CLINICAL CONCEPT CENTERS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used in clinical settings. However, research into the reliability and performance of these models has focused almost entirely on the language substrate, scoring what the model says, while the latent computations that produce these answers stay unexamined. At the same time, mechanistic interpretability has found that the latent space carries a higher fidelity of representation than the text: internal representations not only encode substantially more than the output verbalizes, but the stated reasoning also systematically omits features that causally drive the answer. An evaluation of model behavior in terms of mechanistic interpretability has not been explored in clinical decision support. We believe that it is a necessity, not only for improving model performance, but also to evaluate the safety and alignment of these models as medical devices. In this work, we extend behavioral evaluation into the latent space and ask whether clinical concepts exist as locatable, causally used representations inside open-weight LLMs. Reliable mechanistic interpretability in such settings uses dictionary learning over a labeled text corpus. We reformulate the EHR setting into exactly this approach, with clinical notes as the corpus and expert-assigned ICD billing codes as clinically validated proxies for automated learning of clinical concept centers. We find dedicated clinical concept centers in the latent space of all eleven open models we test. These concept centers are interpretable, firing only on their aligned clinical narratives, and meaningfully and causally drive model behavior in both constrained and open-ended settings. These concept centers are not just analytical representations, but circuits that can be utilized in clinical practice, and we explore their use from the perspective of both evaluation and performance. From the evaluation standpoint, models stay internally coherent and keep using the relevant concept centers even under adversarial role-based priming, while aligned priming improves downstream clinical performance. From a performance perspective, we simulate realistic deployment settings and find that steering models along these centers leads to meaningful downstream improvements. Finally, we conduct a blinded clinician validation and find the activation and usage of these concept centers predicts clinicians preferences.

1 INTRODUCTION

Large language models are increasingly being used in clinical workflows, across settings from medical code prediction (Mullenbach et al., 2018) and question answering over clinical documents (Kweon et al., 2024) to open-ended health conversations (Arora et al., 2025) and inpatient clinical decision support (Nagar et al., 2026), with open medical foundation models released explicitly for clinical application development (Sellergren et al., 2025) and instruction-tuned models whose answers physicians have rated above physician-written ones on most clinical-utility axes (Singhal et al., 2023; 2025). The use of these models as a medical device demands a guarantee that when the model works through a medical problem, it does so in a way that is aligned with how clinicians reason about it, with no implicit computations in the latent space biasing its downstream suggestions. Behavior-oriented benchmarks alone cannot tell us when a model can be trusted. Despite performing well on evaluations in the language substrate, deployed LLMs produce no significant gain in physicians' diagnostic reasoning in a randomized trial (Goh et al., 2024), and models which perform well on exams fall short of clinicians on realistic EHR decisions (Hager et al., 2024).

However, nearly all evaluation of LLMs in clinical settings happens on the language substrate, scoring what the model says as exam answers (Jin et al., 2021; Pal et al., 2022) or physician-rubric grades

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.

over generated conversations (Arora et al., 2025). This is not a faithful window into the model's latent computation. Internal representations encode substantially more than the output verbalizes, to the point that a model can internally encode a correct answer while generating a wrong one (Orgad et al., 2025), and stated chain-of-thought reasoning systematically omits the features that causally drive the answer (Turpin et al., 2023; Lanham et al., 2023). Because these biases or shortcuts may remain hidden in the LLM's latent representations, output-level evaluations alone are insufficient to detect or prevent them, particularly in clinical settings. Interpretability of clinical models has itself largely stayed on the behavioral side, analyzing the language substrate, and mechanistic work in the latents is only beginning (§2). No such work exists that identifies expert-label-grounded clinical concepts inside the model and tests whether they causally influence the model's clinical behavior.

Finding such representations, which we call *clinical concept centers*, has a twofold advantage. From the evaluation perspective, black-box access is insufficient for rigorous audits because root-cause diagnosis needs the internals (Casper et al., 2024), regulators are moving oversight toward transparency about how a deployed model decides (Meskó & Topol, 2023), and our position is that for a model used as a medical device, evaluation in the mechanistic space is as necessary as evaluation in the language substrate. From the performance perspective, reliable concept representations can serve as steerable vectors. Prior work has already shown that interventions on internal representations edit and steer behavior without retraining (Meng et al., 2022; Zou et al., 2023; Rimsky et al., 2024), reinforcement-learning post-training mostly amplifies reasoning paths already present in the base model (Yue et al., 2025), and there is an increasing body of work that uses mechanistic methods to find the reasoning circuits that develop during such post-training and introduce them into models at a fraction of the training cost, though so far by brute-force feature selection (Wang et al., 2025). A well-informed mechanistic approach for improving model capabilities in clinical settings requires the right concept centers to be identified and confirmed to causally influence model behavior, which further motivates this study.

So the question becomes whether clinical concept centers even exist in these models, and how to locate them. Whatever the instrument for interpretable approaches over model latents, from dictionary learning (Templeton et al., 2026; Cunningham et al., 2024) to latent probes (Alain & Bengio, 2016) to direction extraction on the raw residual stream (Marks & Tegmark, 2024), they are most reliable when they run over a text corpus with annotated labels, so that the recovered representations can be named by what they fire on. Automated interpretability can generate such labels with an LLM in the loop, but the generated explanations carry an error signal of their own (Singh et al., 2023; Paulo et al., 2024), which we want to avoid. Healthcare data already maintains this kind of labeled composition. The International Classification of Diseases (ICD) is the standard vocabulary in which diagnoses are recorded, and after essentially every patient encounter a trained medical coder reads the clinical documentation and assigns the codes describing what the encounter was about. The codes form a hierarchy, from thousands of highly specific leaf codes up to broad chapters corresponding to organ systems (circulatory, respiratory, musculoskeletal, and so on), and although their day-to-day purpose is record-keeping and billing, each assignment is an expert's judgment of the diagnostic concepts a note encodes, made at scale as routine clinical practice. Clinical ontologies like the ICD hierarchy are therefore established, expert-maintained terminologies, often used for representation learning in clinical foundation models (Choi et al., 2017). EHR data such as discharge notes already carry these associated codes (Johnson et al., 2023b;a), and the automated-coding literature uses exactly this setting as its ground truth (Mullenbach et al., 2018). We reformulate this EHR setting into the labeled-corpus recipe, with clinical notes as the text corpus and ICD codes as clinically validated labels, so instead of training a model on the ontology, as the foundation-model line does, we take an existing artifact of knowledge and ask whether its latent space already contains these representations. This formulation provides expert-annotated supervision at scale with no labeling effort, and has not been used for model interpretability before, which makes the formulation itself one of the contributions of this work.

Thus, as the foundation, we establish existence, specificity, causal use, and temporal persistence of these concept centers across eleven open-weight Qwen and Gemma models (4B–30B, dense and mixture-of-experts, base, instruction-tuned, and medical variants) (§3). We investigate the use of these concept centers in latent-space evaluation, performance improvements in clinical tasks and for predicting expert preference during clinician validation (§7). This leads to three key contributions:

- **A novel, clinically grounded formulation for latent-space interpretability.** Reformulating EHR notes with expert-assigned ICD codes into a labeled corpus for dictionary learning keeps the probing setting realistic, provides expert-annotated supervision at scale without any labeling effort, and avoids the error signal that automated labeling introduces. The same recipe extends to any domain that maintains an established ontology over its records.
- **Dedicated, separable clinical concept centers in the latent space of open models.** We find the concept centers recoverable in all 11 models tested (best held-out single-feature AUC 0.80–0.92 in the Qwen family), responding to the clinical situation of a note rather than to its vocabulary, and transferring to other clinical settings with 45–86% of excess-over-chance separability retained. These concept centers are used across the depth of the models and persistently through the generation of text for clinical support (§3).
- **A meaningful demonstration of their use for evaluation and performance, validated with clinicians.** Counterfactual role-based priming becomes a latent-space evaluation of whether a model follows its assigned role or its own clinical concepts, showing that the models remain internally coherent under adversarial roles, while aligned priming improves accuracy by $+2.9$ to $+11.2$ points across both model families (§5). Steering the models along the concept centers under a realistic prior improves downstream clinical performance ($+1.1$ to $+2.3$ additional correct diagnoses per hundred cases on the Qwen models). A blinded clinician evaluation shows that the strength of the concept-center activations predict clinician preference, with a 76% agreement on the largest activation gaps, regardless of how the model was prompted (§6–§7).

2 Related work

Clinical LLMs and clinical interpretability. Medical foundation models and instruction-tuned generalists reach expert-level scores on knowledge benchmarks (Singhal et al., 2023; 2025; Sellergren et al., 2025). Evaluation has followed the outputs, from exam-style question answering (Jin et al., 2021; Pal et al., 2022; Kweon et al., 2024) to physician-rubric grading of conversations (Arora et al., 2025), and benchmark performance still fails to translate into clinician benefit or reliable decisions (Goh et al., 2024; Hager et al., 2024). Interpretability has entered this domain only at the edges, decomposing radiology vision encoders (Abdulaal et al., 2024), adapting dictionaries to clinical text (O'Neill et al., 2025), examining reasoning variability (Modi et al., 2026), and localizing triage failures to output formatting (Fraile Navarro et al., 2026). Understanding clinical model behavior has otherwise stayed on the behavioral side, analyzing the language substrate, and no existing latent-space work grounds the recovered features in expert-assigned diagnostic labels and tests them causally against the model's clinical behavior.

Dictionary learning, concept directions, and their critics. Concepts are often carried by linear directions (Park et al., 2024), difference-of-means directions are causally implicated carriers (Marks & Tegmark, 2024), activation-level interventions steer behavior without retraining (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023; Meng et al., 2022), and editing effectiveness dissociates from where probing localizes a concept (Hase et al., 2023), which our results confirm in the clinical domain. A skeptical line reports that SAE probes underperform logistic regression (Kantamneni et al., 2025), that prompting and difference-of-means baselines beat SAE steering and detection (Wu et al., 2025), and that SAE-internal metrics do not guarantee downstream utility (Karvonen et al., 2025). We treat these as design constraints, keep the probe and the raw direction as baselines everywhere, follow the control-task logic of Hewitt & Liang (2019) and the erasure machinery of Belrose et al. (2023) on the removal side, and report the one place our data contradicts the skepticism, the SAE-decoder direction being the stronger causal lever on this task (§6). Appendix A extends this section with the dictionary-caveat and steering-reliability lines that inform our controls.

3 Method

Data and labels. We use MIMIC-IV discharge summaries (Johnson et al., 2023b;a). Each summary contains ICD-9 diagnosis codes assigned by an expert medical coder. Two properties of these codes need consideration. Their ontology is extremely granular, the representations of model latents may not plausibly resolve concepts at that level of granularity, and the codes assigned to an encounter

often include entries copied over from past occurrences or added for billing purposes, which adds its own source of error to the labels. We therefore use only the primary code of each note and group codes at the level of the 17 ICD-9-CM disease chapters, which are broad organ-system categories that behave like department specialties (mapping and counts in Appendix C). After patient-level deduplication and capping at 1,000 notes per chapter, the corpus holds 8,617 notes, split by patient into a selection half used for all direction estimation and hyperparameter choices and an evaluation half on which every reported number is computed (models of 27B parameters and above use a 4,334-note subset with the same balance properties). We study 11 open-weight models across two families, five Qwen models spanning base, chat, medical fine-tuned, and mixture-of-experts variants (8B–30B), and six Gemma models spanning pretrained, instruction-tuned, and medical variants (4B–27B), with dictionaries from the open Qwen and Gemma SAE suites (Deng et al., 2026).

Constructing a concept center. Let $\boldsymbol{h}_\ell(x) \in \mathbb{R}^d$ be the residual-stream activation of note x at layer ℓ, mean-pooled over the note's tokens (the final token of a decoder-only model encodes the next-token prediction rather than a summary of the input, so pooling is the faithful read of what the note activated). An SAE encodes $\boldsymbol{h}$ into sparse features $\boldsymbol{z} = f_{\text{enc}}(\boldsymbol{h}) \in \mathbb{R}^m$, $m \gg d$, and decodes $\hat{\boldsymbol{h}} = \boldsymbol{W}_{\text{dec}}\boldsymbol{z} + \boldsymbol{b}_{\text{dec}}$ (TopK activation for the Qwen suite, JumpReLU (Rajamanoharan et al., 2024) for the Gemma suite). For chapter C we form the difference of mean sparse codes between selection notes inside and outside the chapter and decode it into model coordinates,

$$\boldsymbol{v}_C \;=\; \boldsymbol{W}_{\text{dec}}\Big(\tfrac{1}{|S_C|}\textstyle\sum_{i\in S_C}\boldsymbol{z}_i \;-\; \tfrac{1}{|S_{\neg C}|}\sum_{i\in S_{\neg C}}\boldsymbol{z}_i\Big), \tag{1}$$

which is the mass-mean construction of Marks & Tegmark (2024) adapted into the SAE feature basis, so the direction is built from the sparse features that discriminate the chapter and expressed in the coordinates the model natively computes in. As baselines we keep the raw difference-of-means direction $\boldsymbol{v}_C^{\text{raw}}$ computed directly on $\boldsymbol{h}$ (the strongest simple baseline in the SAE-skepticism evaluations (Wu et al., 2025)) and a multinomial logistic probe over $\boldsymbol{h}$ at every layer (five-fold cross-validated on the selection half), which upper-bounds what any linear read of that depth can achieve (Alain & Bengio, 2016). We then use the projection $\text{proj}(\boldsymbol{h}, \boldsymbol{v}_C) = \boldsymbol{h} \cdot \boldsymbol{v}_C/\|\boldsymbol{v}_C\|$ as the basic measurement for the study.

Probing for concepts centers. We compare the three most common instrumentation methods to probe the latent space of the models. The logistic probe recovers the most separability per layer, but it is a d-dimensional classifier with no interpretability, conflates what is encoded with what the model uses (Belinkov, 2022; Hewitt & Liang, 2019), and information a probe finds can be behaviorally inert (Elazar et al., 2021), so we treat it as a non-interpretable calibration baseline for the other two instrumentation methods. The raw difference-of-means direction is the strongest simple carrier of a concept (Marks & Tegmark, 2024; Wu et al., 2025) and needs no dictionary. The SAE direction of Equation 1 is the most sophisticated, but studies have suggested that it may not be as useful (Kantamneni et al., 2025). However, as our experiments will show, the SAE turns out to be the better instrument for our setting, for three reasons. It decomposes the concept center into named, interpretable monosemantic features whose per-token firings can be monitored during generation (§4.4, §7), a single such feature nearly matches the probe ceiling (§4.1), and the SAE-derived direction is empirically the stronger causal axis than the raw direction on the same models and layers (§6).

4 The foundational setting: Clinical concept centers

For a reliable mechanistic analysis of models, we require two things: *decodability*, meaning the concept can be read out of the latent space as a unique, distinctly separable representation (Alain & Bengio, 2016; Park et al., 2024), and *causal impact*, meaning the representation demonstrably shapes the model's outputs and behaviors (Meng et al., 2022; Hase et al., 2023). To build an informed foundation for clinical concept centers, we investigate the questions outlined in this section as the first step in methodically establishing their existence in the latent space of LLMs.

4.1 Do concept centers exist, and across which architectures and scales?

Clinical concept centers exist in every model, architecture and scale we test. Figure 1 plots the held-out one-vs-rest AUC of the best single SAE feature per chapter against depth, and the profile is

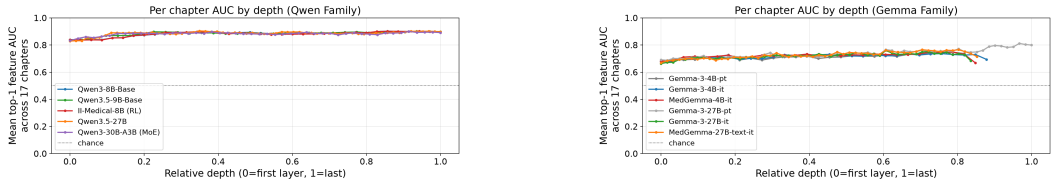

Figure 1: Clinical concept centers exist across architectures and scales. One-vs-rest AUC of the best single SAE feature representing the clinical concept on the held-out set, against relative depth (left, five Qwen-family models; right, six Gemma-family models). The dashed line represents random chance. Gemma curves end before full depth where the JumpReLU dictionaries produce non-finite codes, an SAE instrument limit (only the Gemma-3-27B-pt dictionary stays finite to the last layer).

the same in both families. Separability is nearly uniform across depth, starting from the the earliest layers and holding at almost the same level through the whole range the dictionaries cover. At its best depth, the best single feature reaches AUC 0.80–0.92 in Qwen-family and 0.70–0.90 in Gemma-family models, against a full-rank logistic-probe ceiling of 0.90–0.93 macro AUC on ten of eleven models. Thus, one monosemantic dictionary feature can represent most of what a strongest linear instrument (the linear probe) at that depth can find. Selection optimism is below 0.009 everywhere and a stricter patient-level split leaves held-out AUC above 0.877 (Appendix E), so this separability goes beyond just in-sample selection or within-patient correlation. This near-uniform profile points to a consistent, reusable circuit for the chapter concept through the whole network.

These concept centers have robust properties. They are not determined by the dictionary that found them. We confirmed this with a fresh SAE trained on the medical fine-tuned II-Medical-8B (97.4% variance explained, 0.11% dead features), which recovers chapter directions whose residual-space cosine with the parent dictionary's is $+0.835$ on average and positive on all seventeen chapters. And they are generalizable beyond the data they were discovered with, since applied without refitting to patient-question text they retain 45–86% of excess-over-chance separability across the Qwen family (all permutation $p < 1/2{,}000$, details provided in Appendix H).

4.2 Are the recovered features specifically clinical?

High separability could still come from cheap places like note length, lexical density, or the mere presence of disease names, so before any causal work we ask what the features respond to. To measure specificity, we take the ten highest-AUC features for each concept center and count how many activate under five conditions of decreasing clinical content: the full discharge note, the bare diagnosis text (an isolated label such as "acute myocardial infarction"), the chapter title text (e.g., "Diseases of the Circulatory System"), discharge notes from the adjacent chapter in the ICD-9 ordering (e.g., Respiratory notes probed with the Circulatory chapter's features), and non-clinical control sentences (short texts about geography, sports, and food). On full notes, all ten fire by design. On bare diagnosis labels, activation drops to 2.0–2.7 of ten across Qwen models, on chapter titles to 1.3–2.9, on adjacent-chapter notes to 0.3–0.6, and on non-clinical text to at most 0.016 (Appendix F). Thus these **features use the contextual clinical narrative (not just an isolated concept in the text), discriminate even between neighboring diagnostic categories, and do not activate outside the clinical domain**. This confirms the first requirement we set for a clinical concept center: it is specific, distinct, and unique to its clinical concept.

4.3 Does the model causally use the concept centers for its decisions?

Existence and specificity are representational facts, and representational facts can be inert (Elazar et al., 2021; Hase et al., 2023), so the foundation needs a causal test. We define a seventeen-way ICD classification task as a direct test for the causal effect of these centers on the models. We ask models to classify the principal ICD chapter, with chapters mapped to answer letters (Appendix C). We add $\alpha\hat{\boldsymbol{v}}_C$ (the chapter direction) to the residual stream at the final token position, sweep α over $\{\pm1, \pm3, \pm10, \pm30, \pm100\}$ and ℓ over every layer and select the best (layer, strength) pair. Our analysis reports only the held-out effect of this configuration, to avoid selection bias.

Table 1: The concept centers are causally used across both families. Every row reports ΔP (the change in correct-letter probability) during classification at the configuration selected on the selection half of the notes, evaluated on the held-out half, with bootstrap 95% CIs (5,000 resamples). Chapter-specific is what remains of the injection effect after subtracting what matched-magnitude random directions achieve at the same layer and strength, thus isolating the contribution of the concept identity itself (Appendix G).

Model	Layer	Strength	ΔP	95% CI	Baseline P	Chapter-specific
Qwen3-8B-Base	23	+30	+0.0685	[+0.0652, +0.0719]	0.376	+0.044 (4 seeds)
Qwen3.5-9B-Base	19	+10	+0.0515	[+0.0475, +0.0555]	0.465	+0.033 (3 seeds)
Qwen3-30B-A3B-Base	32	+3	+0.0692	[+0.0645, +0.0739]	0.422	+0.094 (3 seeds)
Qwen3.5-27B	54	+30	+0.0406	[+0.0364, +0.0447]	0.581	+0.057 (3 seeds)
II-Medical-8B	24	+100	+0.0780	[+0.0695, +0.0867]	0.111	+0.064 (4 seeds)
Gemma-3-4B-pt	8	+100	+0.0058	[+0.0053, +0.0062]	0.099	+0.041 (3 seeds)
Gemma-3-4B-it	8	+100	+0.0154	[+0.0115, +0.0194]	0.443	+0.011 (3 seeds)
MedGemma-4B-it[§]	6	+100	+0.0057	[+0.0030, +0.0084]	0.483	−0.003 (3 seeds)
Gemma-3-27B-pt	5	+100	+0.0015	[−0.0005, +0.0036]	0.321	+0.033 (3 seeds)
Gemma-3-27B-it	2	+100	+0.0014	[−0.0005, +0.0036]	0.553	+0.000 (3 seeds)
MedGemma-27B-text-it	11	+100	+0.0193	[+0.0154, +0.0231]	0.496	+0.025 (3 seeds)

[§] 4B dictionary is borrowed from parent model, so its results may reflect the instrument; see Appendix G.

To attribute the steering effect to the clinical concept itself, we compare it against four controls. Each control keeps the layer, strength, held-out notes, and evaluation identical to the main experiment and alters a single component of it. Any effect a control reproduces can therefore be attributed to that component instead of the concept. The controls act at two stages of the pipeline: one alters how the concept center is *discovered*, and three alter the vector *injected* at inference time for steering.

Control at discovery time. Each concept center is the output of a multi-step discovery procedure: we select the features that best separate a chapter's notes from the rest (by AUC) and aggregate them into a unit direction. A procedure this elaborate could return a direction that improves classification even if no real concept underlies it, for instance by favoring features that are broadly active on clinical text. The *permuted-label control* (Hewitt & Liang, 2019) tests this possibility. Before discovery, we randomly shuffle the chapter labels across notes, so that each "chapter" label used for feature selection becomes an arbitrary group of notes with no shared diagnosis. We then rerun the identical discovery procedure on these groups. The resulting control direction shares everything with the concept center except its grounding in real diagnoses. If the concept is what drives the steering effect, this direction should produce little or no improvement.

Controls at inference time. The remaining three controls take the concept centers as given and replace the steering vector added to the residual stream. Each asks whether something other than the matching concept could produce the same gain. The ***wrong-chapter control*** keeps the notes fixed and swaps the direction. For a note whose true chapter is C, we inject the direction of a different chapter $C' \neq C$ and measure the probability of C's answer letter. If the model merely benefits from any push toward clinical content, a mismatched chapter direction should help as much as the correct one. However, if it relies on the *correct* concept, the mismatched direction should fail. The ***random-direction control*** injects random unit vectors at the same layer and strength, repeated over several seeds. Adding any vector of this magnitude perturbs the model, and this control measures how much of ΔP such a generic perturbation achieves. The chapter-specific column of Table 1 is the effect that remains after subtracting this floor. Finally, the ***unembedding control*** addresses the most common shortcut explanation for steering results. The model converts its final residual state into token probabilities through the unembedding matrix. The unembedding row of the correct answer letter is therefore the direction that most directly raises that letter's probability, without involving any clinical representation. If our chapter direction were a disguised version of this output shortcut, it would align with that row, and injecting the row would reproduce the effect.

Table 1 reports our findings. The held-out **steering effect is positive on all 11 models**, with no Qwen confidence interval touching zero, and every Gemma model moving in the predicted direction, so the causal use of the concept centers holds across architectures, scales, and dictionary families (TopK

and JumpReLU alike). Three properties of the effect show that it is the concept doing the work. **It is specific**, since matched-magnitude random directions achieve far less, and on the pretrained Gemma models they actively damage the model while the chapter direction helps. **It is bidirectional**, since negative strengths reduce the correct-letter probability, so the injection moves a representation the model consults rather than adding fluency. And it does not run through the output vocabulary, since the **chapter directions are essentially orthogonal to the answer letters' unembedding rows**, and injecting those rows does not reproduce the effect. The dose-response is an inverted U on every model, matching the pattern activation-steering work reports on general concepts (Rimsky et al., 2024). Details are provided in Appendix G.

What the sweep reveals about depth matters as much as the effect itself. Separability is nearly uniform across depth (§4.1), yet causal effectiveness depends on the location of steering. The best intervention layer does not coincide with the (marginally) most separable layer on any of the five Qwen models, and instead falls in the second half of the network (roughly 60–85% of depth, Table 1). The behavioral swing in this causally effective band is more than an order of magnitude larger than at other depths that read out the concept essentially equally well (0.283 against 0.022 on Qwen3-8B-Base, Appendix G). This confirms in the clinical domain the dissociation Hase et al. (2023) reported for factual editing, where localization does not predict where intervention works, and staged views of transformer depth suggest why. Early layers dissociate and integrate the input, the middle layers carry the complex computation over it, and the later layers govern how the model converts its latent computation into the response it presents to the user (Lad et al., 2024; Skean et al., 2025; Liu et al., 2025). The **concept centers being separable across this whole range** suggests that the models carry them through all of these functional stages, where they may support implicit computation or pass clinical information forward to later layers, even where an injection moves the output little. **The causal influence of an injection is largest where the model is deciding how to verbalize its latent computation**, which is why intervention there helps the most.

4.4 Do concept centers persist through autoregressive generation?

While the ICD prediction experiment establishes causal influence of the concept centers, real-world applications like monitoring a deployed system and steering an open-ended answer need the concept to stay active while text is produced. Transient circuits during token generation can be more akin to noise, and might not influence the quality of outputs during autoregressive generation. We therefore track the projection onto all 17 chapter directions at every token of long-form generation over held-out notes, 170 passes per model of up to 2,000 tokens, with base models continuing the note and chat models writing a prompted clinical assessment of it. Per token we reduce this to the target-chapter dominance, defined as the true chapter's projection minus the mean over all seventeen. Figure 2 shows the dominance as an effect size across passes at each model's most representative layer. **The dominance is positive at every generated position on every model, till end of the generation window**, with Qwen3.5-27B sustaining an effect size near 0.5 from the 100th token to the 2,000th, and the Gemma family showing the same sustained shape at smaller magnitudes. The concept center is therefore a persistent state of the generation, which is what makes per-token monitoring meaningful. Experimental details are provided in Appendix I.

5 The evaluation perspective: Counterfactual role priming

With the foundation in place we can build evaluations that read the latent space instead of only the text. The first is motivated by how clinical models are deployed, since system prompts routinely assign the model a professional identity, and personas change behavior in ways output-level evaluation summarizes only poorly (Salewski et al., 2023; Zheng et al., 2024). The question a mechanistic evaluation can ask is better posed. When the model is told it is a specialist, does it move the corresponding internal clinical concept, and when the assigned role conflicts with the patient in front of it, does the model follow the role or its own internal concepts? The counterfactual design makes this measurable, presenting each note under a matched clinical specialty (the specialty of the note's true chapter), a counterfactual specialty, a randomly drawn clinical specialty, a generalist physician, a non-clinical role, and no role, and measuring both the projection onto the true chapter direction and the downstream accuracy. Projections are read at each model's best causal layer from §4.3.

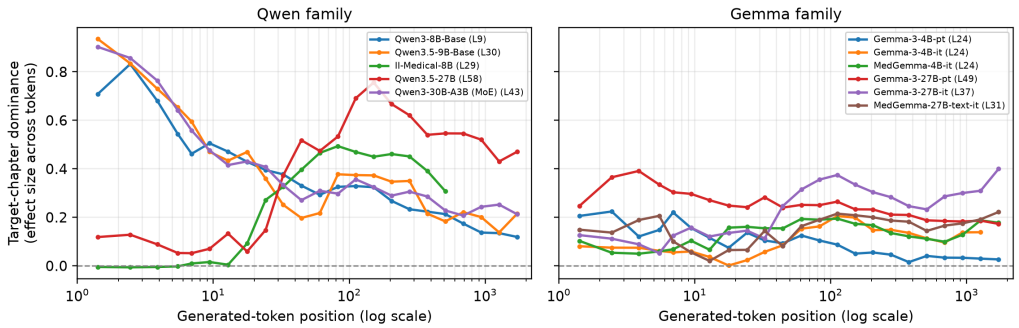

Figure 2: The concept centers persist through generation. Target-concept dominance per generated token as an effect size across 170 passes per model, log-binned out to 2,000 tokens (measurement details in Appendix I).

5.1 Do internal clinical concepts matter more than role based priming?

The matched specialty raises the projection onto the matching chapter direction (paired $t = 2.9$ to 9.5 against the counterfactual, across the ten models of both families) and converts the shift into accuracy, $+2.9$ to $+11.2$ points over the no-role baseline, while the counterfactual condition performs almost similar to the baseline (refer to Table 8 in Appendix J for per-model details). This shows that the **models are internally coherent under adversarial role priming** in this closed task. The wrong specialist role costs little to nothing because the model's own clinical concepts keep it in the right direction, while **the right specialist adds signal on top**. This pattern holds across both families, both scales, and both dictionary types. The six-condition results further show the effect is specialty identity and not clinical framing, because a randomly drawn clinical specialty does nothing (-0.08 projection shift, $p = 0.83$) while matched minus random-clinical is $+1.30$ ($t = +9.57$, $p = 2.3 \times 10^{-19}$, 14 of 17 chapters), the generalist role sitting between (Appendix J). Interestingly, this also rescues a degraded medical fine-tune whose internal chapter separation is intact while its output defaults to a single answer under constrained generation, which is further evidence for the depth picture of §4.3, since introducing the right concept signal at the causally effective depth changes what the model presents to the user (Appendix J).

6 The performance perspective: Model Steering

With the foundation in place and the evaluation angle showing the concept centers at work, can we also leverage them to improve downstream model performance itself? In real clinical deployments there is metadata: we usually know which department the system serves, which specialty the ordering clinician belongs to, and therefore which clinical concepts are relevant before any text arrives, so a realistic intervention is to upweight the matching concept center under that prior, and investigate whether doing so improves the model answer. We outline the two methodological choices for steering in Appendices L and K).

Deployment settings are open-ended and free-form, such as drafting a clinical assessment or advising on a case. We thus hold the injection active at every generated token while the model writes its own diagnosis for 1,624 held-out notes per model and score whether the generated text predicts the correct chapter. The result ordering aligns with our priors. **Steering along the true chapter direction helps the generation**, $+1.1$ to $+2.3$ additional correct diagnoses per hundred cases across the three models tested. **A random-direction ablation does not help, and steering with the wrong chapter hurts the model.** The holds true for all three models, spanning both families, and combines to $p = 0.00046$ (Fisher's method). Table 10 in Appendix L shows the full per-model results.

Table 2: Clinician preference and its agreement with the activation-strength prediction.

			Agreement		
Pairs	**Preferred**	**Gap**	**All**	**Top half**	**Top quarter**
All	—	.058	55.7 (.13)	66.0 (.002)	76.0 (3×10^{-4})
No role vs. adv.	no role 50–20	.082	64.3 (.022)	77.8 (.001)	—
Adv. vs. aligned	aligned 36–24	.057	53.3 (.70)	66.7 (.099)	—
Aligned vs. no role	no role 41–23	.048	48.4 (.90)	54.5 (.73)	—

7 CAN LATENT CONCEPT CENTERS PREDICT CLINICIAN PREFERENCE?

Automated evaluations like the ones above do not always capture the qualities of an output that a practicing clinician would find meaningful, so we conduct a blinded clinician evaluation to test how well our mechanistic findings align with the preferences of expert physicians. From HealthBench (Arora et al., 2025), which benchmarks clinical output quality for health conversations, we filter a specialty cohort matched to the Musculoskeletal chapter. Qwen3.5-27B, the strongest chat model in our pool, answers every case under three role-priming conditions: no role, a matched-specialty role, and an adversarial other-specialty role, with the per-token SAE firings and all-layer chapter projections of every generation recorded. Two clinical raters, a senior orthopaedic resident surgeon and a junior medical student, judge the three pairwise comparisons per case blinded, position-randomized, and with a tie option, giving 228 head-to-head judgments (full protocol, exclusions, and detailed analysis in Appendix M). For every response we reduce the per-token concept trace at the representation-peak layer to its mean target-chapter dominance,

$$\mathrm{dom} \;=\; \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\big(\mathrm{proj}_t(c^*) - \tfrac{1}{17}\sum_c \mathrm{proj}_t(c)\big)\,, \tag{2}$$

which is the average excess activation of the cohort's chapter c^* over the field of all chapters while the response was being written. We hypothesize that *within every pair, the response that used the aligned concept center the most is the one the clinicians find more meaningful.*

Role priming serves as a way to induce genuinely different behaviors on the same evaluation case so that the mechanistic signal has variation to predict. Table 2 reports both which condition the clinicians prefer and how well the mechanistic prediction agrees with them. The prompt-level priming turns out to be an unreliable guide to output quality, while the mechanistic analysis aligns far more closely with the clinicians. The matched-specialty role is not always preferred by the clinicians in open-ended use, **so role-based priming does not guarantee better quality**. Within each pair, we predicted, before any rating, that the clinicians prefer the response that activated the relevant concept center more strongly while it was being generated. Over all pairs, **as the activation strength of the aligned concept center increases, so does the agreement with clinician preference**, from 56% over all pairs to 66% on the half of pairs where the responses' activations separate more, and 76% on the quarter where they separate the most. The rise appears within each comparison type, so it is the activation and not the priming condition doing the predicting, and the one comparison where agreement stays at chance (aligned versus no role) is also the one with the smallest activation gaps. Thus, **the more strongly one response uses the relevant concept center than another, the more consistently clinicians prefer it**, while responses that use it to a similar degree offer no basis for a choice. This suggests that a mechanistic probe of the clinical concept centers can guide abstention and selection for pass-at-k behaviour in real deployment settings, and their usage in LLMs has a meaningful impact on the downstream quality of generated outputs. Notably, this also aligns with the internal coherence we find in §5, where linguistic role-based priming does not guide the downstream quality of the model's answers as much as its own use of the clinical concept centers.

8 CONCLUSION

We conduct a mechanistic study of clinical concept centers in LLMs. We find that they are locatable, specific, causally used by the models, persistent across generation and more effective than the linguistic substrate. We demonstrate rigorous evaluation and performance improvement of models as medical devices using these concepts, and validate it with clinician testing. Our study paves the way for future work in mechanistic interpretability of clinical AI, where the evaluation and usage of a medical model includes the latents in which it computes, not only the text in which it generates.

ETHICS STATEMENT

This work uses MIMIC-IV discharge summaries, which are de-identified and were accessed under PhysioNet credentialed access with the required training and data use agreement. No note text, patient information, or derived data that could support re-identification appears in the paper or the released materials, and the corpus is not redistributed. The HealthBench conversations used in the clinician evaluation are public benchmark data, we generated model responses to them without any training or fine-tuning on the benchmark, and its canary string was respected. The clinician evaluation collected blinded professional judgments about machine-generated text from two clinical raters who participated voluntarily and consented to the use of their anonymized ratings and comments, and no patient data or personal information about the raters was collected. All models studied are open-weight models used within their license terms. No output of this work is intended for clinical use without regulatory review.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code for every experiment in the paper, covering direction construction and its baselines, the specificity firing test, the injection sweep and all four controls, the role-priming conditions, the temporal traces, generation steering, and the clinician-study generation, blinding, and pre-registration pipeline, with a README describing the pipeline order and dependencies. MIMIC-IV requires PhysioNet credentialed access, and §3 with Appendices C and D describes the corpus construction, the patient-level splits, the chapter mapping, and the exact model and dictionary identifiers so the corpus can be rebuilt from the source data. The held-out selection protocol, seeds, and statistical conventions for the causal claims are given in §4.3 and Appendix G, the full clinician-study protocol including the sealed answer key, the frozen pre-registered predictions, and the capability-exclusion list is in Appendix M, and the curated rater workbooks are released with the data so every reported statistic in that section can be recomputed from the same filtered records.

AI USE STATEMENT

We used an LLM-based coding assistant, under continuous author direction, for tasks in both the required and recommended disclosure categories. In the required category, the assistant implemented parts of the experiment and evaluation code, executed the experiment sweeps on our hardware, performed the statistical computations reported in the paper and reformatted the MIMIC-IV notes into the labeled corpus described in §3. All generated code was reviewed by the authors. We have not used generative AI to generate synthetic datasets, to formulate or prove mathematical claims, or to propose the hypotheses under test. We reviewed all AI-assisted work. All experiment code was reviewed by the authors and is included in the supplementary material, every reported number was traced to a stored experiment artifact, all citations were verified against the cited papers, and the clinician study was designed, pre-registered, and adjudicated by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Ahmed Abdulaal, Hugo Fry, Nina Montña-Brown, Ayodeji Ijishakin, Jack Gao, Stephanie Hyland, Daniel C. Alexander, and Daniel C. Castro. An X-Ray is worth 15 features: Sparse autoencoders for interpretable radiology report generation. *arXiv preprint arXiv:2410.03334*, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. Saelens. `https://github.com/decoderesearch/SAELens`, 2024.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 2254–2272, 2024.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 787–795, 2017.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, et al. Qwen-scope: Turning sparse features into development tools for large language models. *arXiv preprint arXiv:2605.11887*, 2026.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Joshua Engels, Logan Riggs, and Max Tegmark. Decomposing the dark matter of sparse autoencoders. *arXiv preprint arXiv:2410.14670*, 2024.

David Fraile Navarro, Berardino Como, Jialei Sheng, and Soundariya Ananthan. Internal representation, not clinical knowledge: Where apparent LLM triage failures originate. *arXiv preprint arXiv:2605.29889*, 2026.

Jack Gallifant, Shan Chen, Kuleen Sasse, Hugo Aerts, Thomas Hartvigsen, and Danielle S. Bitterman. Sparse autoencoder features for classifications and transferability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joseph A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969, 2024.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024.

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 (NeurIPS)*, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*, January 2023a. doi: 10.13026/1n74-ne17. URL `https://doi.org/10.13026/1n74-ne17`. Version 2.2.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023b.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. *arXiv preprint arXiv:2502.16681*, 2025.

Adam Karvonen, Can Rager, Samuel Marks, Neel Nanda, et al. Evaluating sparse autoencoders on targeted concept erasure tasks. *arXiv preprint arXiv:2411.18895*, 2024.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Arthur Conmy, Callum McDougall, et al. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanhyung Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, 2024.

Vedang Lad, Wes Gurnee, and Max Tegmark. The remarkable robustness of LLMs: Stages of inference? In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

William Lehn-Schiøler, Magnus Ruud Kjær, Rahul Thapa, Magnus Guldberg Pedersen, Anton Mosquera Storgaard, Nick Williams, Radu Gătej, Tue Lehn-Schiøler, Andreas Brink-Kjær, Sadasivan Puthusserypady, Sándor Beniczky, James Zou, and Lars Kai Hansen. Mechanistic interpretability of EEG foundation models via sparse autoencoders. *arXiv preprint arXiv:2605.13930*, 2026.

Qi Liu, Haozhe Duan, Jiaxin Mao, and Ji-Rong Wen. How do large language models understand relevance? a mechanistic interpretability perspective. *ACM Transactions on Information Systems*, 44(2):34:1–34:27, 2025.

Michail Mamalakis, Tiago Azevedo, Cristian Cosentino, Chiara D'Ercoli, Subati Abulikemu, Zhongtian Sun, Richard Bethlehem, and Pietro Lio. A monosemantic attribution framework for stable interpretability in clinical neuroscience transformer-based language models. *arXiv preprint arXiv:2601.17952*, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bertalan Meskó and Eric J. Topol. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, 6(1):120, 2023.

Mirage Modi, Jordan E. Krull, David Johnson, Xinyu Wang, Timothy D. Gauntner, Ming Li, Hao Cheng, Anjun Ma, Ping Zhang, Daniel G. Stover, Zihai Li, and Qin Ma. Understanding clinical reasoning variability in medical large language models: A mechanistic interpretability study. medRxiv 2026.01.26.26344845, 2026.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 1101–1111, 2018.

Aishik Nagar, Arun-Kumar Kaliya-Perumal, Yu-Hsuan Han, Andrew Sheng-Han Huang, Kristen Kee, Yushi Cao, Yiming Chen, and Hongchao Jiang. CLR-voyance: Reinforcing open-ended reasoning for inpatient clinical decision support with outcome-aware rubrics. *arXiv preprint arXiv:2605.09584*, 2026.

Neel Nanda and Joseph Bloom. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`, 2022.

Charles O'Neill, Mudith Jayasekara, and Max Kirkby. Resurrecting the salmon: Rethinking mechanistic interpretability with domain-specific sparse autoencoders. *arXiv preprint arXiv:2508.09363*, 2025.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pp. 248–260, 2022.

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)*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2024.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Rohin Shah. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15504–15522, 2024.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Chandan Singh, Aliyah R. Hsu, Richard Antonello, Shailee Jain, Alexander G. Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, 2025.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 139179–139212, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. In *arXiv preprint arXiv:2308.10248*, 2023.

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 (NeurIPS)*, 2023.

Shangshang Wang, Julian Asilis, Ömer Faruk Akgül, Enes Burak Bilgin, Ollie Liu, Deqing Fu, and Willie Neiswanger. Resa: Transparent reasoning models via SAEs. *arXiv preprint arXiv:2506.09967*, 2025.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.13837.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Extended related work

This appendix extends §2 with the lines of work that further shaped our controls and design choices.

SAE features as practical instruments. Gallifant et al. (2025) evaluate SAE features as classifier inputs for safety-critical text classification and find them competitive and transferable across models and languages, which is the supervised regime where we, consistent with Kantamneni et al. (2025), also expect no SAE advantage over a full-rank probe. Movva et al. (2025) use SAE features over text embeddings to generate natural-language hypotheses that predict labels, the closest use to ours in spirit, since both treat the dictionary as a discovery instrument over a labeled corpus. Our formulation differs by grounding the labels in an expert-maintained clinical ontology rather than generating them, and by testing the recovered representations causally rather than correlationally.

Known caveats of sparse dictionaries. Sparse dictionaries absorb sub-concepts into broader features (Chanin et al., 2024), leave a share of the signal unexplained (Engels et al., 2024), learn different features when retrained on the same data (Paulo & Belrose, 2025), and are themselves evaluated through targeted concept-erasure tasks (Karvonen et al., 2024). These caveats motivated three of our design choices, the coarse chapter-level concept granularity (absorption pressure is highest for fine-grained concepts), the dense all-layer raw-residual projections kept alongside every SAE trace (so signal the dictionary misses is not lost), and the dictionary-independence check of §4.1, where a freshly trained dictionary recovers the same chapter directions as the borrowed parent one.

Reliability of steering vectors. Tan et al. (2024) show that steering vectors generalize unreliably, with high variance across inputs and settings, which motivated our relative-dose rule, the mandatory discordance counts before any null is believed (Appendix L), and the multi-seed random controls reported next to every intervention effect. Marks et al. (2025) scale circuit-level analysis with SAE features, which is the natural next granularity below our chapter-level directions and the instrument we expect to need when moving from chapters to finer clinical concepts.

Interpretability of general purpose LLMs. Recent applications of sparse autoencoders (SAEs) in the clinical domain have largely been used for interpretability studies for specific architectures or specific clinical sub-domains (Mamalakis et al., 2026; Lehn-Schiøler et al., 2026). While these studies successfully decompose the models across these specialised settings, their findings do not readily generalise to broader foundation models. We bridge this generalization gap by investigating general purpose large language models, specifically the Qwen and Gemma families, demonstrating that clinically relevant concept centers natively emerge within their latent spaces. Furthermore, we establish the practical utility of these concept centers by leveraging them to improve ICD-9 chapter classification accuracy and align model outputs with clinician preferences.

B Compute and toolkit used

Our experiments run on PyTorch (Paszke et al., 2019) over HuggingFace Transformers checkpoints (Wolf et al., 2020), with Gemma dictionaries loaded through SAELens (Bloom et al., 2024) and hook-level activation access in the style of TransformerLens (Nanda & Bloom, 2022). Experiments were carried out on a combination of Nvidia H100 and A100 GPUs.

C Prompt template and chapter mapping

This appendix fixes the classification task used throughout §4.3 and §5. The prompt template is shown below. The discharge summary text includes all sections as they appear in MIMIC-IV-Note, including the Discharge Diagnosis and Discharge Condition sections, with no truncation or filtering of the clinical narrative.

```
You are a clinical coding assistant. Read the discharge
summary below and identify which ICD-9 disease chapter
the principal diagnosis belongs to.
```

```
Discharge Summary:
{NOTE_TEXT}

Question: Which of the following ICD-9 chapters does the
principal diagnosis fall under?

A. Infectious and Parasitic Diseases (001–139)
B. Neoplasms (140–239)
C. Endocrine, Nutritional, Metabolic, and Immunity (240–279)
D. Diseases of Blood and Blood-Forming Organs (280–289)
E. Mental Disorders (290–319)
F. Diseases of the Nervous System (320–389)
G. Diseases of the Sense Organs (390–459)
H. Diseases of the Circulatory System (460–519)
I. Diseases of the Respiratory System (520–579)
J. Diseases of the Digestive System (580–629)
K. Diseases of the Genitourinary System (630–679)
L. Complications of Pregnancy, Childbirth, Puerperium (680–709)
M. Diseases of the Skin and Subcutaneous Tissue (710–739)
N. Diseases of the Musculoskeletal System (740–759)
O. Congenital Anomalies (760–779)
P. Certain Conditions Originating in Perinatal Period (780–799)
Q. Symptoms, Signs, and Ill-Defined Conditions (800–999)

Answer (letter only):
```

Decoding parameters. Temperature 0.0, top-p 1.0, maximum 10 new tokens. Context length is 8,192 tokens for all models except Gemma-3-27B and Qwen3.5-27B (32,768). Notes exceeding the context length are truncated from the end.

Table 3 lists the seventeen ICD-9-CM disease chapters, their code ranges, and per-chapter counts in the full 8,617-note manifest after per-chapter capping at 1,000.

Letter	Chapter name	ICD-9 range	Count
A	Infectious and Parasitic Diseases	001-139	487
B	Neoplasms	140-239	892
C	Endocrine, Nutritional, Metabolic	240-279	743
D	Blood and Blood-Forming Organs	280-289	312
E	Mental Disorders	290-319	651
F	Nervous System	320-389	568
G	Sense Organs	390-459	89
H	Circulatory System	460-519	1000
I	Respiratory System	520-579	1000
J	Digestive System	580-629	897
K	Genitourinary System	630-679	734
L	Pregnancy, Childbirth, Puerperium	680-709	124
M	Skin and Subcutaneous Tissue	710-739	203
N	Musculoskeletal System	740-759	456
O	Congenital Anomalies	760-779	78
P	Perinatal Conditions	780-799	198
Q	Symptoms and Ill-Defined Conditions	800-999	785
Total			**8,617**

Table 3: ICD-9-CM disease chapters, code ranges, and per-chapter sample sizes in the full manifest. The majority-class baseline (Circulatory and Respiratory at 1,000 notes each) is 11.6%.

D MODELS AND DICTIONARIES

Table 4 gives the exact model identifiers and SAE specifications for the eleven models of §3. The models span 34 to 64 layers and residual dimensions of 2,560 to 5,376. The MedGemma variants

use the dictionaries trained on their Gemma-3 parents (the borrowed-parent configuration discussed in §4.1), and the fresh dictionary for II-Medical-8B was trained on that model's own activations at layer 24 with the parent architecture (TopK, 65,536 features, $K = 100$), reaching 97.4% explained variance with 0.11% dead features.

Model	HuggingFace ID	SAE source	Dictionary size
Qwen3-8B-Base	Qwen/Qwen3-8B-Base	Qwen/SAE-Res-Qwen3-8B-Base-W64K-L0_100	65,536
Qwen3.5-9B-Base	Qwen/Qwen3.5-9B-Base	Qwen/SAE-Res-Qwen3.5-9B-Base-W64K-L0_100	65,536
II-Medical-8B	internlm/II-Medical-8B	Qwen/SAE-Res-Qwen3-8B-Base-W64K-L0_100	65,536
Qwen3.5-27B	Qwen/Qwen3.5-27B	Qwen/SAE-Res-Qwen3.5-27B-W80K-L0_100	81,920
Qwen3-30B-A3B-Base	Qwen/Qwen3-30B-A3B-Base	Qwen/SAE-Res-Qwen3-30B-A3B-Base-W128K-L0_100	131,072
Gemma-3-4B-pt	google/gemma-3-4b-pt	gemma-scope-2-4b-pt-res-all	16,384
Gemma-3-4B-it	google/gemma-3-4b-it	gemma-scope-2-4b-it-res-all	16,384
Gemma-3-27B-pt	google/gemma-3-27b-pt	gemma-scope-2-27b-pt-res-all	16,384
Gemma-3-27B-it	google/gemma-3-27b-it	gemma-scope-2-27b-it-res-all	16,384
MedGemma-4B-it	google/medgemma-4b-it	gemma-scope-2-4b-it-res-all	16,384
MedGemma-27B-text-it	google/medgemma-27b-text-it	gemma-scope-2-27b-pt-res-all	16,384

Table 4: Model and dictionary identifiers. MedGemma dictionaries are borrowed from the corresponding Gemma-3 parents.

E EXISTENCE: HELD-OUT SEPARABILITY AND SPLIT CONTROLS

This appendix carries the split-control detail behind §4.1. Table 5 compares in-sample against held-out one-vs-rest AUC at the selected peak layer. Optimism (in-sample minus held-out) stays below 0.009 in every case and is negative on four of five Qwen models, meaning the held-out AUC often exceeds the in-sample value, a conservative outcome. The patient-level split, which forces all notes of a patient into the same half, guards against within-patient autocorrelation and leaves held-out AUC above 0.877, so the directions generalize across patients and not merely across notes of patients seen in selection.

Table 5: In-sample against held-out one-vs-rest AUC at the selected peak layer. Optimism is in-sample minus held-out. The patient-level split groups all notes of a patient into the same half.

Model	Split	In-sample	Held-out	Optimism
Qwen3-8B-Base (L23)	Item	0.8862	0.8814	+0.005
II-Medical-8B (L24)	Item	0.8868	0.8810	+0.006
Qwen3-8B-Base (L23)	Patient (1,997 patients)	0.8854	0.8772	+0.008

The dictionary-independence check of §4.1 uses the fresh II-Medical-8B dictionary of Appendix D. Projected into residual coordinates, its chapter directions cosine at $+0.835$ on average with those of the borrowed parent dictionary, positive on all seventeen chapters with range $+0.784$ to $+0.865$, so two dictionaries trained on different activation distributions agree on where the chapter lives.

F CLINICAL SPECIFICITY: FULL FIRING-RATE RESULTS

Table 6 reports the condition-wise firing rates behind §4.2 for the Qwen family. For TopK dictionaries a feature counts as active when it appears among the top-K features of the token, and for JumpReLU dictionaries when its pre-activation exceeds the learned threshold. Full notes activate all ten selected features by construction, since the features are chosen as the ten highest selection-half AUC features on discharge notes. The Gemma family follows the same qualitative pattern.

G CAUSAL INTERVENTION: SWEEP DETAIL AND CONTROLS

Magnitudes behind Table 1. Across the four Qwen models with sound baselines, held-out ΔP at the selected configuration is $+0.041$ to $+0.069$, a relative lift of $+7\%$ to $+18\%$ over baselines

Table 6: Mean number of top-10 chapter features activating per condition (of 10), averaged over the seventeen chapters, Qwen family. Non-clinical text is fifteen short sentences on geography, sports, and food.

Model	Full notes	Bare Dx	Chapter title	Adjacent	Non-clinical
Qwen3-8B-Base	10.0	2.7	—	0.6	0.013
Qwen3.5-9B-Base	10.0	2.6	2.4	0.3	0.000
Qwen3.5-27B	10.0	2.7	2.9	0.3	0.008
Qwen3-30B-A3B-Base	10.0	2.0	1.4	0.4	0.016
II-Medical-8B	10.0	2.1	1.3	0.4	0.004

of 0.376–0.581 where uniform chance is 5.9%. II-Medical-8B's near-degenerate baseline can be attributed to post-training issues (a performance of 0.111, with the model answering one fixed letter for most notes) limits the interpretation of its row, and Appendix J treats it separately. The Gemma effects reach $+0.0193$. On the specificity side, the matched-magnitude random perturbation is actively destructive on the pretrained Gemma models (mean -0.035 on the 4B and -0.032 on the 27B across seeds) while the chapter direction helps, so the models tolerate and use their own concept geometry at doses that break them along arbitrary directions.

Sweep and controls. Figure 3 shows the full layer-by-strength sweep behind §4.3 for Qwen3-8B-Base. The positive ridge concentrates in a band of layers even though the AUC profile of §4.1 is nearly flat across depth, which is the depth dissociation of the main text. The layers around the marginal AUC maximum produce a behavioral swing of only 0.022 between the weakest and strongest injection, while earlier layers with essentially the same separability swing 0.283 under identical conditions, a fifteen-fold difference on the same model and direction. The unembedding control extracts the model's unembedding row for the correct answer letter, mean-centers and unit-normalizes it, and injects it at the same layer and strength, and its near-zero cosine with the chapter directions ($+0.003$ on Qwen3-8B-Base at layer 23) together with its failure to reproduce the effect rules out a direct lexical pathway. Benjamini-Hochberg correction at $q = 0.05$ is applied across the roughly 1,500 layer-strength combinations per model rather than across the selected configurations only.

The Gemma rows of Table 1 follow the identical held-out protocol as the Qwen rows, a note-level split with the configuration chosen on the selection half over the positive strengths of the full layer-by-strength sweep and the effect evaluated on the held-out half. Selection optimism on the Gemma rows is between -0.003 and $+0.004$. The Gemma, Qwen3.5-27B, and mixture-of-experts chapter-specific values subtract the mean random-direction effect over seeds 42, 7, and 123 at the reported configuration (random effects per model: -0.035 on Gemma-3-4B-pt, $+0.005$ on Gemma-3-4B-it, $+0.009$ on MedGemma-4B-it, -0.032 on Gemma-3-27B-pt, $+0.001$ on Gemma-3-27B-it, -0.006 on MedGemma-27B-text-it, -0.017 on Qwen3.5-27B, and -0.025 on Qwen3-30B-A3B-Base, whose routing makes arbitrary perturbations costly while the chapter direction helps at the same dose). The remaining Qwen chapter-specific values were measured at each model's earlier best configuration, within a few layers of the reported one. For the MedGemma variants the dictionaries are trained on the Gemma-3 parents rather than on the fine-tuned models themselves, so the model's own medical training additionally shapes how the dictionary interacts with the activations, which is what the MedGemma-4B-it specificity value reflects, the instrument rather than the model. Training dictionaries natively on these variants is the straightforward extension and is left for future work.

H REGISTER TRANSFER

Table 7 carries the full transfer results behind §4.1. Directions fitted on discharge summaries are applied without refitting to note excerpts from a patient-question corpus with a different length distribution and voice. Retention normalizes by the in-corpus ceiling, $\text{Retention} = (\text{AUC}_{\text{out}} - 0.5)/(\text{AUC}_{\text{ceiling}} - 0.5)$, because AUC has a floor of 0.5 rather than 0. Every Qwen entry reaches permutation $p < 1/2{,}000$ under a 2,000-shuffle null.

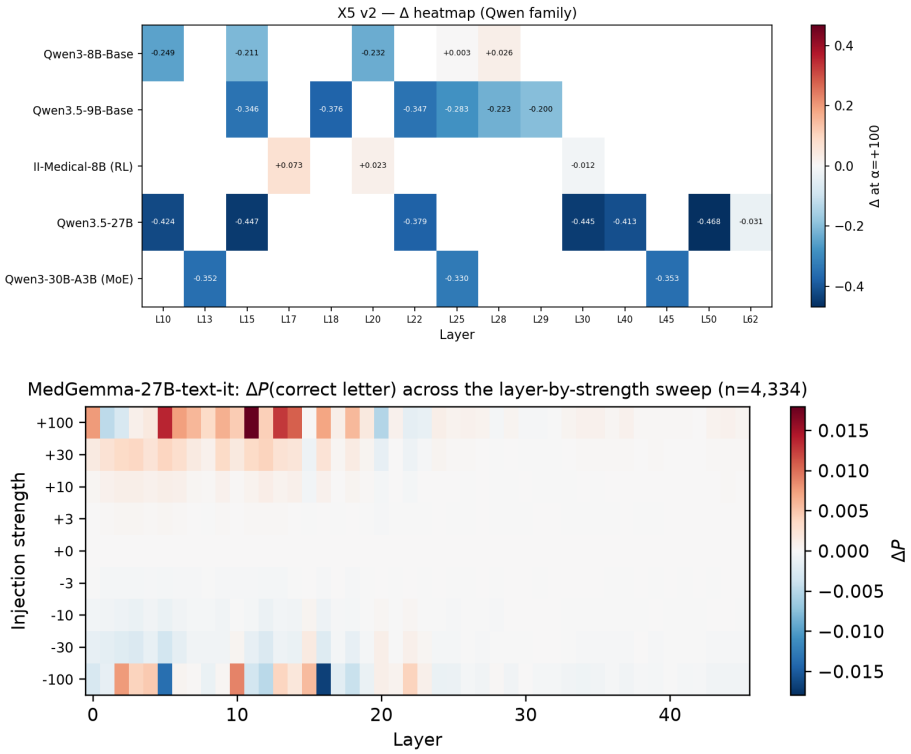

Figure 3: $\Delta P(\text{correct letter})$ over the full layer-by-strength sweep, one model per family: Qwen3-8B-Base (top, held-out evaluation of the selection-half choice) and MedGemma-27B-text-it (bottom, full-manifest sweep). In both families the positive response concentrates in a band of layers even though separability is nearly uniform across depth, and grows with strength up to the swept maximum, the depth dissociation of §4.3.

Table 7: Out-of-register chapter separability, Qwen family. Directions are applied to the new register without refitting.

Model	**Out-of-register AUC**	**In-corpus ceiling**	**Retention**	**Permutation** p
Qwen3.5-27B	0.815	0.868	86%	$< 1/2{,}000$
Qwen3.5-9B-Base	0.756	0.845	74%	$< 1/2{,}000$
Qwen3-30B-A3B-Base	0.709	0.823	65%	$< 1/2{,}000$
II-Medical-8B	0.646	0.691	76%	$< 1/2{,}000$
Qwen3-8B-Base	0.630	0.792	45%	$< 1/2{,}000$

I TEMPORAL PERSISTENCE: MEASUREMENT DETAIL

The 170 passes per model of §4.4 are ten held-out notes per chapter. Base models continue the discharge note itself. Chat models end their generations early under the bare continuation, so they are prompted for a detailed clinical assessment of the note covering the primary diagnosis, the differential, and the management plan, and both registers show the same sustained dominance shape. Generations run to a natural ending capped at 2,000 tokens, and most models write past 1,200 tokens at the median. II-Medical-8B stops at a median of 359 tokens, which is why its curve ends near token 500.

Each model is read at its most responsive layer among those with near-complete finite coverage (at least 95% of the best layer's finite-token count). The coverage condition matters on the two 27B

Gemma dictionaries, whose last layers produce non-finite codes concentrated at early token positions, so an unconstrained effect-size maximum would select a layer that cannot even be evaluated over the opening tokens. The underlying projection grows monotonically with depth (a 130-fold increase from the first to the last layer on Qwen3-8B-Base), consistent with the depth picture of §4.3. II-Medical-8B is the instructive per-model case, near zero over its opening tokens and rising later in the window, consistent with its spikier, content-locked signal, where the projection lights up on chapter-aligned diagnostic tokens.

J ROLE PRIMING: SIX-CONDITION DESIGN AND THE DEGRADED-MODEL RESCUE

Table 8: An assigned specialty moves the model through its concept centers, on both families (deferred from §5.1 for space). Paired t compares matched against counterfactual specialty on true-chapter projections, and accuracy gain is matched specialty over the no-role baseline (paired sign test).

Model	**Paired t (matched vs. counterfactual)**	**Acc. gain (pp)**	**Sign-test p**
Qwen3-8B-Base	+9.18	+5.88	1.8×10^{-4}
Qwen3.5-9B-Base	+8.92	+4.12	2.6×10^{-3}
Qwen3.5-27B	+8.63	+3.82	2.4×10^{-4}
Qwen3-30B-A3B-Base	+9.53	+2.94	0.087
Gemma-3-4B-pt	+3.70	+7.65	8.7×10^{-7}
Gemma-3-4B-it	+2.90	+8.24	5.8×10^{-8}
MedGemma-4B-it	+5.59	+7.35	4.2×10^{-7}
Gemma-3-27B-pt	+3.12	+11.18	4.1×10^{-10}
Gemma-3-27B-it	+5.92	+4.71	1.4×10^{-4}
MedGemma-27B-text-it	+6.80	+5.88	1.9×10^{-6}

The six conditions of §5 are matched specialty, counterfactual specialty, randomly drawn clinical specialty, generalist physician, non-clinical role, and no role, giving 105 role prompts per model over the seventeen chapters plus global controls. The decisive contrasts on Qwen3-8B-Base are random-clinical minus baseline at -0.0815 (paired $t = -0.22$, $p = 0.83$, a randomly assigned clinical specialty does nothing), matched minus random-clinical at $+1.3010$ ($t = +9.57$, $p = 2.3 \times 10^{-19}$, 14 of 17 chapters), and matched minus generalist at $+0.6939$ ($p = 0.054$), placing the unspecified-physician condition between the two, as expected if the effect tracks specialty identity rather than clinical register. The matched-over-random contrast exceeds $t = 7.6$ on four of five models, the accuracy gain of Table 8 holds on 15–16 of 17 chapters per Qwen model, and the counterfactual condition stays within 0.3 accuracy points of the no-role baseline on the Qwen models and within about two points on the Gemma instruct models. Projections in Table 8 are read at each model's best persona-effect layer. Reading depth matters, since the persona-effect peak moves from 83% of depth on Qwen3-8B-Base through 91% (9B) and 94% (30B-A3B) to 97% (27B), while the chapter-encoding peak moves the opposite way with scale (roughly 15% of depth on the 27B model against 85% on the 8B), so any fixed-depth evaluation will miss one of the two signals on most models. The chat-template comparison holds the model fixed and toggles only the template, and the persona-specific mean shift is $3.5\times$ larger without the template on Qwen3.5-27B.

Measurement caveat for role-priming analyses. For those interested in performing this kind of latent-space evaluation, it is important to note that the chat template dilutes the persona signal at the default final-position read. A paired within-model comparison on Qwen3.5-27B, identical prompts with and without the template, shows a $3.5\times$ larger persona-specific mean shift without it, so instruct-tuned deployments carry the same internal effect while showing less of it at the position most evaluations read, and analyses that read only there will underestimate role influence. This is important since the final tokens for the instruct models relate more to formatting and instruction/chat templating instead of the actual clinical content.

Rescuing a degraded model through its concept centers. The medical fine-tune II-Medical-8B is the case where the language substrate and the latent space disagree outright, and it doubles as direct evidence for the depth picture of §4.3. Under the constrained-letter task the model answers one fixed letter for roughly 98% of notes and scores 6.2%, barely above chance, yet its internal chapter separation at the pre-answer position is the strongest we measure anywhere (Welch $t = +40.5$, mean projection $+74$ on the true chapter against -4.9 on wrong ones). The model knows and cannot say, an extreme clinical instance of the internal-knowledge gap Orgad et al. (2025) documented in general domains. From the language substrate this model looks broken, but mechanistically it is a model with intact clinical concepts and a damaged output mapping, and the causal-depth picture says exactly where to intervene, at the layers that govern how the latent computation becomes the response. Injecting the chapter direction there more than doubles the probability of the correct answer, from 0.062 at baseline to 0.135 at the optimum, with the chapter-specific effect $+0.057$ against a five-seed random floor of $+0.016$, roughly thirty-five times the random recovery, converging with the triage-failure localization of Fraile Navarro et al. (2026) from the causal side. The useful operating range is wide, and past the optimum the rescue degrades gradually while never falling below baseline within the swept range. The raw difference-of-means direction also helps ($+0.033$ chapter-specific) but is about 70% weaker, consistent with the axis comparison of Appendix K.

K SAE-DECODER AGAINST RAW DIFFERENCE-OF-MEANS AS THE STEERING AXIS

Table 9 gives the dose curves behind the axis choice of §6, computed on identical models, layers, notes, and doses, so the only moving part is the direction. On Qwen3-8B-Base the raw direction produces no positive dose anywhere on its curve while the SAE-decoder direction peaks at $+0.038$, on II-Medical-8B both rescue the degraded output but the SAE-decoder direction is about 70% more effective, and on the mixture-of-experts Qwen3-30B-A3B the raw direction is destructive at every positive dose tested while the SAE-decoder direction still helps at small doses, with the mixture-of-experts model overall needing far smaller doses than the dense models, consistent with its routing being easier to knock off distribution. Our reading in the main text follows from these curves, the decoder reconstruction concentrates the direction on feature axes that downstream computation consumes, while the raw mean also carries components that correlate with the chapter without taking part in the computation.

Table 9: Change in correct-letter probability against dose for the two steering axes at each model's selected layer. Same notes, layers, and doses within each block.

	Qwen3-8B-Base (L25)		**II-Medical-8B (L17)**		**Qwen3-30B-A3B (L45)**	
α	raw	SAE-dec	raw	SAE-dec	raw	SAE-dec
+3	+0.002	+0.006	+0.001	+0.001	—	—
+10	+0.003	+0.017	—	—	−0.029	+0.009
+30	−0.005	**+0.038**	+0.025	+0.030	−0.153	−0.053
+100	−0.159	+0.003	+0.049	**+0.073**	−0.357	−0.353

L DOSE CALIBRATION FOR GENERATION STEERING

The relative-dose rule of §6 matters because residual-stream norms differ by orders of magnitude across models and grow with depth. As a calibration illustration, on Gemma-3-4B-it, whose residual scale is the largest in our pool (median chapter-contrast norm 666.6 at the steered layer against 18.1 for Qwen3-8B-Base), an absolute dose that steers the 8B Qwen model amounts to $0.02\times$ of the natural between-chapter contrast and changes 13 of 1,624 generated answers against 12 for the random control, an intervention that never fired, while the matched relative dose $\rho = 0.55$ yields $+0.0123$ with 118 discordant answers and the predicted result ordering (Table 10). The rule for any cross-model steering comparison is therefore twofold, express the dose as a fraction of the model's own median chapter-contrast norm at the steered layer, and report the discordance count next to every effect so that an intervention that never fired cannot be read as a null.

Table 10: Steering under the deployment prior (deferred from §6 for space; injection held active through generation, 1,624 held-out notes per model). Δ Frac. is the change in the fraction of generated diagnoses naming the correct chapter under the true-chapter direction, and Discordant counts how many answers changed at all. Doses are relative (ρ of the model's median chapter-contrast norm at the steered layer).

Model	ρ	Δ **Frac.**	p	**Discordant**	**Ordering**
Qwen3-8B-Base	0.55–1.77	+0.0111	0.047	74	true > rand > wrong
Qwen3.5-9B-Base	0.55–1.77	+0.0234	0.001	136	true > rand > wrong
Gemma-3-4B-it	0.55	+0.0123	0.080	118	true > rand > wrong

Two methodological choices thus define the steering appraoch for §6. First, the dose is relative to the model's own representation scale, a fraction ρ of the model's typical between-chapter delta at the steered layer, since residual-stream norms vary by orders of magnitude across models and depths. We report discordance counts (how many outputs changed at all) next to every effect so that an intervention that never fired cannot be mistaken for a null result. Second, the main steering axis is the SAE-decoder direction of Equation 1 and not the raw difference of means. In our ablations, on identical models, layers, and doses, the SAE direction wins every head-to-head comparison we run. Intuitively, this can be explained by the decoder reconstruction expressing the concept along the feature axes the downstream layers actually read, while the raw difference of means also finds components that merely correlate with the chapter without taking part in the latent computation, so part of its dose is spent on directions which the model never uses.

M CLINICIAN STUDY PROTOCOL

Cohort. HealthBench conversations (Arora et al., 2025) are drawn from the consensus, hard, and professional strata, sampled in that priority order, and filtered to a chapter-faithful orthopedic cohort matched to the Musculoskeletal chapter. Chapter faithfulness excludes look-alike neighbors (fractures belong to the Injury chapter and gout to Endocrine, so conversations about them are excluded), red-teaming items are excluded as a different construct, and non-English conversations are removed with a language screen whose short-text false positives are corrected by an English stopword-density rule. The cohort holds about one hundred conversations.

Conditions and generation. Each conversation is answered by Qwen3.5-27B under three conditions, no role, matched-specialty system role, and adversarial other-specialty system role. Generation uses the benchmark's own sampling parameters, all conditions of an item share one random seed so that divergence between responses is attributable to the condition and not to sampling, responses run to a natural ending (items that reach the token cap are regenerated at a doubled cap under the same seed, which reproduces the identical prefix and completes it), and reasoning-trace modes are disabled so raters see final responses only. Every generation records the per-token SAE feature firings at the representation-peak and steering layers and the dense projections onto all seventeen chapter directions at every layer.

Screening and blinding. Cases enter workbooks with response positions randomized per row, condition identities absent, three pairwise judgments per row with an explicit tie option, and a fingerprint that binds each workbook to its sealed answer key. The final workbook carries 41 Qwen3.5-27B cases.

Frozen predictions. Before any rating, every response's per-token trace at the representation-peak layer is reduced to the mean target-chapter dominance of §7, and the frozen rule predicts the higher-dominance response of every pair, with the signed activation gap of every pair stored. Planned analysis on return of the ratings is agreement of blinded preference with the frozen predictions via sign tests on non-tie judgments, stratified by comparison type and model, and inter-rater agreement on the double-rated subset.

Raters, returns, and integrity. The workbooks were completed independently by a senior orthopaedic resident surgeon and a junior medical student, giving 246 head-to-head judgments over the 41 Qwen3.5-27B cases plus a free-text comment field per case. Before analysis, every displayed response in the returned workbooks was re-verified against the stored generation of its keyed (condition, case) by content fingerprint, all matching, so no judgment can be attributed to the wrong condition.

Capability exclusions. Three cases were excluded from every computed statistic because the failures the raters reported trace to a capability the model was never given. The pattern is consistent, since these conversations ask the model about specific literature (one asks about a specific named review, one asks for comparative evidence across three injection-guidance techniques, one asks for the best surgical approach and drew responses citing "highly specific recent studies that ... appear fabricated"), and a model without web access can only fabricate such references. The senior's comment on the evidence-comparison case states the problem plainly, "Fabricated references in all. Although, B got 2 of the references correct," which means ratings on these cases measure verification impossibility rather than response quality. The exclusion list with per-case reasons and curated copies of both workbooks with the exclusions applied are released with the data, and all analysis code reproduces every number in this appendix from those files.

Results in full. After exclusions there are 228 paired judgments over 38 cases, with 34 ties in the pooled set, and ties are informative rather than noise, since raters used them exactly where responses converge. The per-rater agreement of the activation-strength prediction is 57.9% for the junior ($n = 95$) and 53.5% for the senior ($n = 99$), and the gap-stratified values in §7 pool both raters. Inter-rater agreement over the three-way choice including Tie is raw 0.58 with Cohen's kappa 0.43, moderate agreement between two clinicians at very different career stages who never saw each other's sheets.

Agreement by activation gap. Table 11 carries the counts and tests behind the main-text percentages, over all pairs and stratified by comparison type. The stratification shows the prediction is not a relabeling of the priming condition. Among plain-versus-adversarial pairs alone it reaches 64.3% overall and 77.8% on the larger-gap half, while aligned-versus-plain pairs give it almost nothing to separate (median activation gap 0.048 against 0.082 for plain-versus-adversarial), and there it stays at chance, with the small gap known before any rating comes back.

Table 11: Agreement of the activation-strength prediction with blinded clinician preference over non-tie judgments, overall and within the half of pairs whose activation gaps are larger (threshold at each subset's median gap, ties at the threshold included). p is the exact two-sided binomial test against 0.5. On the quarter of pairs with the largest gaps the pooled agreement reaches 76.0% (38 of 50, $p = 3.1 \times 10^{-4}$). Gaps are $|\mathrm{dom}_A - \mathrm{dom}_B|$ from the frozen predictions.

		All pairs			**Larger-gap half**		
Pairs	**Median gap**	Agree	n	p	Agree	n	p
All comparisons	0.058	55.7%	194	0.13	66.0%	97	0.002
Plain vs. adversarial	0.082	64.3%	70	0.022	77.8%	36	0.001
Adversarial vs. aligned	0.057	53.3%	60	0.70	66.7%	30	0.099
Aligned vs. plain	0.048	48.4%	64	0.90	54.5%	33	0.73

Condition-level comparisons (secondary finding). The three priming conditions exist to induce different behaviors on the same case so that the mechanistic signal has variation to predict, and the condition-level preferences are therefore a by-product of the design rather than its aim. Table 12 reports them in full. The plain model beats the adversarially primed one 50 to 20 among non-tie judgments (exact binomial $p = 4.4 \times 10^{-4}$), each rater landing on exactly 25 to 10 independently ($p = 0.017$) with the consensus subset at 20 to 6 ($p = 0.009$), and the plain model also beats the matched-specialty priming, 41 to 23 ($p = 0.033$), the open-ended counterpart of the closed-task gains of §5. The comments say where the two settings diverge, because the persona buys diagnostic focus at the price of invented authority, as documented in the qualitative themes below. The relevant point for the paper is that neither direction of priming tracks quality the way its prompt semantics

promise, while the concept-center activations track it wherever they separate the responses, which is the main-text finding of §7.

Table 12: Blinded clinician preferences on the orthopedic cohort (Qwen3.5-27B), a secondary finding. Wins–losses among non-tie judgments with exact binomial p values. Consensus counts only comparisons where both raters chose the same option.

Comparison	Junior	Senior	Consensus	Aggregate
unsteered > adversarial	25–10 (.017)	25–10 (.017)	20–6 (.009)	50–20 (.0004)
unsteered > aligned	22–10 (.050)	19–13 (.377)	13–5 (.096)	41–23 (.033)
aligned > adversarial	19–9 (.087)	17–15 (.860)	10–5 (.302)	36–24 (.155)

Where the raters disagree, and why. The raters disagree on 48 of the 114 paired comparisons, split between 20 tie-versus-choice mismatches and 28 opposite choices. Reading the comments on the disagreement rows shows the two raters weighting different things rather than reading the responses differently. The senior flags concrete technical faults and missing patient context that decide the comparison outright. On a case asking for a step-by-step technique to elevate a collapsed lateral tibial plateau, the junior chose the aligned response as "the most useful surgical-technique response while retaining appropriate caveats," while the senior chose the plain one because "there are technical flaws in the surgical procedure mentioned." On a case where a man describes his wife's intermittent facial grimacing, the junior preferred the aligned response as the one that "most clearly covers the neurological red flags and escalation criteria," while the senior preferred the plain response because it "is the only one that explicitly checks the wife's awareness." On an under-specified back-pain timeline the senior chose the plain response with the note "As the case conversation is incomplete. Response C asking for more details is the best." The junior, in turn, weights completeness and guideline coverage ("A most closely follows the Jones-criteria/differential/uncertainty requirements"). The disagreements are therefore complementary readings of quality, the junior scoring coverage and the senior scoring correctness and context-seeking, which is also why the consensus subset, where both weightings agree, gives the clearest preference counts in Table 12.

Qualitative themes with the underlying text. Three themes carry most of the signal, and each connects to the mechanistic story. First, persona-induced fabrication of specifics. Asked to draft an inpatient progress note from a trauma-ward recording, the aligned model invents a complete objective section ("**Vital Signs:** Blood Pressure: 128/78 mmHg, Heart Rate: 72 bpm ... Temperature: 98.4°F") where the conversation contains no vitals at all, and the junior chose the plain response because it "avoids inventing exact observations," while the adversarially primed model, answering an emergency physician's question about in-hospital cardiac arrests, confabulates an entire biography ("As a hematologist who has spent decades in the trenches of critical care and resuscitation science..."), marked as "less reliable framing and persona fabrication." The persona instruction is a single sentence, and the model expands it into invented vitals or an invented career, the role pushing content the model cannot ground. Second, the adversarial persona drags the content toward its own specialty, which is the concept intrusion our traces measure, seen by the raters unprompted. On a six-month low-back-pain case with radiation to the leg, the adversarially primed response "overemphasises malignancy/hematologic causes," and on a knee-replacement eligibility question the rater ranked the adversarial response last for its "unnecessary hematology framing." Third, the roles cost calibration. On an adult-onset Still's-like presentation (daily fever spikes with a salmon-pink truncal rash), the plain response "is the most appropriately cautious; B/C move toward treatment/confirmation based on insufficient information," and the same pattern repeats wherever the case under-determines the diagnosis. These comments explain the direction of Table 12, since both roles, aligned and adversarial alike, inject authority the model does not have, the adversarial one additionally injects its own specialty's content, and they also explain why the closed-task gains of §5 do not carry over, because in a seventeen-way classification a persona can only move the answer, while in an open-ended response it also moves the framing, the claimed evidence, and the invented particulars that a clinician reads as untrustworthy.

N EXTENDED DISCUSSION

Across our experiments, one key insight is that the presence of a clinical concept inside a model is spread across its depth. There is a circuit across the layers that encodes the concept center, and it is used at every generation step. The concept is separable almost equally at every depth, yet only some depths are causally potent. The persona effect lives later than the chapter encoding, and it moves later still with scale, from 83% of depth on the 8B model to 97% on the 27B. The final token under a chat template hides what the mid-sequence positions show. The constrained answer responds to a targeted impulse at the injection point, while open-ended generation responds to a sustained one. This fits the staged-inference view of transformer depth (Lad et al., 2024) and the observation that intermediate layers carry the best task representations (Skean et al., 2025), and it explains, without any appeal to model deficiency, why single-layer evaluations in this domain produce false nulls. The shape of our depth results also matches the staged description Liu et al. (2025) give of relevance judgment, where input information is extracted early, processed against the instruction in the middle layers, and converted into the required output by late attention heads. Along these lines, the broad early-to-middle band where the chapter concept stays separable is where the clinical semantics of the note are integrated into the model's reasoning, while the causally effective steering band, sitting at or beyond the middle of the network and reaching the final quarter in the largest models, is where the model decides how those latent computations reach the user. These findings pave the way for a new regime of clinical AI evaluations, with a focus on the latents of the models, which are increasingly becoming more sophisticated and internally coherent, during clinical decision support, and warrant a mechanistic mapping of their latent reasoning in addition to the language substrate.