609/aaai.v39i24.34740.

Jihan Yao, Wenxuan Ding, Shangbin Feng, Lucy Lu Wang, and Yulia Tsvetkov. Varying shades of wrong: Aligning LLMs with wrong answers only. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=p74CpDzw1Y.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Table 3: Matched construction of the three training conditions.

Property	Healthy	Mild	Severe
Parent identity	same	same	same
Target inference	same	same	same
Operator family	–	same	same
Warrant status	valid	invalid, locally plausible	invalid, stronger violation
Non-target content	preserved whenever contract-consistent		

A Controlled Warrant Intervention and Human Quality Audit

DWL constructs matched training counterfactuals in which the validity of support for one designated inference is varied while the underlying reasoning instance remains matched. This appendix specifies the intervention, construction criteria, and human quality audit.

A.1 Reasoning contract and intervention target

Each accepted parent is represented by the experimental dependency

$$P + W_{\text{valid}} \to B \to D \to C, \tag{21}$$

where P denotes premises or source facts, W_{valid} a warrant that licenses a designated bridge conclusion B, and D, C downstream reasoning that substantively depends on B. The intervention target is the edge

$$P + W \to B. \tag{22}$$

We retain a parent only when the designated warrant genuinely supports B and when no independent valid route makes that warrant redundant. This is an experimental abstraction, not a claim that arbitrary natural-language reasoning admits a unique decomposition into P, W, B, D, C. Parents without a sufficiently explicit and manipulable target dependency are excluded.

A.2 Matched Healthy, Mild, and Severe Variants

For each accepted parent, we construct

$$\textsc{Healthy}: \quad P + W_{\text{valid}} \to B, \tag{23}$$

$$\textsc{Mild}: \quad P + W_{\text{mild}} \not\Rightarrow B, \tag{24}$$

$$\textsc{Severe}: \quad P + W_{\text{severe}} \not\Rightarrow B. \tag{25}$$

All three variants share the same parent and target inference. Mild and Severe use the same corruption-operator family: Mild remains locally plausible but fails to license B, whereas Severe introduces a stronger violation along the same corruption axis.

Non-target reasoning content (P, B, D, C) is preserved whenever doing so is consistent with the intervention. Thus, parent identity, target edge, corruption family, and training exposure are matched by design, while wording and length are measured rather than assumed identical.

Severity is defined at the level of data construction, not as a required behavioral ordering. Accordingly, DWL does not assume that every model or metric satisfies $\textsc{Healthy} > \textsc{Mild} > \textsc{Severe}$. We therefore interpret the construction as a controlled warrant-validity-related supervision contrast rather than a context-free "pure logic" variable.

A.3 Corruption Operators

We use ten corruption-operator families: necessary-condition deletion, sufficient–necessary reversal, quantifier shift, scope or modality shift, sign or direction inversion, boundary-condition removal, invalid algebraic transformation, overgeneralization, correlation–causal substitution, and order or dependency corruption.

For each parent, Mild and Severe remain within the same operator family and target edge. Severe increases the strength of the same violation rather than introducing an additional, unrelated error.

Table 4: **Corruption-operator taxonomy.** Mild and Severe vary along the same operator-specific corruption axis.

Operator	Mild → Severe realization
Necessary-condition deletion	relax or omit a required condition → remove a more central requirement
Sufficient–necessary reversal	plausible converse → categorical converse/equivalence
Quantifier shift	nearby count/threshold change → broader change on the same axis
Scope/modality shift	modest broadening/certainty shift → substantially more categorical claim
Sign/direction inversion	reverse one sign or direction → strengthen the same directional error
Boundary-condition removal	relax one restriction → remove a more central boundary condition
Invalid algebraic transformation	nearby-looking invalid manipulation → stronger version of the same manipulation
Overgeneralization	extend to a nearby class → extend substantially farther
Correlation–causal substitution	move toward causal language → treat association as direct/deterministic support
Order/dependency corruption	misassign one dependency → corrupt more structure within the same dependency relation

Table 5: **Major design controls.** Quantities controlled by construction are matched directly; residual differences in wording and length are audited empirically.

Quantity	Control
Number of examples	24,111 examples per condition
Parent / domain distribution	Identical across conditions by matched-parent construction
Target dependency	Same designated $P + W \rightarrow B$ edge within each parent
Corruption family	Mild and Severe use the same operator family
Non-target reasoning content	Preserved whenever consistent with the intervention contract
Training exposure	192 optimizer steps and 3,145,728 processed tokens per condition within each model
Wording and length	Token, character, sentence, and warrant-length differences audited pairwise

A.4 Construction Controls and Corpus Composition

The intervention is specified before linguistic realization. Generative models are used only to instantiate candidate warrants under the fixed intervention contract; they do not select the parent, target edge, operator family, or acceptance criteria. Accepted examples are rendered deterministically thereafter.

The final corpus contains 24,111 matched parents, yielding 24,111 examples per condition. Because all conditions share the same parent set, parent and source-domain distributions are identical by construction. The source composition is 17,810 Mathematics (73.87%), 1,136 Statistics (4.71%), 1,151 Theoretical Computer Science (4.77%), 826 AI/Data Science (3.43%), and 3,188 Science/Engineering (13.22%). The corpus therefore spans five broad domains but is not domain-balanced.

A.5 Human Quality Audit

We manually audit 2,000 sampled matched interventions using pre-specified criteria covering target-edge alignment, Healthy-warrant validity, corruption invalidity, severity ordering, exclusion of alternative valid routes, preservation of non-target reasoning content, and potential semantic or wording-related confounds.

Of the 2,000 audited interventions, 1,952 satisfy all criteria and 48 do not, yielding an overall acceptance rate of

$$\frac{1952}{2000} = 97.6\%. \tag{26}$$

Because the audit protocol did not pre-specify a taxonomy of failure modes, we report the aggregate pass/fail result rather than retrospectively categorizing the 48 failures.

Table 6: **Models and controlled training conditions.**

Model	Layers	Conditions	Seeds
Qwen2.5-3B	36	H/M/S	3
Qwen2.5-7B	28	H/M/S	3
Ministral-3B	26	H/M/S	3
LLaMA-3.2-1B	16	H/S	3

A.6 Representative Interventions

Sign or direction inversion. Consider a Darboux-sum argument in which the ordering between lower and upper quantities licenses a target conclusion about the integral gap. Healthy preserves the valid inequality, Mild reverses a single comparison while remaining locally plausible, and Severe strengthens the same directional error. The underlying mathematical problem, target inference, and surrounding reasoning remain unchanged.

Invalid algebraic transformation. For ordinary least squares,

$$\sum_i (Y_i - \bar{Y})(X_i - \bar{X}) = \sum_i Y_i (X_i - \bar{X}), \tag{27}$$

because

$$\sum_i (X_i - \bar{X}) = 0. \tag{28}$$

Healthy applies this identity correctly. Mild introduces a locally plausible but unjustified correction term, whereas Severe strengthens the same invalid transformation. The variables, estimator, and target inference are held fixed.

B Models, Training, Evaluation, and Evidence Protocol

B.1 Models and Matched Continual Pretraining

We study four pretrained language models from three model families: Qwen2.5-3B, Qwen2.5-7B, Ministral-3B, and LLaMA-3.2-1B. Qwen2.5-3B, Qwen2.5-7B, and Ministral-3B use the full Healthy/Mild/Severe design, whereas LLaMA-3.2-1B serves as a prospective Healthy/Severe replication. For Ministral-3B, the language backbone and tied language-model head are updated while vision-specific components remain frozen.

Within each model, all compared training conditions use the same optimization configuration: BF16 precision, maximum sequence length 512, AdamW with learning rate 5×10^{-6} and weight decay 0.1, effective batch size 32, 192 optimizer steps, and exactly 3,145,728 processed tokens per condition. Independent seeds correspond to separate optimization trajectories; question-level observations from different seeds are not treated as independent training replicates.

B.2 Evaluation Metrics

Reasoning evaluator. The primary evaluator uses explicit inference structures from ProofWriter and EntailmentBank. For a candidate continuation $c = (c_1, \ldots, c_T)$ under context x, we use the average token log-probability

$$s(c \mid x) = \frac{1}{T} \sum_{t=1}^{T} \log p_\theta(c_t \mid x, c_{<t}). \tag{29}$$

Only candidate tokens contribute to the score. We append no EOS token, apply no chat template, use no free-form decoding, and do not truncate candidate continuations.

The primary behavioral metric is the Terminal validity margin

$$m_{\text{term}}(x) = s(c^+_{\text{term}} \mid x) - \max_{c^- \in \mathcal{C}^-_{\text{term}}} s(c^- \mid x). \tag{30}$$

The analogous Temporal next-step margin and candidate accuracy are secondary metrics.

Table 7: **Mechanistic evaluation packets.**

Packet	N	Role
Discovery	300	Exploratory localization / hypothesis formation
Formal	600	Zero-overlap held-out or cross-packet confirmation
Reserve	600	Pre-specified follow-up module / replaceability tests

Clean-NLL. Clean-NLL measures next-token prediction on a fixed clean-text evaluation set. For clean text $x_i = (x_{i,1}, \ldots, x_{i,L_i})$, let

$$N_i = L_i - 1 \tag{31}$$

denote the number of scored tokens, and define

$$\mathrm{NLLSum}_i = -\sum_{t=2}^{L_i} \log p_\theta(x_{i,t} \mid x_{i,<t}). \tag{32}$$

Clean-NLL aggregates uniformly over all scored tokens:

$$\text{Clean-NLL} = \frac{\sum_i \mathrm{NLLSum}_i}{\sum_i N_i}. \tag{33}$$

Lower values indicate better predictive fit to the clean-text distribution. Because tokenization differs across model families, absolute Clean-NLL values across architectures are treated descriptively; degradation is assessed within each model family.

ARC-Challenge and RULER are used as additional collateral controls. None of these collateral metrics is used to select causal regions or parameter-replacement locations.

B.3 Discovery, Formal, and Reserve Packets

The 1,500 Terminal questions are partitioned into mutually disjoint Discovery, Formal, and Reserve packets shared across model families. Their roles are summarized in Table 7.

Packet status is endpoint-specific. A packet previously used for one analysis is not described as globally untouched when a newly defined endpoint is later evaluated on it. This distinction is retained explicitly for subsequent causal-maturation and parameter-replacement analyses.

B.4 Statistical Inference and Evidence Provenance

Primary comparisons are paired at the question level. Unless otherwise stated, 95% confidence intervals use 20,000 paired bootstrap resamples; one-sided paired sign-flip or permutation tests use 100,000 Monte Carlo draws; and Holm correction is applied within each pre-specified testing family.

We distinguish four evidence categories:

1. **Held-out confirmatory:** a pre-specified hypothesis evaluated on a zero-overlap endpoint;
2. **Unexpected held-out:** a held-out result that reverses or differs from the pre-specified prediction;
3. **Exploratory/post-hoc follow-up:** an analysis motivated by an earlier observation unless subsequently evaluated on a newly pre-specified endpoint;
4. **Negative mechanistic:** a pre-specified intervention that fails or reverses its predicted effect.

Unexpected and negative outcomes retain their original evidence status rather than being relabeled after inspection.

B.5 Activation-Transplantation Protocol and Implementation Checks

For a region R, Healthy-to-Severe transplantation is scored as

$$\Delta^R_{H \to S} = m^R_{S \leftarrow H} - m_S, \tag{34}$$

and the reverse direction as

$$\Delta^{R}_{S\to H} = m_H - m^{R}_{H\leftarrow S}. \tag{35}$$

Positive values therefore indicate phenotype-aligned transfer in both directions.

Prediction-span transplantation replaces the residual states used throughout candidate scoring, whereas Boundary transplantation replaces only the residual state immediately preceding the first candidate token.

The finalized transplantation pipelines satisfy two fail-closed implementation checks: transplantation-disabled scoring must exactly reproduce the unpatched evaluator, and self-patching must leave the result unchanged. The maximum numerical discrepancy for these checks is 0.0.

For later cross-runtime directional analyses, newly defined quantities are compared within a single runtime. Small discrepancies relative to historical cached logits are treated as reproducibility diagnostics rather than attributed to a specific hardware cause.

B.6 Independent Validity-Direction Evidence Protocol

The independent validity-direction analyses follow the same construction-and-test procedure for Qwen2.5-3B, Ministral-3B, and LLaMA-3.2-1B. Candidate valid-versus-invalid identity is the only supervision used to construct the direction; Healthy/Severe condition identity and the Terminal-margin gradient are not used for direction fitting.

For each model, the construction layer and comparison depth pair are fixed from the previously established causal-maturation analysis before the independent-direction endpoint is evaluated. The protocol consists of three stages:

1. **A1 – held-out readability:** construct the direction on Discovery300 and evaluate valid-versus-invalid separation on Reserve600;
2. **A2 – frozen depth prediction:** carry the same frozen direction across the pre-specified depth pair and test whether the Healthy–Severe validity-aligned component follows the previously established causal-maturation ordering;
3. **A3 – causal steering:** intervene along the frozen direction on Formal600 and compare the real steering effect against a fixed random orthogonal direction.

After the protocol is fixed, we do not rescan layers, reconstruct the real direction, retune steering strength, reroll the orthogonal control, or replace a failed depth hypothesis. Negative outcomes are retained as negative mechanistic evidence.

For the optimized capture implementation used in the Qwen and LLaMA extensions, single- and dual-layer capture exactly reproduce the historical all-layer capture on the audited inputs (maximum absolute difference $= 0.0$). At $\alpha = 0$, real, orthogonal, and shuffled steering also reproduce the unsteered baseline exactly.

C Full Behavioral Evidence

The primary causal estimand throughout the paper is the matched Healthy–Severe difference. The pretrained checkpoint is used only as a descriptive reference and is not treated as the causal control.

C.1 Final-Checkpoint Terminal Margins

Table 8 summarizes final-checkpoint Terminal validity margins averaged across three training seeds. Qwen2.5-3B exhibits the clearest monotonic ordering, Healthy $>$ Mild $>$ Severe. For Ministral-3B, Healthy is clearly separated from both corrupted conditions, while Mild and Severe do not exhibit a monotonic ordering. Qwen2.5-7B shows a smaller Healthy–Severe difference, and LLaMA-3.2-1B independently reproduces a positive Healthy–Severe contrast.

Table 8: **Final-checkpoint Terminal validity margins, averaged across three training seeds.**

Model	Healthy	Mild	Severe	H–S
Qwen2.5-3B	0.679168	0.659970	0.655072	+0.024095
Ministral-3B	0.760608	0.701354	0.707426	+0.053182
Qwen2.5-7B	0.673616	0.667744	0.669227	+0.004389
LLaMA-3.2-1B	0.732642	–	0.716052	+0.016590

Table 9: **Seed-level Healthy–Severe replication.**

Model	Seed	Healthy	Severe	H–S
Qwen2.5-3B	1	0.676372	0.653638	+0.022734
	2	0.676630	0.655736	+0.020894
	3	0.684501	0.655843	+0.028658
Qwen2.5-7B	1	0.675136	0.668725	+0.006410
	2	0.672806	0.667789	+0.005017
	3	0.672906	0.671166	+0.001741
LLaMA-3.2-1B	1	0.730997	0.715106	+0.015891
	2	0.734398	0.717181	+0.017218
	3	0.732531	0.715870	+0.016661

C.2 Seed-Level Replication

The Healthy–Severe contrast is positive in every seed reported in Table 9. For LLaMA-3.2-1B, the omnibus paired effect is $+0.016590$ with 95% CI $[+0.015003, +0.018182]$.

C.3 Behavioral Emergence During Training

The Healthy–Severe difference emerges before the final checkpoint. For Qwen2.5-3B Seed 2, the Terminal gap is already positive at 25% of the training schedule ($+0.015768$), increases by the 50% checkpoint, and remains positive through the end of training. Ministral-3B similarly shows a positive gap by 25% ($+0.033400$), which increases to $+0.055921$ by 50% before plateauing. The final behavioral difference is therefore not an effect that appears only at the last checkpoint.

C.4 Collateral Metrics and Pretrained Anchors

Clean-NLL provides a control for broad predictive degradation. Across the three Healthy/Mild/Severe models in Table 10, Clean-NLL does not show a severity-dependent deterioration that tracks the Terminal validity contrast.

The effect is also endpoint- and task-dependent. For Qwen2.5-3B Seed 2, Temporal margins are -0.344343, -0.340474, and -0.340667 for Healthy, Mild, and Severe, respectively, and therefore do not follow the Terminal Healthy $>$ Mild $>$ Severe ordering. For Ministral-3B, normalized ARC accuracy instead shows a consistent Healthy $>$ Mild $>$ Severe ordering across seeds (61.576%, 59.841%, and 58.817% when averaged across seeds). Together, these controls do not support a simple universal capability-degradation account of the DWL effect.

The pretrained checkpoint is a descriptive anchor rather than the matched causal control. For Qwen2.5-3B Seed 2, the pretrained Terminal margin is 0.664266, between Healthy (0.676630) and Severe (0.655736). For Qwen2.5-7B, the pretrained margin is 0.617118, whereas both Healthy (0.672806) and Severe (0.667789) are higher. Thus, the position of the pretrained checkpoint need not follow the Healthy–Severe ordering and does not alter the matched DWL estimand $Y_H - Y_S$.

Table 10: **Final Clean-NLL controls.** Lower is better.

Model	Healthy	Mild	Severe
Qwen2.5-3B	1.895453	1.895365	1.895088
Ministral-3B	2.796314	2.790725	2.787541
Qwen2.5-7B	1.813369	1.813071	1.812792

Table 11: **Architecture-relative regions used for coarse causal localization.**

Model	Layers	Later region	Middle control
Qwen2.5-3B	36	L23–L35	L12–L18
Qwen2.5-7B	28	L18–L27	L9–L14
Ministral-3B	26	L17–L25	L5–L9
LLaMA-3.2-1B	16	L10–L15	L5–L8

D Held-Out Causal Expression and Intervention Controls

D.1 Architecture-Relative Regions

The cross-model localization claim concerns architecture-relative coarse regions rather than a universal absolute layer index. The pre-specified later regions and middle controls used for each model are shown in Table 11.

D.2 Region-Level Causal Effects

Table 12 summarizes the coarse causal-expression geometry. For Qwen2.5-3B, the later prediction-span effect has 95% CI $[+0.020914, +0.024602]$. The corresponding intervals are $[+0.003974, +0.007460]$ for Qwen2.5-7B and $[+0.044086, +0.058911]$ for Ministral-3B. Across these three models, all six seed-by-direction later-span point estimates are positive.

LLaMA-3.2-1B provides a prospective replication in which the later region, middle control, and span-versus-boundary comparison were fixed before the corresponding outcomes were inspected. The three pre-specified contrasts are

$$\Delta_{\text{Late,Span}} = +0.020038, \qquad 95\%\ \text{CI} = [+0.017793, +0.022277], \tag{36}$$

$$\Delta_{\text{Late,Span}} - \Delta_{\text{Middle,Span}} = +0.007332, \qquad 95\%\ \text{CI} = [+0.005845, +0.008870], \tag{37}$$

$$\Delta_{\text{Late,Span}} - \Delta_{\text{Late,Boundary}} = +0.021943, \qquad 95\%\ \text{CI} = [+0.019895, +0.024013]. \tag{38}$$

All three contrasts survive Holm correction within the pre-specified family.

D.3 Prediction Span versus Boundary

The position contrast further localizes the coarse causal footprint. For Qwen2.5-3B, the Later-Span minus Later-Boundary effect is

$$+0.019117, \qquad 95\%\ \text{CI} = [+0.017450, +0.020817].$$

For Ministral-3B, the corresponding contrast is

$$+0.044140, \qquad 95\%\ \text{CI} = [+0.037799, +0.050788].$$

For Qwen2.5-7B, all six later-span effects are positive whereas all six corresponding Boundary effects are negative. LLaMA-3.2-1B prospectively reproduces the same Span $>$ Boundary ordering with a contrast of $+0.021943$.

We therefore describe the shared coarse footprint as *prediction-associated*: causal expression is stronger across the candidate prediction span than at the single pre-candidate Boundary state. This comparison does not imply that the tested interventions exhaust every possible positional matching scheme.

Table 12: **Held-out or prospectively pre-specified causal-expression geometry.** Values are source-condition-aligned changes in Terminal validity margin.

Model	Later Span	Middle Span	Later Boundary
Qwen2.5-3B	+0.022735	-0.001623	+0.003617
Qwen2.5-7B	+0.005702	-0.002708	-0.001369
Ministral-3B	+0.051451	≈ 0	+0.007311
LLaMA-3.2-1B	+0.020038	+0.012707	-0.001905

Table 13: **Same-condition activation-transfer background.** Absolute same-condition effects are compared with the corresponding cross-condition reference scale.

Model	Mean $\lvert\Delta_{\text{same}}\rvert$	Max $\lvert\Delta_{\text{same}}\rvert$	Cross-condition reference
Qwen2.5-3B	0.003145	0.008004	0.022735
Qwen2.5-7B	0.001756	0.003489	0.007756
Ministral-3B	0.008215	0.016480	0.059096

D.4 Same-Condition Activation-Transfer Controls

Activation transplantation between independently trained checkpoints can produce generic cross-checkpoint transfer effects even when the training condition is unchanged. We therefore evaluate directed H→H and S→S cross-seed transfers as a same-condition background control.

For all three models in Table 13, both the mean and maximum same-condition transfer magnitudes are smaller than the corresponding cross-condition reference. These controls do not assume zero interaction between independently trained checkpoints; rather, they establish the empirical background scale against which the Healthy–Severe transfer effect is interpreted.

E Causal Maturation and an Independent Validity-Related Direction

The coarse activation-transplantation results identify where the learned Healthy–Severe difference can causally affect behavior, but they do not specify how that difference becomes causally effective through depth. We therefore separate three properties that raw activation distance alone cannot distinguish: (i) the magnitude of the Healthy–Severe representational difference, (ii) its alignment with the local Terminal-validity decision direction, and (iii) its finite causal leverage under intervention.

We first characterize these quantities using the Terminal endpoint and then ask whether the resulting depth organization can be recovered by an independently constructed validity-related direction.

E.1 Decision-Geometry Definitions

For question x, let $h^H_{\ell,p}(x)$ and $h^S_{\ell,p}(x)$ denote the Healthy and Severe residual states at layer ℓ and prediction-associated position p. Let $\mathcal{P}_x$ denote the Prediction-span positions used by the causal intervention.

We define normalized representational discrepancy as

$$R_\ell(x) = \frac{1}{|\mathcal{P}_x|} \sum_{p \in \mathcal{P}_x} \frac{\left\| h^H_{\ell,p}(x) - h^S_{\ell,p}(x) \right\|_2}{\frac{1}{2}\left(\left\| h^H_{\ell,p}(x) \right\|_2 + \left\| h^S_{\ell,p}(x) \right\|_2 \right) + \epsilon}, \qquad R_\ell = \mathbb{E}_x[R_\ell(x)]. \tag{39}$$

Let Δh_ℓ denote the oriented Healthy–Severe residual displacement used in the decision-geometry analysis, and let

$$g_\ell = \nabla_{h_\ell} m_{\text{term}} \tag{40}$$

Table 14: **Qwen2.5-3B Formal600 causal-maturation replication.** Positive contrasts indicate stronger values at L27 except for $G_{21} - G_{27}$.

Contrast	Omnibus effect	95% CI	Replication
$C_{27} - C_{21}$	+0.0158185	[+0.0145974,+0.0170700]	3/3 seeds
$F_{27} - F_{21}$	+0.0147572	[+0.0132942,+0.0162716]	3/3 seeds
$P_{27} - P_{21}$	+0.1368742	[+0.1273946,+0.1462327]	3/3 positive
$D_{27} - D_{21}$	+0.0151866	[+0.0138408,+0.0165674]	3/3 positive
$G_{21} - G_{27}$	+0.1032395	[+0.0986079,+0.1080297]	600/600 questions

denote the local Terminal-validity gradient in the target model. We report

$$C_\ell = \cos(\Delta h_\ell, g_\ell), \tag{41}$$

$$P_\ell = \left\langle \Delta h_\ell, \frac{g_\ell}{\|g_\ell\|} \right\rangle, \tag{42}$$

$$D_\ell = \langle \Delta h_\ell, g_\ell \rangle, \tag{43}$$

$$G_\ell = \|g_\ell\|. \tag{44}$$

Here, R_ℓ measures the normalized magnitude of the Healthy–Severe representation difference, C_ℓ its directional alignment with the local Terminal decision, P_ℓ its signed projection onto the normalized decision direction, and D_ℓ the corresponding first-order decision-weighted displacement. G_ℓ records the local gradient norm.

Finally, F_ℓ denotes the finite bidirectional causal-transplant effect defined using the sign convention in Appendix B. Unlike C_ℓ, P_ℓ, D_ℓ, and G_ℓ, which are local quantities induced by the Terminal endpoint, F_ℓ is obtained from a finite intervention.

These quantities are not assumed to evolve monotonically or in parallel through depth. In particular, a larger R_ℓ need not imply a larger C_ℓ or F_ℓ. Nor do the endpoint-induced quantities define a global or model-independent "validity vector." The independent-direction analyses below are designed specifically to test whether the observed depth organization can be recovered without using the Terminal-margin gradient.

E.2 Qwen2.5-3B: Late Strengthening

Discovery300 indicates increasing residual discrepancy and causal leverage toward later depth, despite parameter displacement being distributed more broadly. The later third is not the most parameter-divergent region. In contrast, activation relative-L_2 increases from 0.010546 in the early region to 0.018607 in the late region and peaks at L26 (0.022369). Within L18–L27, finite causal leverage becomes stably positive around L21 and approaches the full matched Healthy–Severe behavioral gap by L25–L27.

Based on this Discovery analysis, the L21–L27 depth contrast was fixed before evaluation on Formal600 across all three training seeds. Table 14 reports the resulting contrasts.

The primary C and F contrasts replicate across all three seeds, reach the 10^{-5} Monte Carlo resolution, and survive Holm correction. At the same time, gradient norm decreases from L21 to L27 while alignment and finite causal leverage increase.

We therefore describe the Qwen2.5-3B trajectory as *late strengthening*: the learned difference becomes more aligned with the local decision and more causally effective toward later depth. We do not attribute this pattern to increasing gradient magnitude.

E.3 Ministral-3B: Directional Refinement

Ministral-3B provides a direct separation between representational magnitude and causal leverage. Parameter divergence is strongest around L13–L15. Normalized residual discrepancy peaks at L15 (relative-$L_2 = 0.104235$), but finite causal leverage continues to increase toward L19, where it reaches $+0.06696$, while relative-L_2 decreases to 0.087362.

Table 15: **Ministral-3B Formal600 directional-refinement replication.**

Contrast	Omnibus effect	95% CI	Status
$C_{19} - C_{15}$	+0.006507906	[+0.005362275,+0.007693200]	primary, 3/3
$F_{19} - F_{15}$	+0.016947768	[+0.013534994,+0.020474223]	primary, 3/3
$D_{19} - D_{15}$	+0.016381919	[+0.013065210,+0.019840087]	secondary
$P_{19} - P_{15}$	+0.021390351	[+0.018061406,+0.024799412]	secondary
$R_{15} - R_{19}$	+0.017738268	[+0.017270999,+0.018199242]	magnitude decreases
$G_{15} - G_{19}$	+0.725148612	[+0.695323357,+0.755166367]	gradient norm decreases

Table 16: **LLaMA-3.2-1B Formal600 transient-bottleneck replication.**

Contrast	Omnibus effect	95% CI	Role
$C_{10} - C_{15}$	+0.010333779	[+0.009512045,+0.011162512]	primary, 3/3
$F_{10} - F_{15}$	+0.005496543	[+0.004347883,+0.006641205]	primary, 3/3
$D_{10} - D_{15}$	+0.005341028	[+0.004347897,+0.006331590]	secondary
$G_{10} - G_{15}$	+0.473780191	[+0.452631550,+0.495240179]	secondary
$R_{10} - R_{15}$	+0.001022243	[+0.000841483,+0.001200110]	secondary
$P_{10} - P_{15}$	+0.000580957	[-0.000624730,+0.001779791]	descriptive

Following Discovery, the L15–L19 comparison was fixed before the newly defined decision-geometry quantities were evaluated on Formal600. The results are shown in Table 15.

Thus, from L15 to L19, both normalized representational discrepancy and gradient norm decrease, while decision alignment and finite causal leverage increase. This is the operational meaning of *directional refinement* in this paper: the Healthy–Severe difference becomes smaller in normalized magnitude while becoming more strongly organized along a behaviorally relevant direction.

E.4 LLaMA-3.2-1B: Front-Loaded Transient Bottleneck

LLaMA-3.2-1B shows that the direction of the within-region maturation trajectory is itself model dependent. Discovery300 over L5–L15 places the normalized representation peak at L7, the finite causal peak near L9, and the local decision-alignment peak near L10, followed by attenuation toward L15.

The L10–L15 comparison was fixed before the corresponding single-layer decision-geometry quantities were evaluated on Formal600. Table 16 summarizes the result.

The primary C and F contrasts replicate in all three seeds and survive Holm correction. The appropriate claim is therefore not that causal leverage must increase toward the final layer, but that LLaMA-3.2-1B exhibits a *front-loaded transient bottleneck*: causal expression is stronger around L10 and subsequently dissipates toward L15.

Taken together, Sections E.2–E.4 show three distinct maturation profiles: Qwen2.5-3B exhibits late strengthening, Ministral-3B directional refinement, and LLaMA-3.2-1B a front-loaded bottleneck. The common coarse causal footprint therefore does not imply a common depth-wise developmental trajectory.

E.5 Independent Validity Direction: Unified Construction

The decision-geometry analysis above is partly induced by the Terminal endpoint because C_ℓ, P_ℓ, D_ℓ, and G_ℓ depend on the Terminal-margin gradient. We therefore test whether the model-specific depth ordering can be recovered using a direction constructed without either the Terminal-margin gradient or Healthy/Severe condition labels.

For each model, using Discovery300 only and a pre-specified construction layer ℓ^*, we construct

$$v_{\text{valid}} \propto \mathbb{E}_{\text{state},a}\left[\text{centroid}\left(h^{\ell^*}_{\text{valid}}\right) - \text{mean}_j\,\text{centroid}\left(h^{\ell^*}_{\text{invalid},j}\right)\right]. \tag{45}$$

Here $a \in \{H, S\}$ indexes the Healthy and Severe training conditions, which contribute representations symmetrically, and j indexes invalid candidate alternatives. Healthy/Severe identity is never

Table 17: **Pre-specified independent validity-direction protocol.** The construction procedure is shared across models, while the A2 prediction follows each model's previously established causal-maturation profile.

Model	Direction layer	A2 depth prediction	A3 layer
Qwen2.5-3B	L27	$V_{27} > V_{21}$	L27
Ministral-3B	L19	$V_{19} > V_{15}$	L19
LLaMA-3.2-1B	L10	$V_{10} > V_{15}$	L10

used as a class label. The only supervision used to construct v_{valid} is candidate valid-versus-invalid identity.

The construction layer is selected from the previously established causal-maturation profile before the independent-direction endpoint is evaluated: L27 for Qwen2.5-3B, L19 for Ministral-3B, and L10 for LLaMA-3.2-1B. The resulting vector is unit-normalized and frozen before all subsequent evaluation.

For clarity, we use three stages:

1. **A1 – held-out readability:** test whether the frozen direction separates valid from invalid candidates on Reserve600;
2. **A2 – frozen depth prediction:** carry the same frozen direction across the pre-specified depth pair and test whether the Healthy–Severe validity-aligned component follows the model-specific causal-maturation ordering;
3. **A3 – causal steering:** intervene along the frozen direction on Formal600 and compare its effect with a fixed random orthogonal control.

We denote by V_ℓ the Healthy–Severe validity-aligned component obtained by projecting the relevant states onto the frozen v_{valid} direction at layer ℓ. The A2 comparison is therefore fixed by the earlier maturation result rather than chosen after observing the independent-direction measurements.

E.6 Qwen2.5-3B: Independent Validity Strengthens Toward Late Depth

A1: Held-out readability. The L27 direction constructed on Discovery300 separates valid from invalid candidates on Reserve600:

$$\text{Sep}_{\text{real}} = +10.0763, \qquad 95\%\ \text{CI} = [+9.5952, +10.5654]. \tag{46}$$

Relative to the shuffled-direction control,

$$\text{Sep}_{\text{real}} - \text{Sep}_{\text{shuffle}} = +12.6194, \qquad 95\%\ \text{CI} = [+12.2464, +12.9930]. \tag{47}$$

The absolute scale of this separation is architecture dependent and is not compared numerically across models.

A2: Frozen depth prediction. Using the same frozen L27 direction at both L21 and L27,

$$V_{27} - V_{21} = +0.07614, \qquad 95\%\ \text{CI} = [+0.07080, +0.08154]. \tag{48}$$

The contrast is positive for 550/600 questions and in all three training seeds. The seed-level effects are $+0.06982$, $+0.07420$, and $+0.08439$. Thus the independently constructed direction recovers the previously fixed Qwen ordering $C_{27} > C_{21}$ and $F_{27} > F_{21}$.

A3: Causal steering. On Formal600, steering Severe checkpoints along the frozen L27 direction gives

$$\beta_{\text{real}} = +0.006870, \qquad 95\%\ \text{CI} = [+0.006516, +0.007228], \tag{49}$$

and

$$\beta_{\text{real}} - \beta_{\text{orth}} = +0.006674, \qquad 95\%\ \text{CI} = [+0.006306, +0.007056]. \tag{50}$$

The endpoint contrast between $+\alpha_0$ and $-\alpha_0$ is

$$+0.13211, \qquad 95\%\ \text{CI} = [+0.12545, +0.13886]. \tag{51}$$

The orthogonal slope is small but slightly positive:

$$\beta_{\text{orth}} = +0.000196, \qquad 95\%\ \text{CI} = [+0.000002, +0.000387]. \tag{52}$$

Accordingly, the supported claim is that steering along the real direction is substantially more effective than steering along the fixed orthogonal control; we do not claim that the orthogonal effect is indistinguishable from zero.

E.7 MINISTRAL-3B: INDEPENDENT VALIDITY UNDERGOES DIRECTIONAL REFINEMENT

A1: Held-out readability. The frozen L19 direction separates valid from invalid candidates on Reserve600:

$$\text{Sep}_{\text{real}} = +0.979311343, \qquad 95\%\ \text{CI} = [+0.948135800, +1.010388247]. \tag{53}$$

The shuffled-direction control requires a qualification. Before held-out Reserve projection statistics were computed, we identified an error in the initial pseudo-valid construction: pseudo-valid candidates had been sampled only from genuinely invalid candidates, making that realization definitionally anti-valid. Only the shuffled control was corrected; the real validity direction was unchanged.

The corrected shuffled direction remains substantially anti-aligned with the real direction (cosine $= -0.4634$). We therefore treat the real held-out separation as the primary A1 result and use the random orthogonal direction as the primary A3 null control.

A2: Frozen depth prediction. Using the same frozen L19 direction at L15 and L19,

$$V_{19} - V_{15} = +0.016746683, \qquad 95\%\ \text{CI} = [+0.015382843, +0.018118618]. \tag{54}$$

Over the same transition,

$$R_{15} - R_{19} = +0.017996022, \qquad 95\%\ \text{CI} = [+0.017491922, +0.018493219]. \tag{55}$$

Thus the normalized Healthy–Severe representational difference becomes smaller from L15 to L19 while its component along the independently defined validity-related direction becomes larger. The independent-direction analysis therefore reproduces the directional refinement pattern without using the Terminal-margin gradient to construct the direction.

A3: Causal steering. On Formal600,

$$\beta_{\text{real}} = +0.022308048, \qquad 95\%\ \text{CI} = [+0.019834077, +0.024774026], \tag{56}$$

while

$$\beta_{\text{real}} - \beta_{\text{orth}} = +0.021840617, \qquad 95\%\ \text{CI} = [+0.018571176, +0.025048951]. \tag{57}$$

The orthogonal slope is

$$\beta_{\text{orth}} = +0.000467432, \qquad 95\%\ \text{CI} = [-0.001259900, +0.002209228], \tag{58}$$

which is consistent with zero.

E.8 LLAMA-3.2-1B: INDEPENDENT VALIDITY PEAKS AT THE FRONT-LOADED BOTTLENECK

A1: Held-out readability. The L10 direction constructed on Discovery300 separates valid from invalid candidates on Reserve600:

$$\text{Sep}_{\text{real}} = +0.70554, \qquad 95\%\ \text{CI} = [+0.67658, +0.73526]. \tag{59}$$

The real-minus-shuffled contrast is

$$+0.88790, \qquad 95\%\ \text{CI} = [+0.85977, +0.91616]. \tag{60}$$

A2: Frozen depth prediction. The previously established LLaMA maturation profile predicts stronger validity-aligned structure at L10 than at L15:

$$V_{10} > V_{15}. \tag{61}$$

Reserve600 confirms

$$V_{10} - V_{15} = +0.002866, \qquad 95\%\ \text{CI} = [+0.002264, +0.003464]. \tag{62}$$

The contrast is positive for 379/600 questions and in all three training seeds, with seed-level effects $+0.003339$, $+0.002387$, and $+0.002873$.

This comparison is informative because the earlier Terminal-gradient-induced contrast $P_{10} - P_{15}$ is not significant. The independently constructed candidate-identity direction therefore provides a distinct measurement that agrees with the front-loaded bottleneck identified by C_ℓ and F_ℓ.

A3: Causal steering. On Formal600,

$$\beta_{\text{real}} = +0.04006, \qquad 95\%\ \text{CI} = [+0.03677, +0.04343], \tag{63}$$

and

$$\beta_{\text{real}} - \beta_{\text{orth}} = +0.03882, \qquad 95\%\ \text{CI} = [+0.03452, +0.04309]. \tag{64}$$

The endpoint contrast between $+\alpha_0$ and $-\alpha_0$ is

$$+0.05379, \qquad 95\%\ \text{CI} = [+0.04935, +0.05830]. \tag{65}$$

The orthogonal slope is

$$\beta_{\text{orth}} = +0.00124, \qquad 95\%\ \text{CI} = [-0.00125, +0.00378], \tag{66}$$

which is consistent with zero.

E.9 Post-Hoc Construction-Anchor Robustness

The primary A2 analyses construct v_{valid} at the layer selected from the previously established causal-maturation profile. A possible alternative explanation is therefore that the measured depth ordering is favored mechanically by the layer at which the direction was constructed.

We test this possibility with a post-hoc crossed-construction control. For each model, we reconstruct the validity-related direction at the opposite member of the already fixed A2 depth pair and repeat the same depth comparison. We do not scan intermediate layers or alter the predicted ordering. These analyses are robustness checks and do not replace the original pre-specified A2 results.

Qwen2.5-3B. The primary A2 direction is constructed at L27 and gives $V_{27} - V_{21} = +0.07614$. When the direction is instead constructed at L21, the predicted ordering remains:

$$V_{27} - V_{21} = +0.036395, \qquad 95\%\ \text{CI} = [+0.033096, +0.039732]. \tag{67}$$

The contrast is positive in all three training seeds. The reduction in magnitude shows that the construction anchor affects effect size, but the Qwen depth ordering does not depend on constructing the direction at L27.

Ministral-3B. The primary direction is constructed at L19. Under the crossed control, constructing the direction at L15 still gives

$$V_{19} - V_{15} = +0.013845, \qquad 95\%\ \text{CI} = [+0.012898, +0.014791]. \tag{68}$$

The three seed-level effects are $+0.01088$, $+0.01667$, and $+0.01398$. Thus the increase in validity-aligned structure from L15 to L19 cannot be explained solely by constructing the primary direction at L19.

Table 18: **Post-hoc construction-anchor robustness of the independent validity-direction depth ordering.** The original A2 analyses remain the primary pre-specified results.

Model	Anchor reversal	Crossed depth contrast	Outcome
Qwen2.5-3B	L27 → L21	$V_{27} - V_{21} = +0.03640$	3/3 seeds positive
Ministral-3B	L19 → L15	$V_{19} - V_{15} = +0.01384$	3/3 seeds positive
LLaMA-3.2-1B	L10 → L15	$V_{10} - V_{15} = +0.00129$	8/12 seeds positive; mixed

95% CIs are $[+0.03310, +0.03973]$, $[+0.01290, +0.01479]$, and $[-0.00052, +0.00295]$ for Qwen2.5-3B, Ministral-3B, and LLaMA-3.2-1B, respectively.

Table 19: **Cross-model independent validity-direction evidence.** A1 tests held-out readability, A2 the pre-specified model-specific depth prediction, and A3 causal steering relative to a fixed orthogonal control. Absolute effect scales are not compared across architectures.

Model	A1	A2 depth result	A3 real − orthogonal
Qwen2.5-3B	PASS	$V_{27} - V_{21} = +0.07614$	+0.006674
Ministral-3B	PASS	$V_{19} - V_{15} = +0.01675$	+0.021841
LLaMA-3.2-1B	PASS	$V_{10} - V_{15} = +0.002866$	+0.03882

LLaMA-3.2-1B. LLaMA is less stable under construction-anchor reversal. The initial three-seed crossed analysis remains positive on average but does not satisfy the pre-specified robustness criterion. We therefore report a separate post-hoc 12-seed training-randomness stability analysis using the same L15 construction and L10–L15 comparison.

Across 12 independently trained Healthy–Severe seed pairs,

$$V_{10} - V_{15} = +0.001294, \tag{69}$$

with seed-level standard deviation 0.003170 and

$$95\%\ \mathrm{CI}_{\mathrm{seed}} = [-0.000518, +0.002949]. \tag{70}$$

Eight of twelve seeds have the predicted sign, and the exact one-sided 2^{12} sign-flip test gives $p = 0.0913$. Under the fixed stability criterion, this result is therefore **MIXED**, rather than a robust positive replication.

Taken together, the crossed-construction controls weaken a simple "construction layer wins" explanation. The predicted depth ordering survives construction-anchor reversal in Qwen2.5-3B and Ministral-3B. For LLaMA-3.2-1B, however, the crossed result is substantially more sensitive to both the construction anchor and training randomness; we therefore retain its post-hoc robustness status as mixed.

E.10 Cross-Model Independent-Direction Synthesis

Table 19 summarizes the primary independent-direction evidence. Across all three models, A1 establishes held-out readability and A3 shows that intervention along the real direction has a larger causal effect than intervention along the fixed orthogonal control. A2 additionally recovers the model-specific depth ordering established by the earlier causal-maturation analysis.

The cross-model result is not that Qwen2.5-3B, Ministral-3B, and LLaMA-3.2-1B share a single numerical validity vector. They do not share the same depth trajectory either. Rather, the same construction-and-test procedure identifies held-out-readable and causally manipulable validity-related structure in all three models, while the organization of that structure through depth follows the model-specific maturation profile established independently beforehand.

Qwen2.5-3B exhibits late strengthening. Ministral-3B exhibits directional refinement. LLaMA-3.2-1B exhibits a front-loaded validity bottleneck followed by dissipation. The resulting cross-model conclusion is therefore

validity-related structure is procedurally replicable, while its depth organization is model dependent. (71)

Table 20: **Held-out module decomposition.** Positive MLP–Attention values favor MLP.

Model	Δ_{Attn}	Δ_{MLP}	MLP–Attn	Evidence
Qwen2.5-3B	+0.004462	+0.002763	-0.001699	Unexpected Attention-favoring reversal
Qwen2.5-7B	+0.000812	+0.002273	+0.001462	Aggregate MLP-favoring
Ministral-3B	–	–	+0.001917	Seed-dependent crossover; CI includes zero

We do not claim that the three models share the same numerical direction, that validity is uniquely or completely represented by a single dimension, or that Transformer architectures obey a universal depth law.

F Fine-Grained Module Allocation

The architecture-relative coarse causal footprint is more stable than its fine-grained allocation across Attention and MLP modules. We therefore treat module-level decomposition as a finer-grained mechanistic question rather than as part of the shared cross-model claim.

For Qwen2.5-3B, the pre-specified prediction favored MLP over Attention. The held-out result reverses this prediction: all six point estimates favor Attention, and the omnibus MLP–Attention contrast has 95% CI $[-0.002185, -0.001207]$. We therefore retain this result as an *unexpected held-out reversal*, not as a prospectively predicted Attention effect.

Qwen2.5-7B shows the opposite organization. Its MLP–Attention contrast is

$$+0.001462, \qquad 95\%\ \mathrm{CI} = [+0.000756, +0.002164].$$

The aggregate effect favors MLP, although the stricter condition-wise gate is satisfied in five of six conditions.

Ministral-3B is less stable at the module level. Seed 1 is Attention-favoring (-0.005569), whereas Seeds 2–3 are MLP-favoring ($+0.006214$ and $+0.005107$). The omnibus MLP-dominance contrast is

$$+0.001917, \qquad 95\%\ \mathrm{CI} = [-0.001033, +0.004828],$$

and therefore does not establish a consistent module-level preference.

The conclusion is therefore deliberately coarse: the architecture-relative causal footprint is more reproducible across models than the identity of the module carrying the strongest local effect. We do not infer a universal Attention- or MLP-specific locus for DWL.

G Counterfactual Parameter Replaceability and Compatibility Controls

Activation transplantation and parameter replacement answer different causal questions. Activation transplantation asks whether replacing an internal representation is sufficient to transfer the learned Healthy–Severe behavioral difference. Parameter replacement instead asks whether parameters from the Healthy checkpoint can be inserted into the matched Severe checkpoint at a selected location and improve the Terminal validity margin.

For parameter set $\mathcal{P}$, define the hybrid checkpoint

$$\theta_j^{S\leftarrow H} = \begin{cases} \theta_j^H, & j \in \mathcal{P}, \\ \theta_j^S, & j \notin \mathcal{P}. \end{cases} \tag{72}$$

The paired replacement gain is

$$\Delta_{\mathrm{replace}}(x) = m_{\mathrm{term}}\left(x; \theta^{S\leftarrow H}\right) - m_{\mathrm{term}}\left(x; \theta^S\right). \tag{73}$$

For descriptive purposes, we occasionally report the recovery ratio

$$r_{\mathrm{recovery}} = \frac{m_R - m_S}{m_H - m_S}, \tag{74}$$

where m_R denotes the margin of the replacement checkpoint. All statistical inference, however, is based on the absolute paired replacement gain rather than this ratio.

Table 21: **Causal-expression and parameter-replacement geometry.** The table contrasts the coarse activation-level causal footprint with the strongest tested parameter-replacement result.

Model	Coarse causal expression	Best tested replacement	Interpretation
Qwen2.5-3B	L23–L35 later span	L23–L25 MLP, $+0.002671$	Partial co-location; coverage dependent
Ministral-3B	L17–L25 later span	L6–L8 MLP, $+0.007201$	Strong upstream dissociation
LLaMA-3.2-1B	L10–L15 later region	L5–L8 MLP, $+0.014177$	Late prediction fails; upstream replacement is productive
Qwen2.5-7B	L18–L27 later span	No robust positive local replacement	Broad late replacement is negative

G.1 Cross-Model Geometry

The central result is that the location at which the DWL contrast is *causally expressed* need not coincide with the location at which Healthy parameters are *productively replaceable*. Table 21 summarizes the strongest tested replacement result for each model.

G.2 Qwen2.5-3B: Partial Late Replaceability Depends on Coverage

For Qwen2.5-3B, replacing the full L23–L25 MLP block gives

$$\Delta_{\text{replace}} = +0.002671, \qquad 95\%\ \text{CI} = [+0.001801, +0.003573],$$

corresponding descriptively to 11.17% recovery. The corresponding Attention replacement gives

$$+0.001450, \qquad 95\%\ \text{CI} = [+0.000580, +0.002329].$$

An equal-budget sparse MLP replacement containing 28.32M parameters yields only $+0.000124$, with a confidence interval crossing zero. The benefit of the full L23–L25 MLP replacement therefore should not be interpreted as evidence that an individual MLP parameter is intrinsically more effective.

A post-hoc nested-coverage analysis further shows increasing replacement gain as L23–L25 MLP coverage increases from 13.96% to 100%:

$$+0.000124,\ +0.001035,\ +0.001541,\ +0.002281,\ +0.002671.$$

The question-level coverage slope is

$$+0.002771, \qquad 95\%\ \text{CI} = [+0.001891, +0.003634].$$

Because this analysis was motivated after observing the initial replacement result, we treat it as a post-hoc coverage-response analysis and do not claim that every adjacent coverage increment is independently significant.

G.3 Ministral-3B: Late Expression, Upstream Replaceability

Ministral-3B shows the clearest separation between causal-expression depth and productive replacement depth. The Formal screen gives negative replacement effects in the late/core region: Core Attention is -0.001504 and Core MLP is -0.001723. In contrast, the upstream L6–L8 region is positive: Attention gives $+0.005376$ and MLP gives $+0.007264$.

After the upstream location hypothesis was fixed, Reserve600 confirms L6–L8 MLP replacement:

$$\Delta_{\text{replace}} = +0.007201, \qquad 95\%\ \text{CI} = [+0.006248, +0.008165],$$

corresponding descriptively to 13.8% recovery. The pre-specified upstream-minus-core MLP contrast is

$$+0.008988, \qquad 95\%\ \text{CI} = [+0.008000, +0.010002].$$

Reserve600 had previously been used for an activation-module endpoint, but not for selecting the parameter-replacement location. We therefore describe this result as cross-packet replacement-location confirmation rather than as a globally untouched evaluation.

Table 22: **Same-condition parameter-replacement background compared with the corresponding H→S effect.** Ratios are descriptive.

Model / scope	N	Mean signed	Mean $\|\Delta\|$	Max $\|\Delta\|$	H→S effect
Qwen2.5-3B L23–25 MLP	12	-0.000119	0.000677	0.001345	+0.002671
Ministral-3B L6–8 MLP	12	-0.000049	0.001134	0.001950	+0.007201
LLaMA-3.2-1B L5–8 MLP	12	-0.000043	0.001320	0.002868	+0.014177
Qwen2.5-7B L18–27 MLP	12	-0.000377	0.000411	0.001215	-0.002356

G.4 LLaMA-3.2-1B: Failed Late Prediction and Productive Upstream Control

The pre-specified late-MLP replacement hypothesis fails:

$$\Delta_{\text{Late MLP}} = -0.002511, \qquad 95\%\ \text{CI} = [-0.003327, -0.001697]. \tag{75}$$

The pre-specified Late-MLP minus Middle-MLP comparison is also negative (-0.016688).

Unexpectedly, the fixed middle-depth control at L5–L8 produces a large positive MLP replacement effect:

$$+0.014177, \qquad 95\%\ \text{CI} = [+0.012993, +0.015376],$$

corresponding descriptively to 81.36% recovery.

Because a positive middle-region replacement effect was not the prospective prediction, we retain it as secondary exploratory evidence even though the middle control region itself was specified before its outcome was inspected. The result therefore supports a dissociation between where the learned difference is expressed and where parameter substitution can productively alter the Severe checkpoint.

G.5 Qwen2.5-7B: No Robust Local Rescue Under the Tested Operators

Qwen2.5-7B provides the negative parameter-replacement case. Neither local core nor middle replacement yields a robust positive effect. Moreover, increasing late-MLP coverage makes the outcome more negative:

$$\beta_{\text{coverage}} = -0.003239, \qquad 95\%\ \text{CI} = [-0.004356, -0.002114]. \tag{76}$$

Full L18–L27 MLP replacement is worse than the L18–L19 core replacement by

$$-0.002700, \qquad 95\%\ \text{CI} = [-0.003710, -0.001702].$$

The three upstream L6–L8 MLP effects are also negative, and soft interpolation at $\alpha \in \{0.25, 0.50, 0.75\}$ averages -0.000226, with a confidence interval crossing zero.

Full-late H→S replacement is negative in all three seeds: -0.002078, -0.002359, and -0.002630, with mean -0.002356. We therefore report *no robust local rescue under the tested parameter-replacement operators*. This is a statement about the tested interventions, not a claim of fundamental irreparability.

G.6 Same-Condition Parameter-Compatibility Controls

A parameter transplant between independently trained checkpoints can itself introduce generic cross-trajectory compatibility effects. We therefore repeat the same parameter scope and replacement operator within ordered H→H and S→S cross-seed pairs.

For Qwen2.5-3B, same-condition effects have mixed signs (5 positive and 7 negative), and the confirmed H→S replacement effect is 1.99× the largest same-condition absolute effect. The corresponding ratios are 3.69× for Ministral-3B and 4.94× for LLaMA-3.2-1B; the LLaMA same-condition controls are also sign-balanced (6/6).

Qwen2.5-7B is particularly useful because the cross-condition effect is negative. Its H→H full-late swaps have mean signed effect -0.000140 and mean absolute effect 0.000179, while S→S swaps

have mean signed effect -0.000614 and mean absolute effect 0.000643. The pooled mean absolute same-condition effect is 0.000411 and the largest absolute control effect is 0.001215, both smaller than the mean H$\rightarrow$S effect in absolute magnitude.

These controls show that generic cross-seed compatibility effects are nonzero but smaller than the principal cross-condition replacement effects. They do not establish that every replacement effect is purely condition-specific.

Taken together, the parameter experiments support a distinction between *causal expression* and *parameter replaceability*: a representation may causally carry the DWL contrast at one depth even when productive parameter substitution occurs elsewhere, or fails under the tested local replacement operators.