RATIONAL MANIFOLDS: THE SPATIAL ORGANIZATION AND CONTROL OF LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

The thinking traces of large language models (LLMs) interleave behaviors such as planning, exploring and verifying, yet little is known about how these behaviors are arranged in latent space. Linear representations provide a framework to explain and control model behavior but the directions it relies on are often unstable. We provide a mechanistic recount of LLM reasoning by interpreting regions instead of directions as the semantic units of latent space, partitioning activations via a Mixture of Factor Analyzers (MFA). Our experiments reveal that these regions correspond to recognizable problem-solving behaviors with complex nonlinear boundaries and whose local concept directions consistently disagree with global ones. Causal interventions show that regions offer limited control over individual behaviors but can have a strong impact on reasoning performance: diverting traces away from a small subset of load-bearing regions critically degrades accuracy, but carefully redirecting the path a model takes through specific regions can also make reasoning shorter and more accurate. This mechanism is found to differ geometrically from linear steering. Together, these findings suggest that localized descriptors, such as latent space regions and the paths through them, can inform new methods for both interpretation and control of reasoning. [1]

1 INTRODUCTION

Let large language models (LLMs) explore. Much of their recent improvements comes from allowing them to generate intermediate "thinking" steps—planning, revisiting earlier decisions, and checking provisional answers—before committing to an answer (Jaech et al., 2024; DeepSeek-AI, 2025; Yang et al., 2025). Exploring, however, is both *expensive* and *error-prone*: models can spend an exorbitant number of tokens pursuing unproductive paths or repeating themselves (Chen et al., 2025b; Sui et al., 2025), while hallucinations can compound across steps (Zhang et al., 2024). Although these reasoning traces read like step-by-step accounts of thought, the text alone reveals little about the internal mechanisms governing them. Our understanding of these mechanisms remains limited, which restricts our ability to detect and correct reasoning failures. In this paper, we study reasoning from the latent space perspective, asking how reasoning trajectories can be *interpreted*, and to what extent they can be *controlled*.

The Linear Representation Hypothesis (LRH) is the current prevailing paradigm for how LLM representations are structured. It posits that concepts are encoded as directions in latent space (Park et al., 2024; 2025), successfully explaining atomic concepts such as numbers, years and spatial coordinates (Heinzerling & Inui, 2024; Gurnee & Tegmark, 2024). This view also suggests a simple control mechanism: steering activations linearly along a direction during generation can strengthen or suppress the corresponding behavior. Such interventions have also been used to steer complex behaviors such as sycophancy, hallucination, refusal and corrigibility (Rimsky et al., 2024).

In practice, recovered directions are often unstable: different prompts can produce conflicting directions (Braun et al., 2025; Tan et al., 2024); small input perturbations can reduce their efficacy (Haller et al., 2026); and locally effective directions may become misaligned to the global one, weakening or even reversing its effect (Raval et al., 2026; Li et al., 2026a). Several independent works have set out to refine the hypothesis accordingly, proposing alternative definitions for concepts either as

[1] Code, data and results will be publicly available.

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.

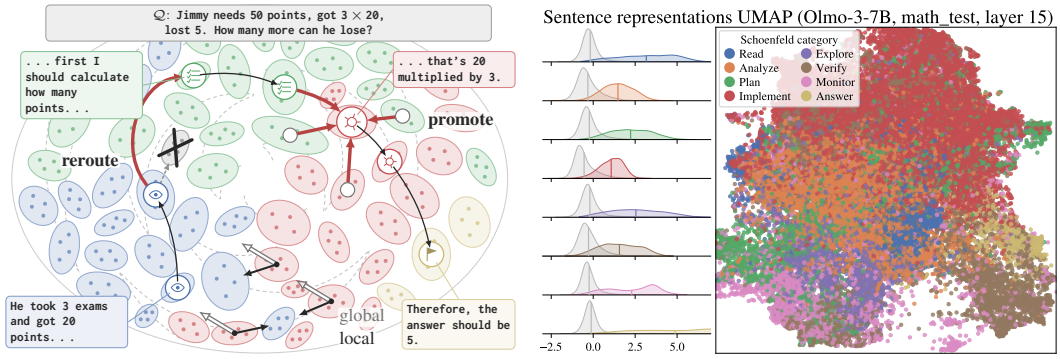


Figure 1: **Left:** working hypothesis of this paper and proposed interventions. The latent space is partitioned into functional regions traversed during reasoning, and local mean-difference directions between functional types disagree with the global ones. We intervene by *rerouting*, blocking a region and redistributing its traffic, or *promoting*, pulling traffic into a region. **Right:** sentences of the same episode category occupy preferred areas of the latent space without being linearly separable. Example of latent space structure in sentence representations and their distribution projected along LDA direction, with color representing Schoenfeld episode category.

convex polytopes, weighted Voronoi cells, ellipsoids, or Gaussian components (Fel et al., 2026; Ma & Wolfinger, 2026; Postmus & Abreu, 2024; Shafran et al., 2026). These approaches all define a concept to be a bounded region of the space rather than a globally valid direction. We argue this region-based view of the latent space provides a promising description of complex, nonlinear behaviors. Reasoning is a natural candidate on which to test this: a reasoning trace is a highly structured object in latent space, composed of several functional units, whose path has a complex geometry that reflects how the reasoning unfolds (Minegishi et al., 2025; Zhou et al., 2026). We therefore set out to investigate whether regions can serve as semantic units of reasoning, and whether such a description can inform fine-grained control methodologies.

We operationalize this view by fitting a Mixture of Factor Analyzers (MFA; Shafran et al., 2026) on sentence-level representations of thinking traces so that its components partition the space into coherent regions. This unsupervised process recovers regions that we interpret as nodes of a directed graph which thinking traces traverse during reasoning (Figure 1). Our analysis covers eight reasoning models from five families, ranging from 1.5B to 32B parameters, on three datasets covering both mathematical and general reasoning (§3.2).

First, we ask whether there is a correspondence between thinking steps and the structure of their underlying activations. We annotate thinking traces following the Schoenfeld taxonomy of human problem-solving (Schoenfeld, 1985; Li et al., 2026b), so called *episode categories*. We find that different behaviors occupy preferred areas of the latent space (Figure 1). While **global directions separate behaviours only partially, region-specific directions separate them more reliably across every model and dataset we consider**. This suggests that a single set of global directions, as commonly assumed in interpretability work, is an insufficient description of behavioral concepts. Second, we test whether these regions play a causal role in reasoning using intervention. **Only some regions turn out to be causally active, and only one episode category can be induced this way**, which we interpret as a recovery mechanism of the model. Finally, we investigate what the traversal of specific regions contributes to the final output, and whether this trajectory can be redirected towards better reasoning. This is achieved by *rerouting* reasoning away from specific regions or *promoting* passage through them. Reasoning trajectories are first observed to be brittle: **preventing the model from accessing a small set of load-bearing regions collapses accuracy**. Directed at the right regions, **the same interventions can improve both efficiency and accuracy.** This mechanism is found to differ geometrically from activation addition.

To the best of our knowledge, our work is the first region-based analysis of reasoning traces and their functional role, and the first demonstration that interventions on these regions can improve reasoning efficiency and accuracy. Together, our findings suggest there are significant gaps in our understanding of LLM reasoning, gaps that can be filled by novel localized descriptors.

2 Background and Related Work

Our work draws on two lines of interpretability research: one studying how reasoning unfolds in the latent space of a model, and one studying how concepts are geometrically represented within that space. We first outline how the two together motivate a region-based analysis of reasoning (§2.1), and then describe the Mixture of Factor Analyzers, the method we use to partition the latent space into regions (§2.2). A more detailed review of the literature is provided in Appendix A.

2.1 Interpreting Reasoning

Recent interpretability work has studied the activations associated with LLM chains-of-thought as geometric objects. Some have managed to relate properties of the resulting generation such as correctness, length and logical structure to measurable properties of the trajectory itself such as velocity, curvature, and displacement (Jiang et al., 2026; Vilas et al., 2026; Zhou et al., 2026). Others have decoded the topology of this underlying structure, assigning functional roles to each part (Minegishi et al., 2025; Yu et al., 2025).

Other works in activation steering take on controllability of the traces. Steering vectors have since been used to control reasoning, both by eliciting or suppressing behaviors such as reflection and backtracking (Venhoff et al., 2025; Chang et al., 2025) and by shortening traces without degrading accuracy (Chen et al., 2025a; Huang et al., 2025; Eisenstadt et al., 2025). This use of steering vectors relies on the Linear Representation Hypothesis (LRH), which posits that language models encode concepts as directions in their latent space (Park et al., 2024; Modell et al., 2025), and which is well supported for atomic concepts (Marks & Tegmark, 2024; Gurnee & Tegmark, 2024; Heinzerling & Inui, 2024). Initial evidence, however, suggests that a single global direction often gives an incomplete description of a concept. Some concepts have multi-dimensional or nonlinear structure that can only be recovered jointly from several directions (Engels et al., 2025; Tan et al., 2024). Directions estimated from one set of inputs also transfer poorly, proving brittle under superficial changes to the input and inconsistent across the prompts used to construct them (Haller et al., 2026; Braun et al., 2025). Both observations are consistent with a concept structure that varies across space, in which case a concept may be better described by the region it occupies and its local directions of variation.

Several recent works adopt this view, modeling concepts as convex polytopes, weighted Voronoi cells or ellipsoids (Fel et al., 2026; Ma & Wolfinger, 2026; Postmus & Abreu, 2024). Relevant for this work, Shafran et al. (2026) fit a mixture model to the activation distribution whose components each cover a region of the space. Reasoning traces are a natural object of study for this region-based view: over the course of a single trace, activations repeatedly cross multiple regions of the latent space (Minegishi et al., 2025), and their associated reasoning steps take on distinct functional roles (Yu et al., 2025). To the best of our knowledge, no previous work has studied these regions as monosemantic, functional objects during reasoning, nor explored their causal role in this domain.

2.2 Mixture of Factor Analyzers

Among these region-based approaches, we adopt the Mixture of Factor Analyzers (MFA; Shafran et al., 2026). Fitted without supervision, the MFA has been shown to reconstruct activations as faithfully as sparse autoencoders while requiring far less data, and to recover semantically distinct regions of the latent space directly. Concretely, MFA is a generative model that explains a distribution over $\mathbb{R}^d$ as a mixture of K Gaussian components with density:

$$p(x) = \sum_k \pi_k \mathcal{N}(x; \mu_k, C_k). \tag{1}$$

Here π_k is the weight of component k and $\mu_k \in \mathbb{R}^d$ is its mean. Each covariance is low-rank-plus-diagonal, $C_k = W_k W_k^\top + \Psi$, with $W_k \in \mathbb{R}^{d \times R}$, $R \ll d$ its *loading matrix*. The diagonal Ψ, shared across components, holds the remaining per-coordinate variance. Equivalently, points are generated by component k as $x = \mu_k + W_k z + \epsilon$ with latent factors $z \sim \mathcal{N}(0, I_R)$ and noise $\epsilon \sim \mathcal{N}(0, \Psi)$. This provides a two-level description of the data: each Gaussian covers a coherent region of space, while the columns of W_k capture the local variation of points inside it. A fitted MFA assigns each

point to a region through the posterior responsibilities r_k and represents it in local coordinates $\hat{z}_k$,

$$r_k(x) = \frac{\pi_k \mathcal{N}(x; \mu_k, C_k)}{\sum_j \pi_j \mathcal{N}(x; \mu_j, C_j)}, \qquad \hat{z}_k(x) = (I_R + W_k^\top \Psi^{-1} W_k)^{-1} W_k^\top \Psi^{-1}(x - \mu_k), \quad (2)$$

so that each point admits a linear decomposition $x \approx \sum_{k=1}^{K} r_k(x)\, \rho_k(x)$ in terms of its per-component reconstruction $\rho_k(x) = \mu_k + W_k \hat{z}_k(x)$. Shafran et al. (2026) already exploit this for steering. We build onto this framework and show how it can be applied to reasoning in § 3.

3 Charting the Regions of Thought

This section describes the intervention methodology we follow. We fit an MFA on the sentence representations of thinking traces, so that its components partition the activation space into regions and interpret each trace as a walk over these regions. We label the regions based on their functional role, then intervene on how reasoning traverses these regions to establish their causal role.

3.1 Methods

Partitioning the latent space into regions via MFA We record activations of a thinking trace from the residual stream at layer ℓ and split them into sentences via a fixed set of delimiters (`.`, `!`, `?`, `\n`) on their corresponding generated tokens. Sentence representations are obtained via mean-pooling following prior works (Minegishi et al., 2025; Zhou et al., 2026; Venhoff et al., 2026), $x_s = \frac{1}{L_s} \sum_{m=1}^{L_s} h_{s,m}$, with $h_{s,m} \in \mathbb{R}^d$ the activations for the m-th token of sentence s and L_s the token length of sentence s.

We fit an MFA on $\{x_s\}$, gathered across multiple traces in a given task. We identify a MFA component with the *region* of activation space it covers, and with the graph node introduced below, indexing all three by the same $k \in \{1, \ldots, K\}$ and writing k for the region itself. Sentence s is assigned to the region that best explains its representation, $k_s = \arg\max_j r_j(x_s)$. We call its reconstruction $\rho_{k_s}(x_s)$. We refer to the sequence $(x_1, \ldots, x_S)$ as the *reasoning trajectory* and to the sequence of regions $(k_1, \ldots, k_S)$ it traverses as the *reasoning walk*.

By interpreting regions as nodes of a directed graph, it is possible to define edges whose weights are given by the transition probabilities of observed walks. Let N_{ij} be the number of transitions from region i to region j in a given corpus of traces. Row-normalizing these counts gives a first-order Markov estimate of the transition probabilities, $P(j \mid i) = N_{ij} / \sum_{j'} N_{ij'}$. These estimates are used in the reroute and promote interventions defined below.

Attributing functional roles to regions Among existing taxonomies of reasoning behavior (Appendix A), we choose the Schoenfeld episode categories (Schoenfeld, 1985; Li et al., 2025) as somewhat more established and for which an annotation framework exists (ThinkARM; Li et al., 2026b). It segments reasoning steps into eight categories: *reading* (RDN), *analyzing* (ALS), *planning* (PLN), *implementing* (IMP), *exploring* (EXP), *verifying* (VER), *monitoring* (MON) and *answering* (ANS) later added by Li et al. (2026b). Following ThinkARM, we annotate traces with LLM-as-a-judge using `Qwen3.8-27B`. Validated against the human-annotated gold labels released with ThinkARM, the judge reaches a Cohen's κ of 0.71, which indicates substantial agreement. We then attribute to each episode category e every region k in which steps of that category occur more often than across all annotated steps, that is, every region with $p(e \mid k) > p(e)$, where $p(e)$ is the fraction of annotated steps labeled e and $p(e \mid k)$ the same fraction among the steps assigned to k. A region may therefore be attributed to several categories.

Performing single-region interventions We steer an activation h from its region i towards a target region j as $h' = h + \alpha\big(\rho_j(h) - \rho_i(h)\big)$, with $\alpha = 1$. We use this to control region traversal in two ways: **rerouting** (R) steers the walk away from a region k^* into one of its successors, effectively setting $P(k^* \mid i)$ to zero for each region i; **promoting** (P) steers the next sentence into k^* whenever the current region i admits a transition to it, $P(k^* \mid i) > 0$, raising that probability to one. During generation, regions are assigned from the running mean of the current sentence, and interventions act on every token until the next delimiter. Steering is confined to the thinking trace and stops once the model terminates its reasoning, so the final answer is generated without any intervention.

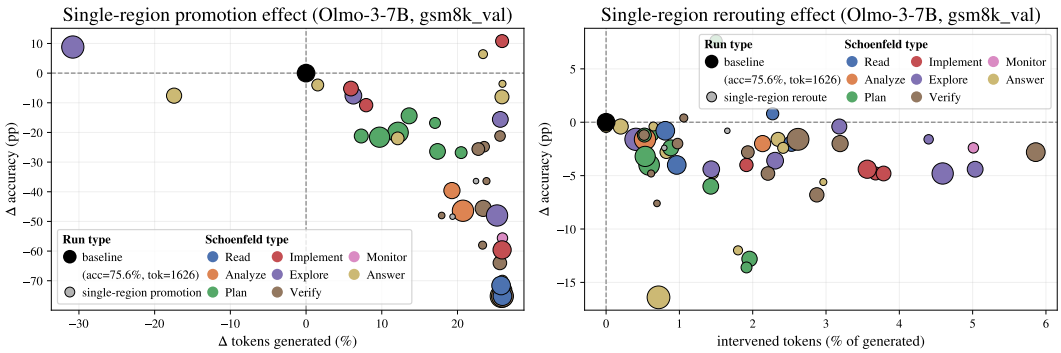


Figure 2: **Left:** promoting a small subset of regions improves accuracy *and* token usage. **Right:** rerouting a handful of load-bearing regions lowers accuracy by more than ten points while affecting under two percent of generated tokens. Each point represents one single-region intervention; with color representing the region's majority episode category, and dot size the relative population of the corresponding intervened region before intervention.

Composing multi-region interventions For each model–dataset pair, we sweep both interventions over all K regions on the validation split, intervening on a single region per run, and select subsets of regions to intervene on jointly according to several selection criteria. Five of these criteria use the outcome of the sweep: each scores every region based on its single-region run and selects the $n = 5$ highest-scoring regions. Three of them select regions whose intervention improves reasoning: ACC ranks regions by accuracy gain, TOK by reduction in generated tokens, and JNT by the combination of the two. The other two select regions where intervening on a small fraction of tokens causes a large drop in accuracy: LOAD ranks regions by the accuracy lost per fraction of intervened tokens, and FLIP by the net number of answers turned from correct to incorrect per intervened token. Additionally, the region attributions defined above provide further criteria to select every region attributed to a specific category. We use them to reroute the walk away from a single episode category in §4.2. Appendix B details the full steering method and formally defines each criterion.

Figure 2 shows a representative single-region sweep on the validation set. Most single-region promotions reduce performance, but a small subset improves accuracy, token usage, or *both*. Rerouting away from specific regions lowers accuracy far beyond their share of intervened tokens. For every model–dataset pair, each selected configuration is applied jointly on the test split. § 4 examines how multiple single-region intervention effects compound.

3.2 Experimental Setup

Models We test 8 reasoning models: `SmolLM3-3B`, `Gemma-4-E2B`, `DeepSeek-R1-Distill-Qwen-1.5B`, `Olmo-3-7B-Think`, and `Qwen3` at 4B, 8B, 14B, and 32B. This selection covers five different families, a distilled model, and several model scales in the same family. Traces are generated with each model's chat template and default decoding settings, with thinking enabled and a budget of 2048 tokens.

Datasets We evaluate on three benchmarks of increasing difficulty: GSM8K (Cobbe et al., 2021), consisting of grade-school arithmetic word problems; MATH (Hendrycks et al., 2021), consisting of competition mathematics problems answered by a free-form symbolic expression; and GPQA (Rein et al., 2023), consisting of graduate-level multiple-choice science questions, which we include to assess how regions generalize outside mathematical reasoning. For GSM8K and MATH we sample training and validation instances from the benchmark's own training split, while for GPQA we draw them from SuperGPQA (Team et al., 2025). For all datasets we sample 1000 training and 250 validation instances, and evaluate on test sets of 1319 (GSM8K), 5000 (MATH), and 448 (GPQA) instances. We fit the MFA and transition matrices on activations from the training split, select intervention regions on the validation split, and report final results on the test split.

MFAs and interventions Following Shafran et al. (2026), we fit one MFA per model–dataset pair, with $K = 50$ components of rank $R = 25$, trained for 400 epochs with a batch size of 256 and a learning rate of 10^{-3}. We set K to the largest value for which an exhaustive component-wise

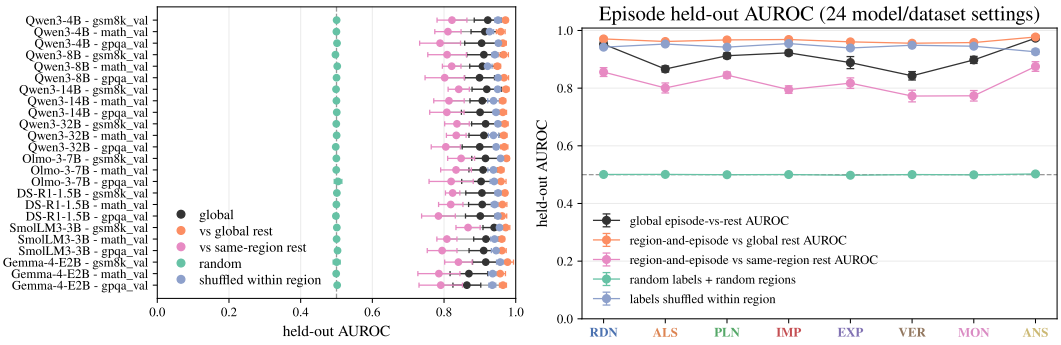


Figure 3: Episode categories overlap less within a single region than across the whole space, largely because of region membership itself. AUROC of discriminating an episode category from the remaining episodes by projecting held-out traces on their difference-in-means direction. Fitted globally or within regions, against two label-permutation baselines. **Left:** per model–dataset pair. **Right:** per episode category.

ablation fits our compute budget. All interventions, and all visualizations unless otherwise specified, are performed at relative depth $\ell = 0.5$, following common practice of attributing behaviors to intermediate layers. We first sweep promote and reroute interventions over single regions on the validation split, one run per region, and use the measured effects to select region subsets according to the multi-region criteria of §3.1. Each selected intervention is then evaluated on the test split.

Baselines We compare accuracy improvement and reasoning length reduction against two steering methods respectively: Contrastive Activation Addition (CAA; Rimsky et al., 2024) and Overclocking (OCK; Eisenstadt et al., 2025). The first steers generation along the difference-in-means direction between activations of correct and incorrect traces. The second along a "direction of progress" obtained via regression, effectively accelerating reasoning. For both baselines, intervention strength is selected on the validation split and performance reported on the test split. More details on hyperparameters are provided in Appendix D.

4 RESULTS

The following section reports our findings on the structure of latent space for reasoning tasks. We first use MFA to study how the latent space is organized, and then measure how effectively region-based interventions alter reasoning and its effectiveness.

4.1 EPISODE CATEGORIES CLUSTER INTO REGIONS WITH NONLINEAR BOUNDARIES

Our first finding, illustrated in Figure 1, is that regions separate functionally distinct behaviors which align with the episode categories of classical human problem-solving analysis (Schoenfeld, 1985). Prior work has located this functional distinction in attention heads (Bogdan et al., 2025) or in steering directions (Venhoff et al., 2025; Chang et al., 2025; Lippl et al., 2026). We complement these findings with evidence that regions also act as semantic units.

Regions are not uniform: most contain several episode categories (Figure 15). Inside a region, however, episodes overlap less than globally. We measure overlap by the AUROC of discriminating a category from the remaining episodes along their difference-in-means direction, scored on held-out traces. Restricting a category to a single region lowers its overlap across all 24 model–dataset pairs and every category, mostly through region membership itself: permuting labels within each region retains most of this gain (Figure 3).

The directions themselves are only valid locally. Comparing directions computed for each region, alignment drops sharply with the Wasserstein-2 distance between the corresponding MFA components (Figure 11). Local directions disagree among themselves and with the global one, while global directions fitted on random halves of the activations agree closely with each other (Figure 4). This suggests that linearity may only apply locally, explaining previous results (Haller et al., 2026; Braun

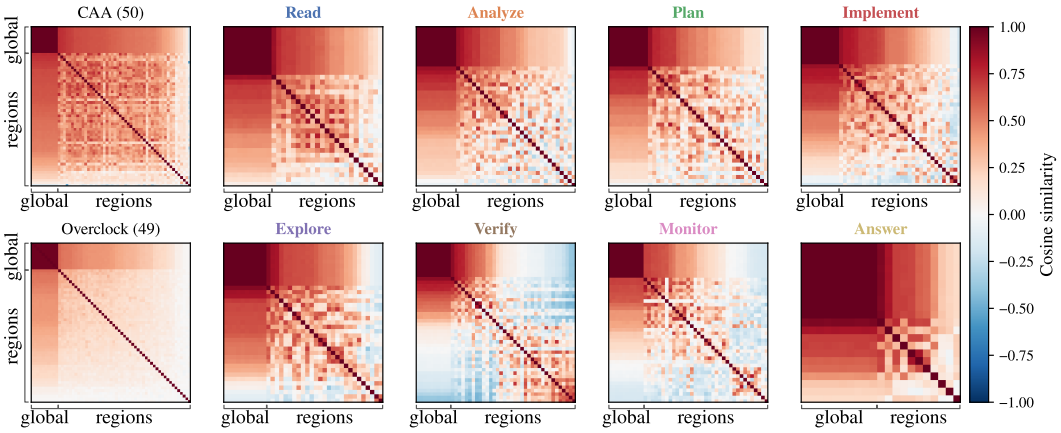


Figure 4: Local region-wise directions disagree with each other and with the global ones, forming orthogonal and antiparallel groups, while global directions agree closely among themselves. Each panel shows pairwise cosine similarities among 10 global directions, each fitted on a random half of the activations, and one local direction per region. Episode panels use difference-in-means between the category and all other categories; the CAA and OCK panels use the direction each method computes (**Qwen3-14B**, MATH).

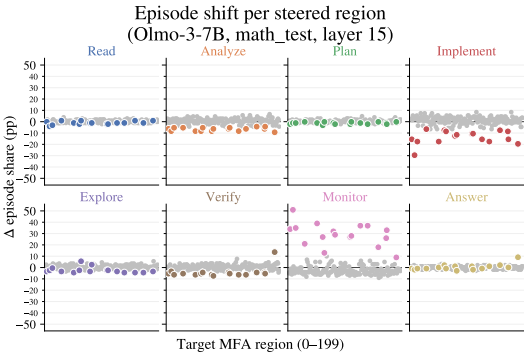


Figure 5: Steering makes only the monitoring category more frequent, others cannot be induced this way. Change in the share of episode categories after intervention, for each target region (**Olmo-3-7B**, MATH, MFA with $K = 200$, $\ell = 0.5$). Colored points mark significant regions.

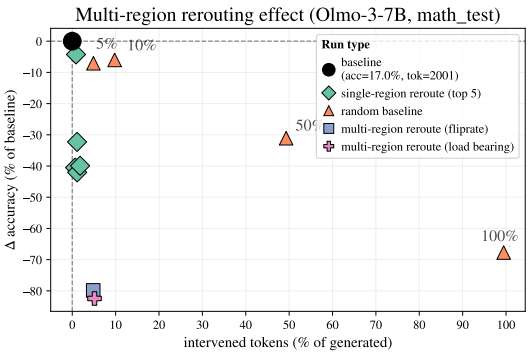


Figure 6: Rerouting load-bearing regions lowers accuracy more than an equal-scale noise baseline applied to a substantially larger fraction of reasoning tokens. LOAD and FLIP multi-region interventions against noise added to increasing proportions of the trace (**Olmo-3-7B**, MATH).

et al., 2025): different parts of the latent space encode the same behavior differently. This structure may also predict effectiveness for nonlinear steering methods (Raval et al., 2026).

4.2 Region interventions trigger a recovery mechanism

Regions are also causally active: rerouting the walk away from every region attributed to an episode category, as selected by the episode criteria of §3.1, changes the relative frequency of that category (Figure 9). The effect is partial and not uniform in sign: rerouting reduces relative frequency only for some episodes and sharply increases it in others.

To study this behavior more in detail, we run an experiment inspired by interchange intervention (Wu et al., 2023) on **Olmo-3-7B** and MATH. We cut unintervened test traces at a random sentence boundary, regenerate only the next sentence while steering every generated token towards a target region, and label the new sentence with its episode category. A finer MFA with $K = 200$ components is used, each region is chosen as the target for 100 intervention instances, and the experiment is repeated at relative depths $\ell \in \{0.3, 0.5, 0.7\}$. Monitoring is the only episode category that steering makes consistently more frequent, while implement, analyze and verify steps become rarer (Fig-

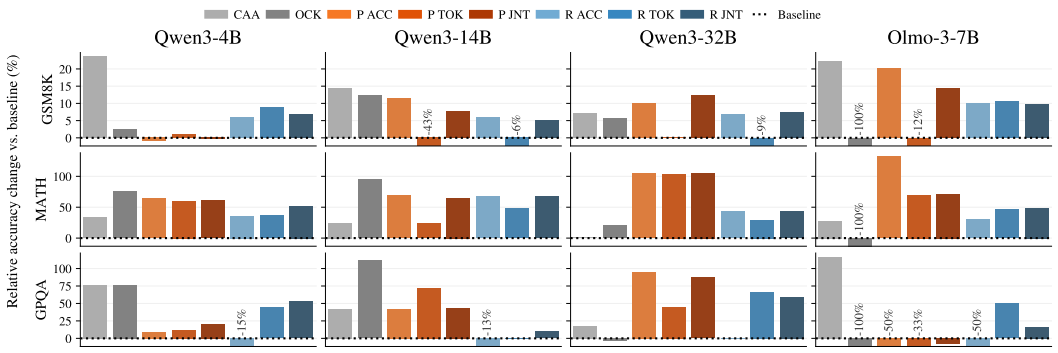


Figure 7: Region-based steering delivers performance comparable with linear steering. Relative accuracy improvement over the unintervened baseline across models, tasks and intervention types.

ure 5). Moreover, this pattern is consistent across layers (Figure 8). No region increases the frequency of any of the remaining categories. Further details are provided in Appendix C.

These results suggest that a region may not be a self-contained unit of behavior: steering towards most regions does not significantly change the episode of the next sentence, and the regions that do change it almost all make the same category more frequent. We speculate that this reflects a recovery mechanism that the model triggers in response to excessive disruption, which would also explain the higher frequency of monitoring behavior observed in Figure 9. These results point to a different and possibly more granular control mechanism, which a finer partition of the space should expose.

4.3 A SMALL SET OF REGIONS IS LOAD-BEARING

A small subset of regions carries critical information the model needs to compute the answer, and rerouting away from it collapses task accuracy. The effect is shown in Figure 2: a handful of regions cause a drop in accuracy of more than ten points while affecting under two percent of the generated tokens. The effect accumulates when regions are ablated together. Figure 6 shows the effect of LOAD and FLIP multi-region interventions. A random ablation is shown for reference: we add i.i.d. noise to every reasoning token in the trace, sampled from $\mathcal{N}(0, \Sigma)$ with Σ the covariance of the recorded activations at that layer. The noise intervention is less damaging than the load-bearing one despite being twenty times more frequent. The effect is model and task dependent (Table 4) and localized to specific regions (Figure 2). **`Gemma-4-E2B`** stands apart as extremely robust to this intervention. We interpret this as evidence of redundancy: more than one region accounts for a given function, so ablating one leaves an equivalent alternative reachable, a behavior resembling the Hydra effect (McGrath et al., 2023).

4.4 PROMOTING THE RIGHT REGIONS SHORTENS AND IMPROVES REASONING

The same machinery improves reasoning when it is pointed at the right regions. Figure 2 shows that promoting a small number of regions raises accuracy, shortens the trace, or both. Composing interventions on beneficial regions (ACC, TOK and JNT criteria) transfers these gains to the test split (Figures 7 and 12). The corresponding token usage is shown in Figure 13 and the full results in Table 3. Promotion gives the larger improvements on MATH, where promotion raises the relative accuracy of the **`Qwen3`** models and **`Olmo-3-7B`** by over fifty percent. On GSM8K the accuracy improvements are smaller and sit alongside token reductions of up to a half for the larger **`Qwen3`** models. Most models and settings achieve below-chance accuracy on GPQA. This signals many necessitate resources beyond our allowed token budget. Nevertheless, interventions largely have a positive effect, confirming their efficacy even on this challenging task. **`Gemma-4-E2B`** is again mostly unaffected, a possible consequence of its per-layer embeddings (Team et al., 2026)

Against the baselines, P ACC improves on the unintervened run in 18 of the 22 available settings and exceeds CAA in 11 of them. The per-sentence directions induced by promotion are unaligned with the CAA direction (Figure 10), so the two methods reach their gains through different mechanisms. The advantage is task dependent: P ACC prevails on all 7 MATH settings, on 3 of the 8 on GPQA, but

only on 1 of the 7 on GSM8K. On length reduction, promoting for brevity shortens GSM8K traces more than Overclocking on six of the eight models. These reductions are not obtained without loss: in 6 out of 23 available settings for P TOK and in 9 of 24 for OCK, accuracy is critically reduced by more than 10% relative to baseline. The accuracy and token usage columns of Table 3 should therefore be read together. No method ultimately prevails, and we speculate this is because of the specific latent space geometry induced by each task: CAA and OCK may be more effective for those tasks where correctness or task progress are linearly distributed, while a task manifold with high curvature may favor region-based intervention.

5 Discussion & Conclusion

This work studies region-based attribution and intervention as an instrument for interpreting reasoning. We partition the latent space without supervision, interpret a thinking trace as a walk over the resulting regions, and measure the causal contribution each one makes. We attribute changes in accuracy to individual regions by sweeping both interventions over all K regions. All interventions are performed at a fixed intermediate layer, following prior work that attributes reasoning behaviors to intermediate layers. This choice makes our exhaustive ablation computationally tractable. The steering experiment of §4.2 produces the same pattern at multiple layers, which suggests that regions play similar roles at different depths. Identifying the layer at which region-based intervention is most effective remains an open question. The number of regions K is likewise set to the largest value for which the exhaustive sweep fits our compute budget. Predicting the effect of intervening on a region without running the intervention, for instance from existing signals of reasoning quality (Jiang et al., 2026), would make finer partitions practical to study and may reveal further effects.

Thinking traces concentrate in regions that align with human episode categories, and behavior is described more faithfully inside a single region than over the whole distribution, showing reasoning resides on a strongly curved manifold. These regions are causally active, yet rerouting a walk away from one changes behavior in ways its attribution does not predict, so no region can be interpreted in isolation. The walk itself is brittle: a small part of the region graph accounts for the model's ability to answer at all, while another part can be used to shorten and correct reasoning.

Attributing behaviors to regions proves easier than controlling them. Interventions largely cause corrective action in the model instead of the targeted behaviour. A similar phenomenon was observed in concurrent works (McKenzie et al., 2026). We speculate the culprit may be a mismatch between attribution and control: flow control may be exerted at a sub-sentence granularity, or take place earlier during generation. Future work should attribute exactly where within a thinking trace such decisions reside. Region-level attribution also invites mechanistic analysis over the latent space itself. Having established that region traversal causally affects downstream accuracy, one may ask for the minimum functional set a model must visit to solve a task, and which of these atoms recur across tasks. The question parallels circuit discovery (Conmy et al., 2023), with regions of activation space in place of model components.

Our results show that one direction may be an insufficient descriptor for behavioral concepts. This disagreement between local and global descriptors calls for a closer inspection of latent space geometry. Highly structured manifolds are known to arise for comparably simpler atomic properties (Engels et al., 2025; Tiblias et al., 2026). Complex behaviors such as toxicity, sycophancy, and faithfulness should admit a likewise complex description, yet they are commonly reduced to a single direction fitted over the whole space. Future work should therefore put more effort into decomposing complex behaviors into atomic steps and describing each one locally, a task for which region-based methods might be better equipped.

AI use statement

In this work, we used generative AI tools to expand bullet-point outlines into a rough first draft of the paper and to polish its writing, to obtain feedback on the research methodology and the experiments, and to review the final draft. Generative AI tools also assisted our literature search, extending it from the papers we already knew to related works we had not yet encountered. We have not used generative AI tools to identify gaps in the literature, to decide on the experimental setup, to define the methods, or to interpret the results and derive our findings. The choice of

which works are relevant and appear in the paper was made by the authors, and no citation was added without human supervision. Additionally, we used generative AI tools to implement most of the codebase underlying this research, including the experiment runner and the visualizations throughout the paper, and to adapt the codebases of prior work (MFA (Shafran et al., 2026) and ThinkARM (Li et al., 2026b)) to our setting. Following Li et al. (2026b), we annotated reasoning traces with an open-weight LLM used as a judge, and validated its annotations against the original human-annotated gold labels, obtaining a Cohen's κ of 0.71, which indicates substantial agreement. All AI-assisted work was reviewed, and the main author manually verified the correctness of the AI-generated code, both the code implementing our methods and the code producing the visualizations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Our experiments use exclusively open-weight models and publicly available datasets. Whenever a dataset is subsampled, the sample is drawn with a fixed random seed, so that the training and validation splits are reproducible. Section 3.2 lists the models, datasets, split sizes and MFA hyperparameters. Appendix B formally defines the region interventions and the criteria used to select regions for multi-region interventions, while Appendix C details the target-region experiment. The code to reproduce our results, along with detailed instructions, is available at `https://anonymous.4open.science/r/thought-trajectories-52EB`. It contains the implementation of our interventions as well as the code producing every figure and table in the paper.

References

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought Anchors: Which LLM Reasoning Steps Matter?, 2025. URL https://arxiv.org/abs/2506.19143.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (Un)Reliability of Steering Vectors in Language Models. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*, 2025. URL https://doi.org/10.48550/arxiv.2505.22637.

Fu-Chieh Chang, Yu-Ting Lee, and Pei-Yuan Wu. Unveiling the Latent Directions of Reflection in Large Language Models. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025. URL https://arxiv.org/abs/2508.16989.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Isaac Bloom. A Is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=R73ybUciQF.

Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. SEAL: Steerable Reasoning Calibration of Large Language Models for Free. In *Second Conference on Language Modeling (COLM)*, 2025a. URL https://openreview.net/forum?id=klPszYDIRT.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT Think That Much for 2+3=? On the Overthinking of Long Reasoning Models. In *Forty-Second International Conference on Machine Learning (ICML)*, 2025b. URL https://proceedings.mlr.press/v267/chen25bx.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, 2021. URL https://arxiv.org/abs/2110.14168.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL https://doi.org/10.48550/arxiv.2304.14997.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 2025. URL https://doi.org/10.1038/s41586-025-09422-z.

Julie Delon and Agnès Desolneux. A Wasserstein-type Distance in the Space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*, 13(2), 2020. URL https://doi.org/10.1137/19M1301047.

Roy Eisenstadt, Itamar Zimerman, and Lior Wolf. Overclocking LLM Reasoning: Monitoring and Controlling Thinking Path Lengths in LLMs, 2025. URL https://arxiv.org/abs/2506.07240.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Thomas Fel, Binxu Wang, Michael A. Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep Singh Lubana, Talia Konkle, Demba E. Ba, and Martin Wattenberg. Into the Rabbit Hull: From Task-Relevant Concepts in DINO to Minkowski Geometry. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=pbbRmnFkjG.

Satvik Golechha, Lucius Bushnaq, Euan Ong, Neeraj Kayal, and Nandi Schoots. Intricacies of Feature Geometry in Large Language Models. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=Ut3ml7Hdwx.

Wes Gurnee and Max Tegmark. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Patrick Haller, Mark Ibrahim, Polina Kirichenko, Levent Sagun, and Samuel J. Bell. LLM Knowledge Is Brittle: Truthfulness Representations Rely on Superficial Resemblance. In *Third Conference on Language Modeling (COLM)*, 2026. URL https://openreview.net/forum?id=WcJN5D82ZV.

Benjamin Heinzerling and Kentaro Inui. Monotonic Representation of Numeric Attributes in Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL https://aclanthology.org/2024.acl-short.18.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://www.semanticscholar.org/paper/57d1e7ac339e783898f2c3b1af55737cbeee9fc5.

Yihuai Hong, Meng Cao, Dian Zhou, Lei Yu, and Zhijing Jin. The Reasoning-Memorization Interplay in Language Models Is Mediated by a Single Direction. In *Findings of the Association for Computational Linguistics: ACL 2025 (Findings of ACL)*, 2025. URL https://aclanthology.org/2025.findings-acl.1111.

Yao Huang, Huanran Chen, Shouwei Ruan, Yichi Zhang, Xingxing Wei, and Yinpeng Dong. Mitigating Overthinking in Large Reasoning Models via Manifold Steering. In *Neural Information Processing Systems (NeurIPS)*, 2025. URL https://doi.org/10.48550/arXiv.2505.22411.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI O1 System Card, 2024. URL https://arxiv.org/abs/2412.16720.

Seogyeong Jeong, Jaehui Hwang, Dongyoon Han, Geonmo Gu, Alice Oh, and Taekyung Kim. Beneath the Surface of Chains-of-Thought: A Mechanistic Interpretation of Reasoning Operations in LLMs, 2026. URL https://arxiv.org/abs/2609.04753.

Xinyan Jiang, Ninghao Liu, Di Wang, and Lijie Hu. Beyond Scalars: Evaluating and Understanding LLM Reasoning via Geometric Progress and Stability, 2026. URL https://arxiv.org/abs/2603.10384.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering Vector Fields for Context-Aware Inference-Time Control in Large Language Models, 2026a. URL https://arxiv.org/abs/2602.01654.

Ming Li, Nan Zhang, Chenrui Fan, Hong Jiao, Yanbin Fu, Sydney Peters, Qingshu Xu, Robert Lissitz, and Tianyi Zhou. Understanding the Thinking Process of Reasoning Models: A Perspective from Schoenfeld's Episode Theory. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://aclanthology.org/2025.emnlp-main.922.

Ming Li, Chenrui Fan, Yize Cheng, Soheil Feizi, and Tianyi Zhou. Schoenfeld's Anatomy of Mathematical Reasoning by Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026 (ACL)*, 2026b. URL https://doi.org/10.18653/V1/2026.ACL-LONG.1513.

Samuel Lippl, Thomas McGee, Kimberly Lopez, Ziwen Pan, Pierce Zhang, Salma Ziadi, Oliver Eberle, and Ida Momennejad. Algorithmic Primitives and Compositional Geometry of Reasoning in Language Models, 2026. URL https://arxiv.org/abs/2510.15987.

Chunwei Ma and Russell Wolfinger. Laguerre Geometry for Interpreting Large Language Models, 2026. URL https://arxiv.org/abs/2607.10578.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *First Conference on Language Modeling (COLM)*, 2024. URL https://doi.org/10.48550/arxiv.2310.06824.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The Hydra Effect: Emergent Self-repair in Language Model Computations. https://arxiv.org/abs/2307.15771v1, 2023. URL https://arxiv.org/abs/2307.15771.

Alex McKenzie, Keenan Pepper, Stijn Servaes, Martin Leitgab, Murat Cubuktepe, Mike Vaiana, Diogo de Lucena, Judd Rosenblatt, and Michael S. A. Graziano. Endogenous Resistance to Activation Steering in Language Models, 2026. URL https://arxiv.org/abs/2602.06941.

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Topology of Reasoning: Understanding Large Reasoning Models through Reasoning Graph Properties. In *Neural Information Processing Systems (NeurIPS)*, 2025. URL https://doi.org/10.48550/arXiv.2506.05744.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The Origins of Representation Manifolds in Large Language Models, 2025. URL https://arxiv.org/abs/2505.18235.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *ICML'24*. JMLR.org, 2024. URL https://doi.org/10.48550/arxiv.2311.03658.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The Geometry of Categorical and Hierarchical Concepts in Large Language Models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Joris Postmus and Steven Abreu. Steering Large Language Models Using Conceptors: Improving Addition-Based Activation Engineering. In *MINT: Foundation Model Interventions*, 2024. URL https://doi.org/10.48550/arxiv.2410.16314.

Shivam Raval, Hae Jin Song, Linlin Wu, Abir Harrasse, Jeff M. Phillips, Fazl Barez, and Amirali Abdullah. Curveball Steering: The Right Direction To Steer Isn't Always Linear, 2026. URL https://arxiv.org/abs/2603.09313.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

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 (ACL)*, 2024. URL https://aclanthology.org/2024.acl-long.828.

Alan H. Schoenfeld. *Mathematical Problem Solving*. Academic Press, 1985. URL https://www.sciencedirect.com/book/monograph/9780126288704/mathematical-problem-solving.

Or Shafran, Shaked Ronen, Omri Fahn, Shauli Ravfogel, Atticus Geiger, and Mor Geva. From Directions to Regions: Decomposing Activations in Language Models via Local Geometry. In *Forty-Third International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=Tz7n7pX9cO.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=HvoG8SxggZ.

Daniel Chee Hian Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the Generalisation and Reliability of Steering Vectors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. URL https://doi.org/10.52202/079017-4417.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, et al. Gemma 4 Technical Report, 2026. URL https://arxiv.org/abs/2607.02770.

M.-A.-P. Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, David Ma, Yuansheng Ni, Haoran Que, et al. SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL https://doi.org/10.52202/085713-3766.

Federico Tiblias, Irina Bigoulaeva, Jingcheng Niu, Simone Balloccu, and Iryna Gurevych. Hypothesis-Driven Feature Manifold Analysis in LLMs via Supervised Multi-Dimensional Scaling. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=vCKZ40YYPr.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Understanding Reasoning in Thinking Language Models via Steering Vectors. In *Workshop on Reasoning and Planning for Large Language Models at ICLR 2025*, 2025. URL https://openreview.net/forum?id=OwhVWNOBcz.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Base Models Know How to Reason, Thinking Models Learn When. In *Forty-Third International Conference on Machine Learning (ICML)*, 2026. URL https://doi.org/10.48550/arXiv.2510.07364.

Martina G. Vilas, Safoora Yousefi, Besmira Nushi, Eric Horvitz, and Vidhisha Balachandran. Tracing the Traces: Latent Temporal Signals for Efficient and Accurate Reasoning. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=Ytlj8Ckv9n.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, NIPS '22. Curran Associates Inc., 2022. URL https://doi.org/10.52202/068431-1800.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at Scale: Identifying Causal Mechanisms in Alpaca. In *Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL https://doi.org/10.48550/arxiv.2305.08809.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Sheldon Yu, Yuxin Xiong, Junda Wu, Xintong Li, Tong Yu, Xiang Chen, Ritwik Sinha, Jingbo Shang, and Julian McAuley. Explainable Chain-of-Thought Reasoning: An Empirical Analysis on State-Aware Reasoning Dynamics. In *Findings of the Association for Computational Linguistics: EMNLP 2025 (Findings of EMNLP)*, 2025. URL https://aclanthology.org/2025.findings-emnlp.904.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How Language Model Hallucinations Can Snowball. In *Forty-First International Conference on Machine Learning (ICML)*, 2024. URL https://doi.org/10.48550/arxiv.2305.13534.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru Zhang. The Geometry of Reasoning: Flowing Logics in Representation Space. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=ixr5Pcabq7.

A Extended Literature Review

A.1 Reasoning as a Trajectory

A *reasoning trace* is the sequence of intermediate text generated by a chain-of-thought model (Wei et al., 2022) prior to its final answer. Recent interpretability works read the latent activations associated to such a trace as a literal trajectory and study its geometry. Jiang et al. (2026) show that correct traces exhibit high displacement and low curvature, while incorrect traces show the reverse. Vilas et al. (2026) complement this by finding that successful trajectories cover greater distances, advance more directly towards the final state, and deviate less from the direct path. Zhou et al. (2026) find that traces sharing the same underlying logical structure have similar velocity and curvature regardless of surface wording, while traces on the same topic occupy similar positions in the space, suggesting that logic is encoded in the dynamics of the trajectory and content in its location. Minegishi et al. (2025) take a complementary view, representing the reasoning process as a graph by clustering sentence representations and studying its topological properties, and find that models distilled from larger teachers exhibit more and larger cycles in this graph. Similarly, Yu et al. (2025) model reasoning steps as state transitions in a Markov chain and identify different functional roles for each state.

Alongside these trajectory-level accounts, a second line assigns functional roles to reasoning steps directly and formalizes them into taxonomies of reasoning behavior. Venhoff et al. (2025) annotate sentences with a set of step categories and derive one steering vector per category. Bogdan et al. (2025) reuse that taxonomy to ask which steps matter, quantifying the importance of each sentence by resampling it and finding that a small number of sentences, together with attention heads that spike in response to them, carry the final answer. Jeong et al. (2026) instead annotate traces with reasoning operations derived from Pólya's problem-solving framework and show that these operations are geometrically separable in the model's hidden representations. The taxonomy we adopt (§3.1) predates all of these: it comes from the mathematics education literature and was developed to describe human mathematical problem solving.

A.2 The Linear Representation Hypothesis

The *Linear Representation Hypothesis* (LRH) posits that language models encode concepts as directions in their latent space, organized into structured multidimensional manifolds (Park et al., 2024; Modell et al., 2025). Evidence for this view is well-established for atomic concepts: Marks & Tegmark (2024) show that true and false statements are linearly separable across diverse topics; Gurnee & Tegmark (2024) recover geographic coordinates and timeline positions from linear

probes trained on entity activations; Heinzerling & Inui (2024) find that numeric attributes such as birth year and population are monotonically encoded in the residual stream; and Tiblias et al. (2026) provide statistically significant evidence that feature manifolds reflect the topology of the underlying data distribution across different models. These results provide compelling evidence that atomic concepts form complex, highly structured geometries.

A.3 Steering Along Linear Directions

These directional representations motivate *activation steering*, which has been used to elicit or suppresses a target behavior by editing latent activations at inference time. Contrastive Activation Addition (CAA; Rimsky et al., 2024) is the canonical instance: it computes the difference in mean between two contrastive sets of activations and adds the resulting vector during generation. The same recipe carries over to chain-of-thought reasoning at the granularity of individual reasoning steps. Venhoff et al. (2025) promote or suppress individual step categories with one CAA vector per category, Chang et al. (2025) find suppressing reflection and backtracking easier than eliciting them, and Lippl et al. (2026) recover compositional geometric structure among functional step types. A second line steers for efficiency: Chen et al. (2025a) shift generation away from reflection and transition steps towards execution, Huang et al. (2025) steer within a low-dimensional PCA subspace to curb overthinking, and Eisenstadt et al. (2025) shorten traces along the direction that encodes relative position within the thinking block, each reporting reduced token usage without degrading accuracy and without any parameter update. A further direction separates reasoning-intensive from memorization-intensive processing (Hong et al., 2025). High-level reasoning behaviors therefore leave steerable signatures in activation space.

A.4 Limitations of the Linear View

Despite these successes, the LRH has known limitations. Not all features conform to linear directions: Engels et al. (2025) show that days of the week and months of the year form circular manifolds, and demonstrate that these circular structures are irreducible to individual directions. The truth direction has been proven brittle in out-of-domain settings driven by superficial transformations to the input (Haller et al., 2026). Golechha et al. (2025) show that the reported orthogonality and polytope structure of categorical and hierarchical concepts (Park et al., 2024; 2025) follow trivially from high dimensionality, appearing even for random concept categories and untrained unembeddings. More critically for steering, Tan et al. (2024) find that many behaviors are encoded by more than one direction or by combinations of opposed sub-behaviors, and Braun et al. (2025) find steering vectors for the same concept disagree between variations of the prompts used to construct them. It is increasingly evident from these results that linearity might not be a global property. This gap motivates the search for representations that localize a behavior to a bounded region of latent space.

A.5 From Directions to Regions

Early steps have been taken in characterizing concepts as *regions* of latent space rather than single directions. Fel et al. (2026) describe features as convex polytopes so that every activation is a convex combination of archetypal examples from its distribution. Ma & Wolfinger (2026) partition the space into weighted Voronoi cells, each an atomic concept, with higher-level concepts formed by groups of cells that share the atomic concepts lying in their intersection. Shafran et al. (2026) fit a Mixture of Factor Analyzers (MFA) to the activation distribution and recover both narrow components tied to atomic, often syntactic features and broad ones covering thematic topics with local semantic variation. We adopt their MFA and describe it in §2.2.

B Region Steering

Performing single-region interventions Using the per-region reconstruction ρ_k defined above, we steer activations from a source region i towards a target region j as:

$$h' = h + \alpha\big(\rho_j(h) - \rho_i(h)\big). \tag{3}$$

We fix $\alpha = 1$ unless specified otherwise. This has the effect of shifting an activation into another region and relocate it relative to its mean. We combine several such interventions to control region traversal in two ways:

- **Reroute** (R): the walk is steered away from a region k^*. Whenever a sentence is assigned to k^*, a successor $j \neq k^*$ is drawn from $P(\cdot \mid k^*)$, renormalized over the remaining regions, and the generation steered towards it. This sets the effective $P(k^* \mid i)$ to zero at every region i, passing that mass to the successors of k^*.
- **Promote** (P): the walk is steered into a region k^*. Whenever the region of the preceding sentence admits a transition into k^*, that is $P(k^* \mid k_{s-1}) > 0$, the current sentence is steered into k^*. This raises the effective $P(k^* \mid i)$ to one at every region i that admits such a transition.

In both cases the target is sampled once per sentence and the intervention applies to all tokens generated until the next delimiter. Note that while MFA is trained on sentence-level activations, interventions take place at the token level. For region attribution and intervention purposes, during generation we consider the running mean $\tilde{x}_s = \frac{1}{\tilde{L}_s} \sum_{m=1}^{\tilde{L}_s} h_{s,m}$ of the sentence currently being generated as an approximation of x_s. Interventions are applied only when $\tilde{x}_s$ triggers the promote or reroute conditions. Interventions are limited to the thinking trace: steering stops once the model terminates its reasoning, and the final answer is generated by the unintervened model conditioned on the steered trace. Any change in the answer therefore follows from the altered reasoning, not from steering the answer tokens directly.

Composing multi-region interventions For each model–dataset pair we sweep both interventions over all K regions, performing one single-region run per region on the validation split. We combine the results of this sweep with the episode annotations described above to select a subset of regions to intervene on jointly. To do this, the first five criteria below rank regions and select the n best-scoring ones ($n = 5$ unless otherwise specified), while the episode occupancy criteria select every region attributed to a category. We define the following criteria:

- **Accuracy** (ACC): regions are ranked by the accuracy gain $\Delta\mathrm{acc}$, measured in percentage points relative to the unintervened run. Only regions with $\Delta\mathrm{acc} > 0$ are retained.
- **Brevity** (TOK): regions are ranked by reduction in generation length $\Delta\mathrm{tok}$, expressed as a percentage of the tokens generated by the unintervened run. Only regions with $\Delta\mathrm{tok} < 0$ are retained.
- **Joint** (JNT): regions are ranked by $\Delta\mathrm{acc} - \Delta\mathrm{tok}$, so that raising accuracy and shortening the trace are required to hold simultaneously. Only regions with $\Delta\mathrm{acc} > 0$ and $\Delta\mathrm{tok} < 0$ are retained.
- **Load-bearing** (LOAD): regions are ranked by the accuracy lost per unit of intervention, $-\Delta\mathrm{acc}/f_k$, where f_k is the percentage of intervened tokens over total generated ones. Only regions with $\Delta\mathrm{acc} < 0$ are retained.
- **Flip rate** (FLIP): regions are ranked by the net number of answers turned incorrect per intervened token, $\left(q_k^- - q_k^+\right)/T_k$, where q_k^- is the number of questions the unintervened run answers correctly and the intervened run does not, q_k^+ the number for which the reverse holds, and T_k the number of intervened tokens. Only regions with a positive ratio are retained.
- **Episode occupancy** (RDN, ALS, PLN, IMP, EXP, VER, MON, ANS): an episode e is attributed every region k with $p(e \mid k) > p(e)$, where $p(e)$ is the fraction of all labeled steps that are labeled e and $p(e \mid k)$ the same fraction among the steps assigned to k. Eight region sets are thus obtained, one per episode category, and a region may be attributed to several episodes. These criteria select the regions rerouted in §4.2.

C TARGET-REGION INTERVENTIONS

This appendix details the single-sentence experiment of §4.2. We use `Olmo-3-7B-Think` on MATH and fit an MFA with $K = 200$ components on the training split, separately at relative depths $\ell = 0.3$, 0.5 and 0.7. Each unintervened test trace is cut at a uniformly sampled sentence boundary, keeping the exact generated prefix and chat template. From this prefix we generate a single sentence, up to the first delimiter, and steer every generated token from the current region i towards a target region j with Eq. 3. The target is drawn from all K regions regardless of transition probability $P(j \mid i)$.

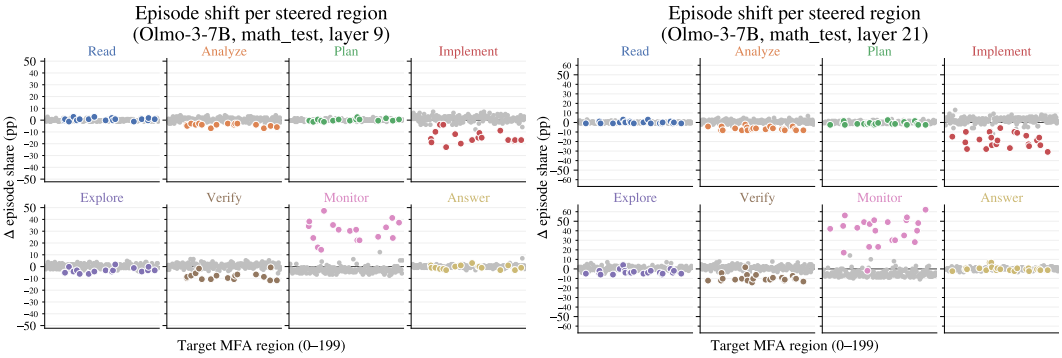


Figure 8: The effect of target-region interventions is consistent across layers: at both depths, the significant regions make monitoring sentences more frequent and implementing sentences less frequent. Same as Figure 5, at relative depths $\ell = 0.3$ (**left**) and $\ell = 0.7$ (**right**).

Each region is the target of 100 different intervention instances, giving 20,000 sentences per layer. Finally, the judge of §3.1 labels both generation and preceding context.

At each layer, we test the association between the target region and the episode category of the generated sentence with a χ^2 test, and report Cramér's V as its effect size. To identify individual regions, we compare the episode distribution of each region with the pooled distribution of all other regions and apply the Benjamini–Hochberg correction across the 200 tests. Because this test compares full distributions, the category that a significant region makes more frequent (Figure 5) is reported descriptively.

Table 1 reports the results. We first observe that the edit reaches the intended region, with 71 to 83% of the post-edit token assignments matching the target. The association between target region and episode category is significant at every layer, but its effect size is small ($V \approx 0.01$). After correction, 17, 17 and 22 of the 200 regions remain significant at $\ell = 0.3$, 0.5 and 0.7, respectively. Of these 56 regions, all but two increase the frequency of monitoring sentences; the other two, region 191 at $\ell = 0.5$ and region 78 at $\ell = 0.7$, increase the frequency of verifying sentences.

Table 1: Target-region interventions on `Olmo-3-7B`, MATH, MFA with $K = 200$. Intervention Success Rate (ISR) is the fraction of post-intervention token assignments that match the target region; p and V are the χ^2 p-value and Cramér's V of the association between target region and episode category. The last column counts the regions that remain significant after Benjamini–Hochberg correction, split by the category they make more frequent.

ℓ	ISR	p	V	Significant regions
0.3	70.8%	9.5×10^{-68}	0.0095	17
0.5	82.6%	4.7×10^{-41}	0.0089	17
0.7	81.0%	1.8×10^{-291}	0.0124	22

D HYPERPARAMETERS

For both baselines, the intervention strength α is swept on the validation split, and the selected value is used for the corresponding test run. For CAA, α scales the difference-in-means direction $\bar{h}_{\text{correct}} - \bar{h}_{\text{incorrect}}$ between activations of correct and incorrect traces, and is swept over $\{0.25, 0.5, 1, 2\}$. This range is compatible with the multipliers used by Rimsky et al. (2024). For OCK, α scales the direction of progress w fitted by ridge regression with a fixed regularization strength $\lambda = 10^{-10}$, and is swept over $\{25, 50, 100, 200\}$. This range is compatible with the strengths used by Eisenstadt et al. (2025). As with region-based interventions, both baselines are applied at relative depth $\ell = 0.5$ and only within the thinking trace. Table 2 reports the selected values.

Table 2: Intervention strength α selected on the validation split for each baseline, model and dataset.

	CAA			OCK		
Model	GSM8K	MATH	GPQA	GSM8K	MATH	GPQA
`Qwen3-4B`	2	2	2	50	50	50
`Qwen3-8B`	2	2	2	100	100	100
`Qwen3-14B`	2	2	2	200	200	200
`Qwen3-32B`	2	1	2	200	200	25
`Olmo-3-7B`	2	2	2	200	25	25
`DS-R1-1.5B`	2	2	1	200	200	200
`SmolLM3-3B`	0.5	1	1	25	25	25
`Gemma-4-E2B`	0.25	0.5	2	200	200	25

E Additional Findings & Ablations

This section contains ablations on the number of MFA components K, the depth at which promotion is applied, and the probability with which each eligible sentence is promoted, as well as general observations we make from our experiments. Unless stated otherwise, settings follow §3.2.

Finer partitions yield purer regions As K grows, regions become more informative of the episode category with diminishing returns, while the overlap between categories within each region decreases for most of them (Figure 17). We interpret this as evidence that thinking activations are organized at several levels of granularity: mixed regions such as those in Figure 15 split into purer ones, reminiscent of feature absorption in sparse autoencoders (Chanin et al., 2025).

Regions are shared across tasks and diverge with depth Figure 18 shows that the regions recovered by MFAs are similar for the same layer across different datasets. This similarity quickly weakens with layer distance, even within the same dataset. This suggests that regions reflect a task-independent organization of the latent space that becomes more task-specific in deeper layers, although we note this divergence may also be explained by growth of activation norms with depth.

Promotion improves accuracy across layers We repeat the single-region promotion sweep for `Olmo-3-7B` on MATH at $\ell \in \{0.3, 0.7\}$ (Figure 19). Also in these two settings, a subset of interventions is observed to improve accuracy. This shows that the benefits of promotion can transfer to other layers, although further evidence is necessary to establish this in general.

Larger models may benefit more from promotion For `Qwen3-32B` on MATH, single-region promotions span a wide range of effects and trace a distinct "sickle" shape (Figure 14). Compared to other observed settings, far more regions show a net-positive effect to both accuracy and token usage. Albeit limited, this evidence suggests that larger models are less negatively affected by interventions and stand to benefit more from altering their traversal of the latent space.

Promotion probability controls the intensity of the intervention By default, P ACC steers every sentence that meets the promotion condition. We instead steer each such sentence with probability p, effectively controlling the intensity of the intervention. For `Olmo-3-7B` on MATH, accuracy increases monotonically with p, with small gains at low probabilities and the largest ones as p approaches one. This suggests that the effect of promotion accumulates over the trace, and that p offers a way to tune its strength (Figure 20).

F Additional Figures & Tables

Table 3: Promotion more than doubles MATH accuracy for **Qwen3-32B** and **Olmo-3-7B**, and roughly halves GSM8K token usage for **Qwen3-8B** and **Qwen3-32B**. Accuracy and average token count per model, task and intervention type, against the unintervened baseline and the CAA and OCK baselines (test split). A dash marks settings where no region satisfied the selection criterion.

		GSM8K		MATH		GPQA	
		acc % ↑	avg tokens ↓	acc % ↑	avg tokens ↓	acc % ↑	avg tokens ↓
Qwen3-4B	baseline	71.3	1607 ± 24	21.5	1968 ± 6	7.6	1982 ± 22
	CAA	**88.2** +16.8	**1202** ± **25** -25.2%	**28.7** +7.2	**1905** ± **8** -3.2%	**13.4** +5.8	**1827** ± **37** -7.8%
	OCK	**73.0** +1.7	**783** ± **25** -51.3%	**37.7** +16.2	**1583** ± **14** -19.6%	**13.4** +5.8	**1618** ± **53** -18.4%
	P ACC	70.7 -0.7	1606 ± 25	**35.3** +13.8	**1592** ± **16** -19.1%	**8.3** +0.7	**1952** ± **26** -1.5%
	P TOK	**72.0** +0.7	**1610** ± **25** +0.2%	**34.5** +12.9	**1591** ± **16** -19.2%	**8.5** +0.9	**1946** ± **26** -1.8%
	P JNT	71.1 -0.2	1617 ± 25 +0.6%	**34.5** +13.0	**1599** ± **16** -18.7%	**9.2** +1.6	**1947** ± **26** -1.7%
	R ACC	**75.7** +4.3	**1462** ± **27** -9.0%	**29.3** +7.7	**1864** ± **10** -5.3%	6.5 -1.1	1989 ± 20 +0.4%
	R TOK	**77.7** +6.4	**1362** ± **29** -15.2%	**29.3** +7.8	**1832** ± **11** -6.9%	**10.9** +3.3	**1899** ± **34** -4.1%
	R JNT	**76.2** +4.8	**1463** ± **27** -9.0%	**32.7** +11.2	**1824** ± **10** -7.3%	**11.6** +4.0	**1907** ± **33** -3.7%
Qwen3-8B	baseline	70.6	1645 ± 23	19.3	1981 ± 5	7.6	1999 ± 17
	CAA	**89.3** +18.7	**1197** ± **24** -27.2%	**25.6** +6.3	**1936** ± **7** -2.2%	**10.5** +2.9	**1933** ± **27** -3.3%
	OCK	**79.7** +9.1	**1000** ± **29** -39.2%	**42.3** +23.0	**1709** ± **12** -13.7%	**13.8** +6.2	**1886** ± **32** -5.7%
	P ACC	**77.3** +6.8	**799** ± **18** -51.4%	**31.1** +11.8	**1248** ± **17** -37.0%	**10.9** +3.3	**1946** ± **27** -2.6%
	P TOK	**75.7** +5.2	**788** ± **17** -52.1%	**27.5** +8.2	**1808** ± **12** -8.7%	**9.8** +2.2	**1945** ± **27** -2.7%
	P JNT	**75.4** +4.9	**792** ± **17** -51.8%	**31.1** +11.7	**1259** ± **17** -36.5%	**11.2** +3.6	**1952** ± **26** -2.4%
	R ACC	**75.8** +5.2	**1429** ± **25** -13.1%	**34.8** +15.5	**1793** ± **11** -9.5%	**8.5** +0.9	**1956** ± **26** -2.1%
	R TOK	67.8 -2.8	1392 ± 30 -15.3%	**31.7** +12.4	**1836** ± **10** -7.3%	**9.4** +1.8	**1942** ± **29** -2.9%
	R JNT	**75.7** +5.2	**1425** ± **25** -13.4%	**33.4** +14.1	**1817** ± **11** -8.3%	**9.4** +1.8	**1945** ± **28** -2.7%
Qwen3-14B	baseline	80.4	1440 ± 26	25.2	1940 ± 7	13.4	1944 ± 27
	CAA	**92.0** +11.5	**1091** ± **24** -24.3%	**31.3** +6.1	**1886** ± **8** -2.8%	**19.0** +5.6	**1860** ± **35** -4.3%
	OCK	**90.5** +10.0	**783** ± **20** -45.7%	**49.2** +24.0	**1568** ± **13** -19.2%	**28.4** +15.0	**1577** ± **49** -18.9%
	P ACC	**89.8** +9.3	**1383** ± **24** -4.0%	**42.6** +17.4	**1462** ± **16** -24.6%	**19.0** +5.6	**1840** ± **37** -5.4%
	P TOK	46.1 -34.3	562 ± 17 -61.0%	**31.0** +5.9	**1603** ± **16** -17.4%	**23.0** +9.6	**1591** ± **53** -18.2%
	P JNT	**86.7** +6.2	**1337** ± **26** -7.1%	**41.5** +16.3	**1466** ± **16** -24.5%	**19.2** +5.8	**1835** ± **38** -5.6%
	R ACC	**85.3** +4.9	**1211** ± **28** -15.9%	**42.2** +17.1	**1661** ± **14** -14.4%	11.6 -1.8	1974 ± 23 +1.5%
	R TOK	75.7 -4.7	1197 ± 30 -16.9%	**37.1** +12.0	**1524** ± **16** -21.4%	13.4	1949 ± 27 +0.3%
	R JNT	**84.5** +4.1	**1178** ± **29** -18.2%	**42.0** +16.8	**1655** ± **14** -14.7%	**14.7** +1.3	**1941** ± **28** -0.2%
Qwen3-32B	baseline	83.2	1437 ± 24	25.7	1951 ± 6	14.1	1954 ± 24
	CAA	**89.2** +6.0	**1289** ± **23** -10.3%	**26.1** +0.4	**1943** ± **7** -0.4%	**16.5** +2.5	**1938** ± **25** -0.8%
	OCK	**88.0** +4.8	**1362** ± **25** -5.2%	**30.9** +5.2	**1916** ± **7** -1.8%	13.6 -0.4	1955 ± 24 +0.1%
	P ACC	**91.4** +8.3	**843** ± **19** -41.3%	**52.6** +26.9	**1466** ± **15** -24.9%	**27.5** +13.4	**1736** ± **43** -11.1%
	P TOK	83.2	753 ± 20 -47.6%	**52.4** +26.7	**1474** ± **15** -24.4%	**20.3** +6.2	**1836** ± **36** -6.0%
	P JNT	**93.5** +10.3	**762** ± **18** -47.0%	**52.8** +27.1	**1466** ± **15** -24.8%	**26.3** +12.3	**1826** ± **35** -6.5%
	R ACC	**88.9** +5.8	**1204** ± **25** -16.2%	**36.7** +11.0	**1806** ± **11** -7.4%	14.1	1962 ± 24 +0.4%
	R TOK	75.6 -7.6	1110 ± 24 -22.8%	**33.3** +7.5	**1686** ± **13** -13.6%	**23.4** +9.4	**1837** ± **37** -6.0%
	R JNT	**89.2** +6.1	**1108** ± **24** -22.9%	**36.9** +11.2	**1804** ± **11** -7.5%	**22.3** +8.3	**1839** ± **38** -5.9%
Olmo-3-7B	baseline	72.2	1673 ± 22	17.0	2001 ± 4	2.7	2035 ± 8
	CAA	**88.2** +16.0	**1385** ± **22** -17.2%	**21.7** +4.7	**1976** ± **5** -1.3%	**5.8** +3.1	**1988** ± **19** -2.3%
	OCK	0.0 -72.2	3 ± 0 -99.8%	0.0 -17.0	1936 ± 13 -3.2%	0.0 -2.7	1665 ± 72 -18.2%
	P ACC	**86.7** +14.5	**1126** ± **30** -32.7%	**39.3** +22.3	**1929** ± **8** -3.6%	1.3 -1.3	2037 ± 8 +0.1%
	P TOK	63.6 -8.6	1239 ± 31 -25.9%	**28.8** +11.9	**1982** ± **6** -0.9%	1.8 -0.9	2039 ± 7 +0.2%
	P JNT	**82.5** +10.3	**1118** ± **28** -33.2%	**29.1** +12.1	**1984** ± **6** -0.9%	2.5 -0.2	2036 ± 8
	R ACC	**79.4** +7.2	**1464** ± **27** -12.5%	**22.2** +5.2	**1936** ± **8** -3.3%	1.3 -1.3	2041 ± 7 +0.3%
	R TOK	**79.9** +7.7	**1376** ± **29** -17.7%	**24.8** +7.9	**1929** ± **8** -3.6%	**4.0** +1.3	**2023** ± **14** -0.6%
	R JNT	**79.3** +7.1	**1474** ± **27** -11.9%	**25.3** +8.3	**1928** ± **8** -3.7%	**3.1** +0.4	**2020** ± **13** -0.7%
DS-R1-1.5B	baseline	70.0	1416 ± 27	30.2	1853 ± 9	2.7	1837 ± 42
	CAA	**74.8** +4.8	**1213** ± **26** -14.4%	**31.5** +1.3	**1818** ± **10** -1.9%	**4.0** +1.3	**1613** ± **55** -12.2%
	OCK	**74.9** +4.9	**720** ± **19** -49.1%	**48.0** +17.9	**1352** ± **16** -27.1%	**10.0** +7.4	**1263** ± **63** -31.2%
	P ACC	**74.6** +4.6	**1288** ± **27** -9.1%	**41.0** +10.8	**1613** ± **14** -13.0%	**7.1** +4.5	**634** ± **53** -65.5%
	P TOK	55.7 -14.3	654 ± 17 -53.8%	19.3 -10.8	1653 ± 14 -10.8%	**7.8** +5.1	**1118** ± **68** -39.1%
	P JNT	**74.4** +4.4	**1311** ± **27** -7.4%	**41.8** +11.7	**1613** ± **14** -12.9%	**10.5** +7.8	**649** ± **55** -64.6%
	R ACC	**77.2** +7.2	**1047** ± **30** -26.1%	**39.9** +9.7	**1665** ± **13** -10.1%	1.3 -1.3	1874 ± 38 +2.0%
	R TOK	**72.3** +2.3	**950** ± **28** -32.9%	**43.3** +13.1	**1574** ± **15** -15.1%	**4.7** +2.0	**1618** ± **61** -11.9%
	R JNT	**74.8** +4.8	**1073** ± **31** -24.3%	**40.4** +10.3	**1679** ± **13** -9.4%	**2.9** +0.2	**1829** ± **43** -0.4%
SmolLM3-3B	baseline	86.7	881 ± 25	40.0	1736 ± 11	6.9	1939 ± 28
	CAA	**88.1** +1.4	**781** ± **21** -11.4%	**43.6** +3.6	**1671** ± **12** -3.7%	**10.7** +3.8	**1837** ± **37** -5.3%
	OCK	0.0 -86.7	2048 ± 0 +132.3%	0.0 -40.0	2048 ± 0 +18.0%	0.0 -6.9	2048 ± 0 +5.6%
	P ACC	84.5 -2.3	956 ± 28 +8.5%	**44.9** +4.9	**1377** ± **16** -20.6%	**7.6** +0.7	**1933** ± **29** -0.3%
	P TOK	84.2 -2.6	821 ± 25 -6.9%	**44.6** +4.6	**1368** ± **16** -21.2%	**13.2** +6.2	**1689** ± **49** -12.9%
	P JNT	86.4 -0.3	879 ± 25 -0.3%	**44.4** +4.5	**1376** ± **16** -20.8%	5.6 -1.3	1914 ± 32 -1.3%
	R ACC	**87.1** +0.4	**864** ± **26** -2.0%	**45.6** +5.6	**1275** ± **18** -26.5%	**11.2** +4.2	**1656** ± **56** -14.6%
	R TOK	84.5 -2.3	506 ± 22 -42.6%	**45.3** +5.4	**1243** ± **19** -28.4%	**10.7** +3.8	**1507** ± **62** -22.3%
	R JNT	86.6 -0.2	782 ± 25 -11.3%	**44.6** +4.6	**1321** ± **18** -23.9%	**12.3** +5.4	**1620** ± **58** -16.5%
Gemma-4-E2B	baseline	78.6	944 ± 18	38.1	1577 ± 14	12.7	1838 ± 32
	CAA	**79.5** +0.8	**949** ± **18** +0.5%	**38.5** +0.4	**1580** ± **14** +0.1%	**14.1** +1.3	**1769** ± **33** -3.7%
	OCK	15.8 -62.8	1651 ± 36 +74.9%	4.8 -33.4	1613 ± 21 +2.3%	9.2 -3.6	1821 ± 33 -0.9%
	P ACC	NA	NA	NA	NA	10.0 -2.7	1865 ± 29 +1.5%
	P TOK	77.1 -1.5	867 ± 21 -8.2%	NA	NA	10.9 -1.8	1828 ± 31 -0.5%
	P JNT	NA	NA	NA	NA	9.2 -3.6	1852 ± 30 +0.8%
	R ACC	78.1 -0.5	941 ± 18 -0.3%	**38.6** +0.5	**1569** ± **14** -0.5%	10.5 -2.2	1846 ± 32 +0.5%
	R TOK	78.4 -0.2	935 ± 19 -1.0%	**38.3** +0.2	**1559** ± **15** -1.2%	9.8 -2.9	1843 ± 32 +0.3%
	R JNT	**78.8** +0.2	**940** ± **18** -0.5%	38.0 -0.1	1571 ± 15 -0.4%	11.6 -1.1	1839 ± 32 +0.1%

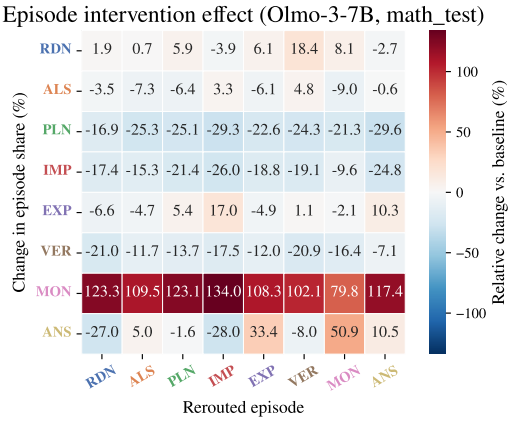


Figure 9: Rerouting changes the relative frequency of an episode category, though not uniformly in sign. Reroutes target the regions attributed to each category.

Table 4: Rerouting load-bearing regions lowers GSM8K accuracy by up to 46 points, while **Gemma-4-E2B** stays within three points of its baseline on every task. Accuracy under LOAD and FLIP reroute interventions per model and task, against the unintervened baseline (test split).

		GSM8K acc % ↑	MATH acc % ↑	GPQA acc % ↑
Qwen3-4B	baseline	71.3	21.5	7.6
	R LOAD	61.2 -10.2	**21.6** +0.1	**8.3** +0.7
	R FLIP	62.6 -8.7	14.1 -7.5	**9.8** +2.2
Qwen3-8B	baseline	70.6	19.3	7.6
	R LOAD	54.5 -16.1	**19.4** +0.1	6.9 -0.7
	R FLIP	69.5 -1.1	11.4 -7.9	4.9 -2.7
Qwen3-14B	baseline	80.4	25.2	13.4
	R LOAD	56.6 -23.9	17.5 -7.6	12.3 -1.1
	R FLIP	59.2 -21.2	19.3 -5.9	12.5 -0.9
Qwen3-32B	baseline	83.2	25.7	14.1
	R LOAD	51.1 -32.1	22.2 -3.6	**14.3** +0.2
	R FLIP	56.0 -27.1	21.9 -3.8	**15.4** +1.3
Olmo-3-7B	baseline	72.2	17.0	2.7
	R LOAD	39.7 -32.5	3.0 -14.0	0.0 -2.7
	R FLIP	26.2 -46.0	3.4 -13.6	0.0 -2.7
DS-R1-1.5B	baseline	70.0	30.2	2.7
	R LOAD	32.7 -37.3	3.5 -26.7	**3.6** +0.9
	R FLIP	38.1 -31.9	9.7 -20.5	**3.1** +0.4
SmolLM3-3B	baseline	86.7	40.0	6.9
	R LOAD	83.1 -3.6	27.4 -12.6	5.6 -1.3
	R FLIP	85.2 -1.5	30.3 -9.7	5.1 -1.8
Gemma-4-E2B	baseline	78.6	38.1	12.7
	R LOAD	78.4 -0.2	36.1 -2.0	9.8 -2.9
	R FLIP	**78.8** +0.1	37.3 -0.8	10.9 -1.8

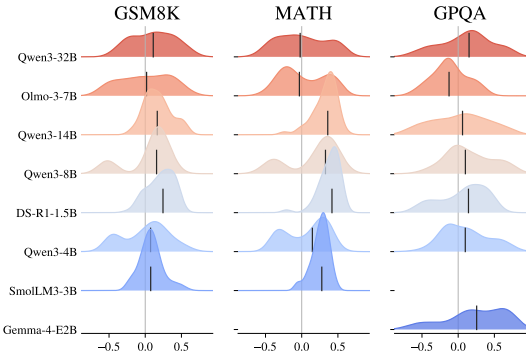


Figure 10: Region promotion induces steering directions unaligned with CAA, indicating a different mechanism. Distribution of cosine distances between the CAA direction and the per-sentence steering directions induced by region promotion, per model and dataset (test split).

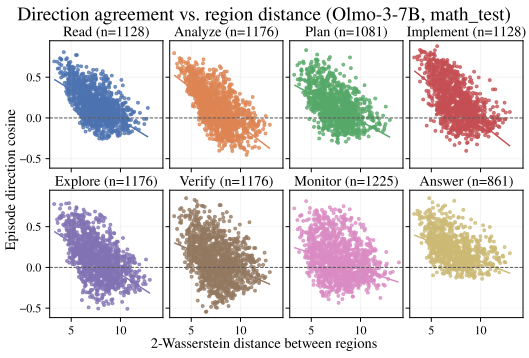


Figure 11: Agreement between region-wise episode directions decays with distance, as expected of a smoothly curved manifold. Cosine similarity between region-wise episode category directions against the 2-Wasserstein distance between the corresponding MFA components (`Olmo-3-7B`, MATH, test split). Each point is a pair of regions, and n is the number of region pairs compared for that category.

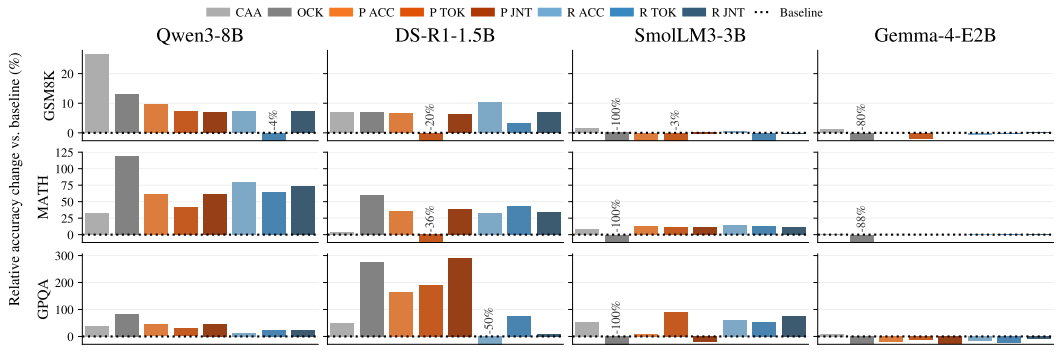


Figure 12: Relative accuracy improvement over the unintervened baseline for the remaining four models, across tasks and intervention types (test split). Continuation of Figure 7.

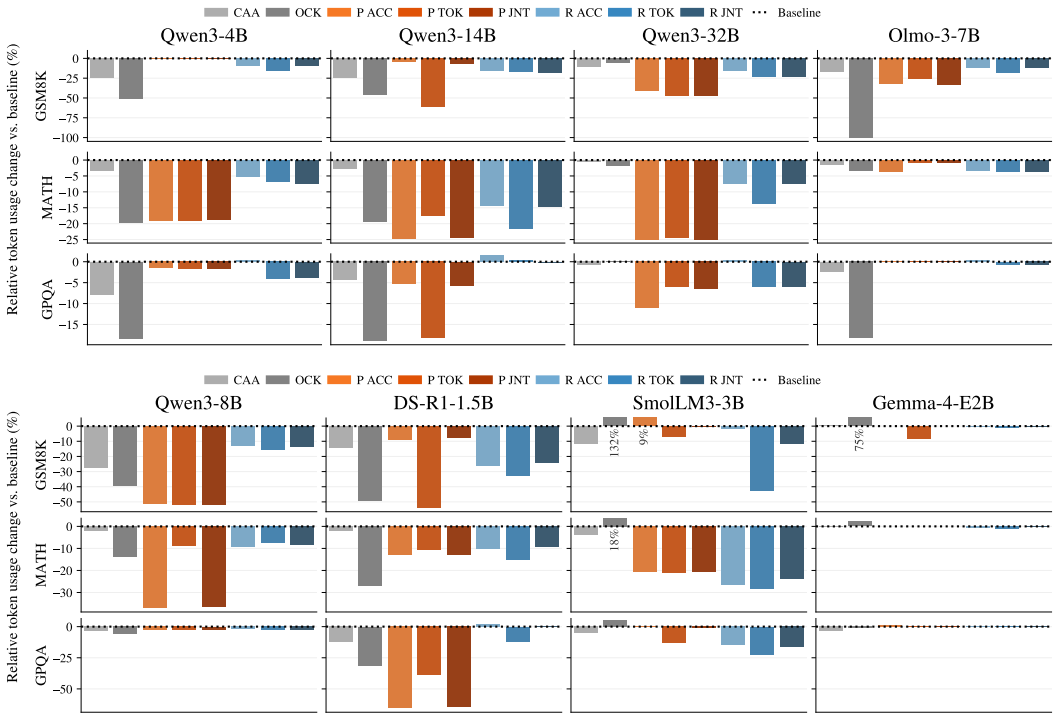


Figure 13: Token usage falls most on GSM8K, by roughly half for **Qwen3-8B** and **Qwen3-32B**, while several settings leave trace length unchanged. Relative token usage change against the unintervened baseline, across all eight models, tasks and intervention types (test split).

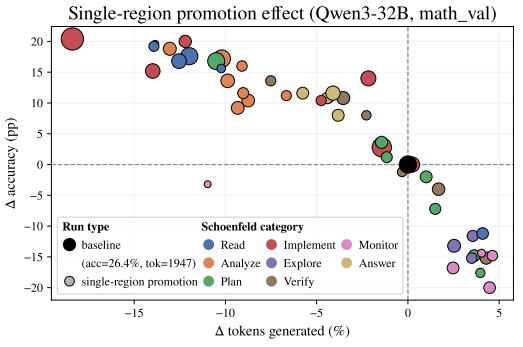


Figure 14: Promotions trace a sickle: regions that shorten the trace improve accuracy, while those that lengthen it lower accuracy. Single-region promotion sweep on MATH (**Qwen3-32B**, validation split). Colors and dot sizes as in Figure 2.

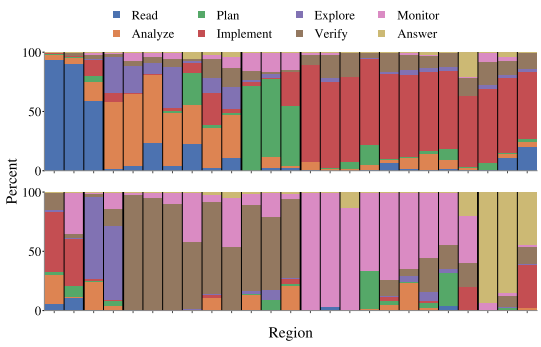


Figure 15: Only a few regions are monosemantic; most host several episode categories. Episode composition of 50 MFA regions sorted by dominant category (**Olmo-3-7B**, GSM8K).

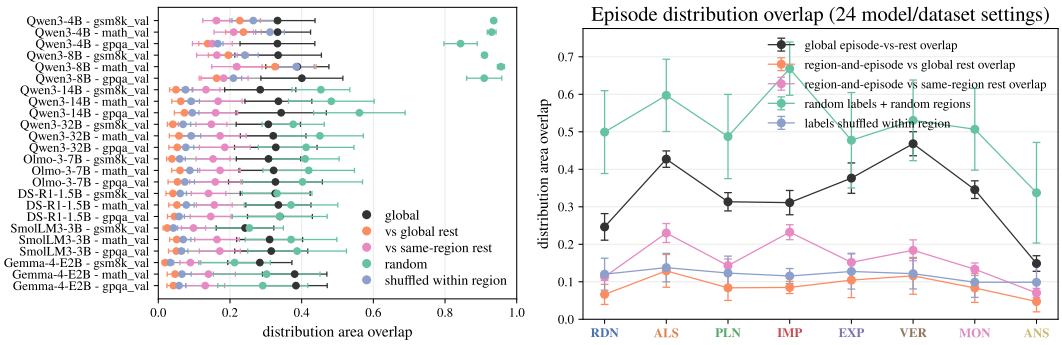


Figure 16: Measured as histogram overlap, episode categories also overlap less within a single region than across the whole space. Area overlap between the projected distributions of an episode category and the remaining episodes, with the setup of Figure 3. The estimate is biased towards zero for small samples such as single regions.

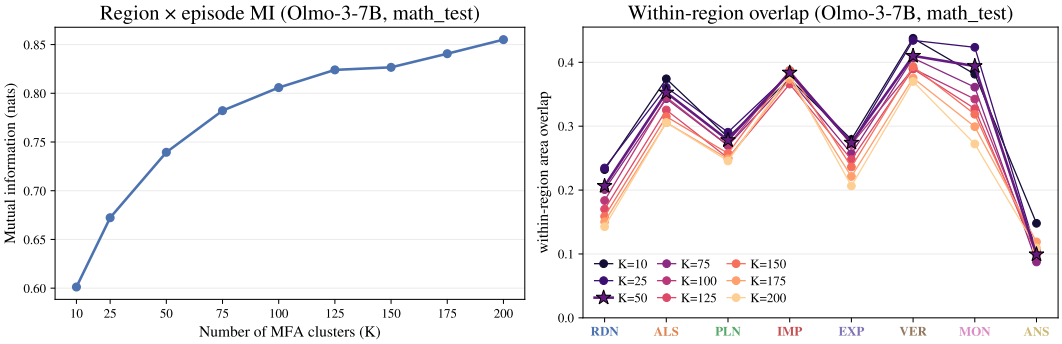


Figure 17: **Left:** mutual information between region and episode category rises with the number of MFA components K, with diminishing returns beyond $K \approx 100$. **Right:** within-region overlap decreases with K, for every episode category. Both panels show the effect of the number of MFA components K for **`Olmo-3-7B`** on MATH at relative depth 0.5 (test split).

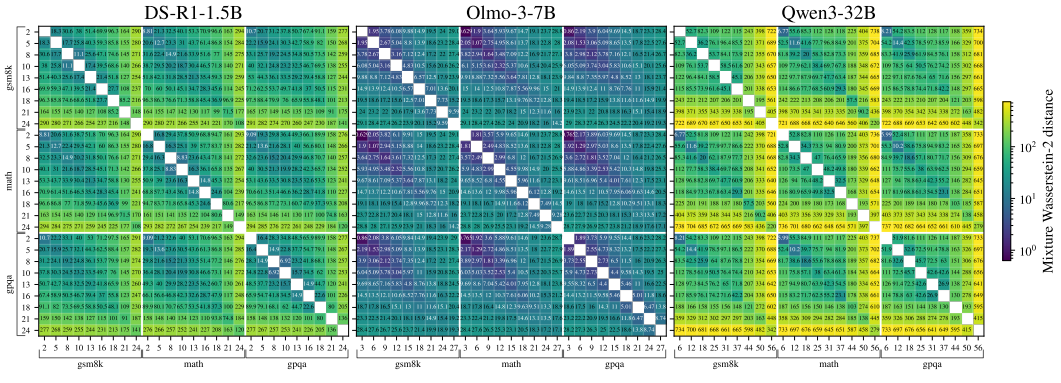


Figure 18: MFAs identify similar regions across different tasks, but they progressively diverge with depth. Mixture Wasserstein-2 distance (Delon & Desolneux, 2020) between MFAs trained on activations from different layers and datasets.

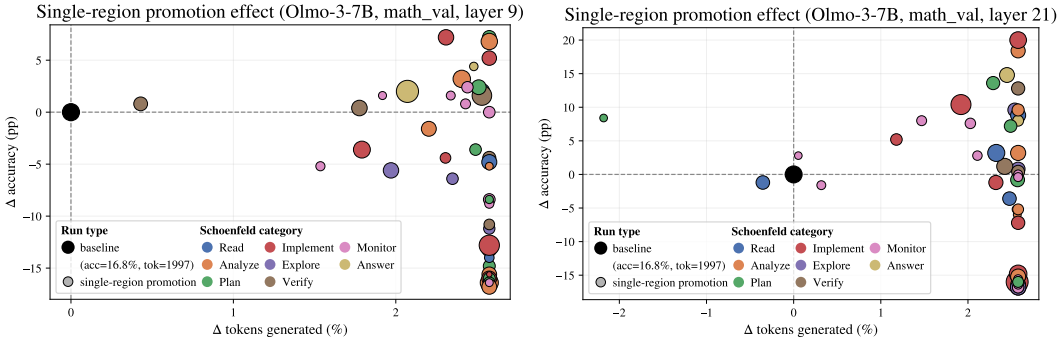


Figure 19: For **Olmo-3-7B** on MATH, promoting a small subset of regions also improves accuracy at $\ell \in \{0.3, 0.7\}$. Single-region promotions for **Olmo-3-7B** on MATH (validation split) at relative depth **Left:** $\ell = 0.3$ and **Right:** $\ell = 0.7$. Colors and dot sizes as in Figure 2.

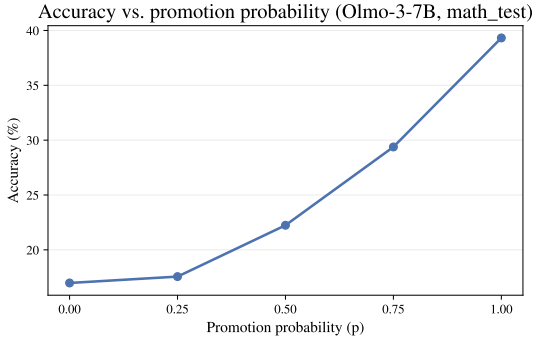


Figure 20: Accuracy increases monotonically with the frequency of P ACC interventions. Accuracy of **Olmo-3-7B** on MATH as a function of intervention probability p.