CHARACTERIZING ERRORS IN SMALL MEDICAL LANGUAGE MODELS: GEOMETRY, PREDICTION, AND CAUSAL SENSITIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

What do differences between correct and incorrect medical answers reveal about a language model's errors? Geometric separation, predictive internal signals, and intervention effects offer different evidence, yet none alone establishes that a model recognizes or can correct its mistakes. We introduce MIRAGE, a framework that tests these claims separately on shared questions and responses, using coverage checks, predictive baselines, and controlled interventions. Across five medical checkpoints and four datasets, we collect 208,896 responses across calibration and evaluation. Incorrect responses are more dispersed in 545 of 808 eligible evaluation pairs, but this geometric comparison covers only 12.6% of the 6,400 evaluated pairs. On 278 matched pairs, masking explicit answers reduces biomedical Fisher label propagation macro-F1 from 0.807 to 0.670, revealing a substantial contribution from answer content. Internal states predict errors on unseen questions, while gains beyond answer type vary across visual tasks and fitting partitions. Interventions produce protocol dependent effects, and a supplementary restoration study finds no observed benefit in sampled answer recovery. Together, these results identify three obstacles to interpreting correctness signals: selective coverage, dependence on answer content and task structure, and a gap between output sensitivity and demonstrated repair. By turning these obstacles into explicit testbeds, we provide a reproducible framework for testing what geometric, predictive, and intervention evidence can establish about model errors, while keeping benchmark agreement separate from clinical explanation.

1 INTRODUCTION

Medical language models can answer questions using both text and images, and open checkpoints make their responses and internal states available for study (Labrak et al., 2024; Sellergren et al., 2026b; Team et al., 2025). Yet benchmark accuracy alone provides a limited account of their errors: models can produce incorrect answers with plausible explanations, and agreement with an answer key does not establish sound clinical reasoning (Singhal et al., 2023; Pal et al., 2023; Pandit et al., 2025; Chen et al., 2025). Understanding these failures requires asking what response patterns and internal measurements actually reveal about correctness.

Several approaches offer useful but different evidence. Repeated generation and response embeddings reveal consistency, uncertainty, and geometric differences between correct and incorrect answers (Manakul et al., 2023; Farquhar et al., 2024; Ricco et al., 2026; Ceballos-Arroyo et al., 2024). Internal probes test whether model representations contain information associated with truthfulness or errors (Burns et al., 2022; Chen et al., 2024). Interventions test whether changing an input or internal state changes an output (Zhang & Nanda, 2023; Liu et al., 2025). These observations are sometimes interpreted as evidence that a model *knows when it is wrong*. However, geometric separation may reflect explicit answers or formatting, decodable information need not guide generation, and an intervention may change an output without explaining or correcting an error (Hewitt & Liang, 2019; Belinkov, 2022; Zhang & Nanda, 2023). Medical sparse-autoencoder studies further connect reasoning-error categories to internal interventions, and physician–model comparisons offer an independent view of evidence relevance (Liu et al., 2025; Hao et al., 2025). The central challenge is to distinguish what each measurement supports.

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

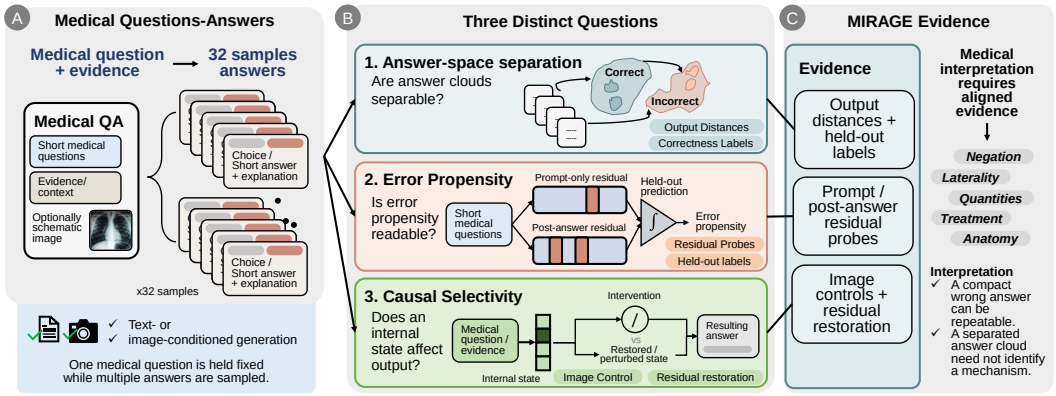


Figure 1: **The MIRAGE audit.** (A) Multiple answers are sampled for a fixed medical question and evidence. (B) Three evidence levels are tested separately: descriptive answer-space separation, predictive error propensity, and causal sensitivity to a specified intervention. (C) Each level is checked against medical distinctions such as negation, laterality, quantities, treatment, and anatomy.

To support this, we introduce MIRAGE (**M**edical **I**nference **R**epresentation **A**nalysis of **G**eometry and **E**rrors), a framework that audits these claims on linked questions, responses, and internal states (Figure 1). *Descriptive geometry* asks how correct and incorrect responses to the same question differ in their embedding space. *Predictive accessibility* asks whether internal states predict errors on questions excluded from fitting. *Causal sensitivity* asks whether a specified change to evidence or an internal state changes a specified output. These claims require distinct controls: sufficient responses from both correctness classes, separate fitting and evaluation questions, and matched intervention controls with execution checks. Evidence for one claim does not establish the others.
We evaluate five medical checkpoints across MedMCQA, MedQA, VQA-RAD, and English SLAKE. The primary collection contains 208,896 responses, including 204,800 responses from 6,400 evaluation question and model pairs. Throughout, correctness means agreement with the benchmark reference. Geometry analyses compare full responses with views that remove explicit answers, while tracking eligibility and encoder consistency. Internal probes distinguish two targets: states before generation predict whether a question produces mostly incorrect responses, whereas states after a generated prefix predict individual response correctness. Visual propensity analyses additionally compare residual features with the dataset's answer type labels. Separate control subsets test image changes and residual restoration against sham and random interventions.

Our contributions are:

1. **A coordinated audit with distinct claims.** MIRAGE links response geometry, prediction on unseen questions, and controlled interventions through shared identifiers, while retaining separate targets, controls, and units of analysis.
2. **Evidence that geometric patterns depend on coverage and answer content.** Only 808 of 6,400 evaluation pairs meet the requirement of at least five parsed responses in each correctness class. Incorrect responses are more dispersed in 545 eligible pairs, but the direction is not universal. On 278 matched pairs, masking explicit answers reduces biomedical Fisher label propagation macro F1 from 0.807 to 0.670, showing that answer content contributes to separation within questions.
3. **Limits on predictive and causal interpretations.** Residual probes recover information about errors, but visual propensity gains beyond answer type vary across checkpoints, datasets, and fitting partitions. Image changes and residual restoration affect answers or reference answer scores, yet the effects depend on the intervention protocol and execution path. The sampled intervention follow up does not establish reliable answer recovery.

Together, these findings identify concrete alternatives to interpreting a measured signal as a general representation of medical error. Answer content can support separation, task structure can support prediction, and broad perturbations can produce intervention effects. MIRAGE makes these alternatives testable through coverage reporting, matched baselines, disjoint evaluation, and stability checks. Its linked records and reproducible analyses support model auditing and mechanistic hypothesis generation, with clinical interpretation requiring evidence beyond benchmark agreement.

2 Related Work

Medical Models and Correctness. Medical language models now span biomedical adaptation, instruction tuning, and multimodal reasoning (Labrak et al., 2024; Ankit Pal, 2024; Sellergren et al., 2026b; Team et al., 2025). Common benchmarks such as MedQA and MedMCQA measure agreement with examination answer keys, while VQA-RAD and SLAKE add image-grounded and short-answer settings (Jin et al., 2021; Pal et al., 2022; Lau et al., 2018; Liu et al., 2021). Broader evaluations show that answer accuracy does not capture hallucination, explanation quality, or clinical safety (Singhal et al., 2023; Pal et al., 2023; Pandit et al., 2025; Chen et al., 2025). These targets are related but require different labels and evaluation protocols. MIRAGE therefore treats benchmark correctness as a controlled target rather than a proxy for clinical validity, and it is important when interpreting geometric or internal signals as evidence about model errors.

Response Geometry and Internal Signals. Repeated generation provides a simple way to study response consistency and uncertainty: SelfCheckGPT uses agreement across samples, while semantic uncertainty methods distinguish surface variation from changes in meaning (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). APORIA studies the geometry of response embeddings through within- and cross-class distances and supervised separation (Ricco et al., 2026); sentence and biomedical encoders provide different representation spaces (Reimers & Gurevych, 2019; Gu et al., 2021). Yet such geometry can also reflect answer format, repeated text, explanation length, or encoder choice. Internal probes offer another view, but information that can be decoded from a representation need not be information used by the model (Alain & Bengio, 2016; Hewitt & Liang, 2019; Belinkov, 2022). Causal methods such as activation patching and inference-time intervention provide stronger tests of internal influence, while also showing that intervention choices can affect conclusions (Li et al., 2023; Zhang & Nanda, 2023). MIRAGE brings these views together on a common medical cohort, with controls that separate response-space separation, held-out error prediction, and causal sensitivity. A more detailed discussion of related work is provided in §A.

3 Preliminaries

✓ **Questions, responses, and correctness.** Let q denote a medical question, v its image when available, and m a fixed generator checkpoint. The evidence is $e = (q, v)$ for visual questions and $e = q$ otherwise. Under fixed stochastic decoding settings, we sample $r_i \sim p_m(\cdot \mid e)$ for $i = 1, \ldots, n$. A deterministic parser extracts an answer a_i. For parsed responses, $y_i = 0$ denotes agreement with the benchmark reference and $y_i = 1$ denotes disagreement. Unparsed responses are excluded from geometric labels but count as nonmatches in overall reference match rates. Thus, correctness refers to benchmark agreement; normalization of open visual answers does not establish clinical equivalence. For an encoder f and text view τ, we define $x_i = f(\tau(r_i))/\|f(\tau(r_i))\|_2$. We distinguish full responses, explicit answer masking, and separately extracted rationales. Let $I_G = \{i : y_i = 0\}$ and $I_H = \{i : y_i = 1\}$, and define the indexed collections $G_q = (x_i)_{i \in I_G}$ and $H_q = (x_i)_{i \in I_H}$. These collections retain repeated responses and identical embeddings. Geometry requires $|I_G|, |I_H| \geq 5$; pairs without this support remain in coverage reporting.

✓ **Distance distributions and direction.** For an indexed collection $A = (x_i)_{i \in I}$, let $D(A) = \{\{\|x_i - x_j\|_2 : i, j \in I,\ i < j\}\}$ be the multiset of distances between distinct response indices. Let P_G and P_H be the empirical distributions of $D(G_q)$ and $D(H_q)$, and let P_{GH} be the empirical distribution of all distances between the two classes. Following APORIA (Ricco et al., 2026), we compare the within class distance distributions using $W_q = W_1(P_G, P_H) = \int_{\mathbb{R}} |F_G(t) - F_H(t)|\, dt$, where F_G and F_H are their empirical CDFs. This measures transport between distance distributions. We also report the signed gap $\Delta_q = \mathbb{E}_{d \sim P_H}[d] - \mathbb{E}_{d \sim P_G}[d]$ and the normalized contrast $\mathrm{HII}_q = W_q/(\mathrm{med}(P_G) + \mathrm{med}(P_H) + 10^{-9})$. Positive Δ_q indicates greater dispersion among incorrect responses. W_q captures differences in dispersion or distribution shape without specifying their direction. HII can be unstable when both medians approach zero, so we interpret it alongside W_q, Δ_q, and response duplication. Distances sharing a response are dependent; inference therefore uses response labels and question or image groups as the relevant units.

✓ **Observation: separation need not imply unequal dispersion.** Suppose all correct responses have embedding u and all incorrect responses have embedding $v \neq u$, with at least two responses per class. Both within class distance distributions equal δ_0, giving $W_q = \Delta_q = 0$, yet a hyperplane orthogonal to $v - u$ through $(u + v)/2$ perfectly separates the classes. Thus, separability and dispersion measure different properties, and neither alone establishes error prediction on unseen questions.

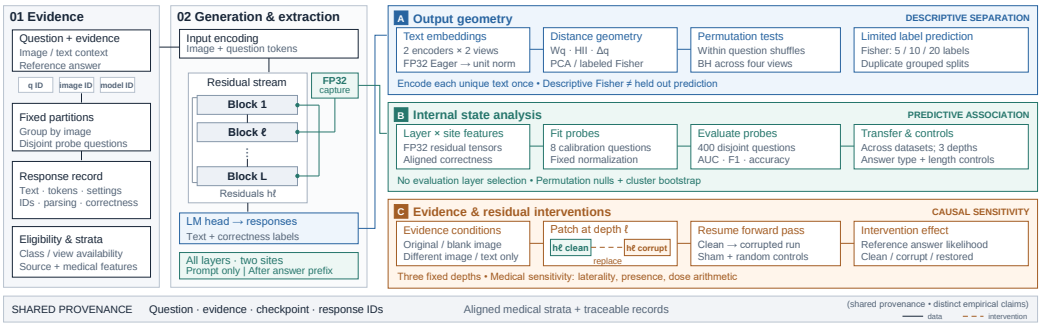


Figure 2: **MIRAGE architecture.** Shared records connect three analyses with distinct targets and controls: response geometry, prediction from internal states, and sensitivity to controlled interventions.

✓ **Internal states and prediction targets.** Let $h^{\mathrm{pre}}_{q,\ell}$ denote the residual state after decoder block ℓ at the final prompt position, and let $h^{\mathrm{post}}_{qi,\ell}$ denote the state at the final position of a short replayed generated prefix. The state before generation is shared across sampled responses. For the parsed response indices I_q, define $\widehat{\rho}_q = |I_q|^{-1} \sum_{i \in I_q} y_i$ and $z_q = \mathbf{1}[\widehat{\rho}_q > 1/2]$, provided $|I_q| > 0$. The first probe predicts z_q, indicating whether a majority of parsed responses are incorrect; ties receive label zero. The second predicts the individual response label y_i. These are distinct targets derived from sampled responses. We evaluate both on questions excluded from fitting, keep shared images within partitions, and apply answer type and response length controls where relevant.

✓ **Intervention quantities.** For a reference answer a^* containing T tokens, define its mean log probability under teacher forcing as $s(e) = T^{-1} \sum_{t=1}^{T} \log p_m(a^*_t \mid e, a^*_{<t})$. Let c denote a specified corruption of the evidence or internal computation, and let $s_c(e)$ be the resulting score. Restoring the clean residual at layer ℓ and token position j gives the effect $R_{\ell,j} = s_c(e; h_{\ell,j} \leftarrow h^{\mathrm{clean}}_{\ell,j}) - s_c(e)$. A sham reinserts the corrupted residual. A matched random control adds a random vector to that residual with norm $\|h^{\mathrm{clean}}_{\ell,j} - h^{\mathrm{corrupt}}_{\ell,j}\|_2$. These comparisons measure sensitivity of a specified reference answer score. They do not establish improved sampled accuracy, and the original reference need not remain correct after an evidence change.

4 Methodology

MIRAGE links questions, evidence groups, checkpoints, and responses through fixed identifiers; and shared records connect text, token IDs, decoding settings, parsing outcomes, and benchmark agreement across geometry, prediction, and intervention analyses. Figure 2 summarizes the pipeline.

4.1 Output geometry and permutation inference

For each eligible question and representation, we compute W_q, HII_q, and Δ_q. PCA provides an unsupervised view; the APORIA Fisher projection uses all labels in the displayed response set and serves only as a descriptive visualization. We assess W_q by shuffling correctness labels within each question, preserving class counts and recomputing both distance distributions. With B permutations, $p_q = (1 + \sum_{b=1}^{B} \mathbf{1}[W_q^{(b)} \geq W_q])/(B+1)$. Shuffling response labels retains dependence among distances sharing a response. The null is label exchangeability within a question, rather than equality of dispersion alone. Benjamini–Hochberg correction pools eligible tests across four representations and both calibration and evaluation rows within each model and dataset cell; reported findings use evaluation rows. We separately assess a global testing family and class support thresholds from 2-6. All primary geometry and revised Fisher analyses use corrected embeddings. Each exact text view is encoded once using FP32 Eager attention, padding to 512 tokens, masked mean pooling, and unit normalization, then reused for matching responses. This corrects a diagnostic failure in which identical texts received different embeddings across batches under SDPA. Original embeddings are retained for comparison.

4.2 Fisher projection and limited-label classification

We use the APORIA Fisher implementation (Ricco et al., 2026), with biased empirical within class covariance and trace scaled ridge regularization ($\lambda = 10^{-3}$). Unlike the descriptive plots, classification trials fit projections and class references using only labeled training responses. We compare Wasserstein and centroid propagation with nearest five class distance and majority baselines.

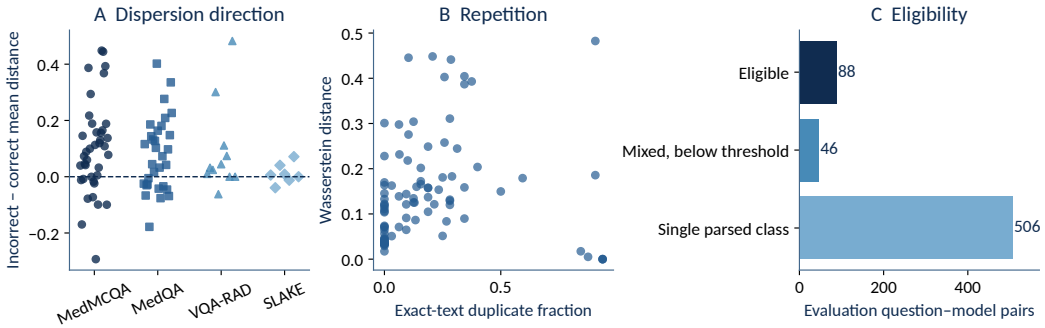


Figure 3: **Raw biomedical geometry across all 640 evaluation question–model pairs.** Points in A–B represent the 88 eligible pairs. Positive signed gap means larger incorrect-answer spread; the unsigned Wasserstein statistic also detects reversed or shape-only differences.

At a budget of 20 biomedical labels, we also compare identity, whitened one dimensional PCA, and random one dimensional projections. Alternative ridge values are assessed only on calibration data. Each eligible question contributes up to three fixed splits and three label subset draws per split at budgets of 5, 10, and 20 responses. Sampling guarantees two labeled responses per class before filling the remaining budget. These budgets therefore assume class coverage and do not measure the full cost of acquiring labels.

We compare stratified response splits with splits that keep whitespace normalized duplicate texts together. Duplicate groups follow the evaluated text view, preventing identical masked or rationale inputs from crossing folds. We retain up to three valid splits from 100 fixed candidates. Trials are averaged within each question and model pair before averaging pairs. The primary matched comparison uses the same 278 pairs across the four encoder and text view combinations.

4.3 Controlled analyses and robustness checks

✓ **Medical descriptors.** Source fields and explicit lexical rules describe specialty, anatomy, modality, task, answer type, negation, laterality, quantities, treatment, and diagnosis. We report coverage before subgroup results. These overlapping descriptors support descriptive analysis; they are not expert judgments of clinical severity. Missing fields remain unknown (§B.2.1).

✓ **Residual probes.** We store residual states before generation and after a short generated prefix in FP32. Probes fit on eight calibration questions and evaluate on 400 disjoint questions, keeping shared images within partitions. States before generation predict the majority error label; states after a prefix predict individual response correctness. Controls include answer type for visual propensity and response length and text features for response correctness. We report AUC, macro F1, accuracy, permutation nulls, and bootstrap intervals over question or image groups, without selecting layers on evaluation data (§B.2.2).

✓ **Transfer and interventions.** We test transfer across datasets at three fixed decoder depths. Paired original, blank, swapped image, and text only inputs measure evidence dependence. Residual restoration compares clean replacement with sham and matched random controls using reference answer log probability. A separate sampled follow up examines answer recovery, while execution replay checks score stability (§B.2.3).

✓ **Additional sensitivity analyses.** Fixed partitions test fitting stability, alternative predictors, representation baselines, and null refits. A small controlled language suite tests laterality, presence or absence, and dose arithmetic using paired counterfactual prompts and interventions at specified positions and depths. These exploratory checks do not constitute a new confirmatory cohort (§B.2.4).

5 Experimental Design

✓ **Checkpoints and tasks.** We evaluate MedGemma-4b-it, MedGemma-1.5-4b-it, Lingshu-7B, BioMistral-7B, and OpenBioLLM-8B on MedMCQA, MedQA-USMLE, VQA-RAD, and English SLAKE. The resulting 16 compatible combinations support text comparisons across five checkpoints and visual comparisons across three multimodal checkpoints. We treat checkpoints as fixed model identities, without assuming independent architectures. Repository revisions and the compatibility matrix are recorded in §C.1.

✓ **Main cohort and generation.** Each dataset contains 408 fixed questions: eight for calibration and

Table 1: **Paired Fisher label-propagation comparison.** The same 278 question–model pairs support all four encoder/text views; trials are averaged within pairs and pairs weighted equally. Full responses retain explicit answers; rationale views remove the parsed answer line.

Selected Encoder / Seletcted Text	Pairs Evaluated	Mean macro-F1 Score	Median macro-F1 Score
PubMed / Full	278	0.807	0.822
PubMed / Rationale	278	0.666	0.630
General / Full	278	0.818	0.840
General / Rationale	278	0.665	0.631

400 for evaluation. Each compatible question and checkpoint pair receives 32 stochastic responses, yielding 208,896 primary responses: 4,096 for calibration and 204,800 across 6,400 evaluation pairs. The visual evaluation sets contain 195 image groups for VQA-RAD and 88 for SLAKE. Shared images and duplicate normalized questions remain within partitions, with cohort identities fixed by hashes. Generation uses fixed prompts, temperature 1.0, top $p = 0.95$, and a limit of 192 new tokens, without selecting or resampling responses based on correctness. Cohort construction and answer formats are detailed in §C.

✓ **Representations and supplementary controls.** We analyze full, answer masked, and rationale text views with biomedical and general encoders, alongside residual states captured before generation and after a short generated prefix. Supplementary studies use smaller subsets, typically 40 evaluation questions per compatible combination, to test evidence dependence, intervention behavior, and execution stability. Image controls compare original, blank, swapped image, and text only inputs; residual controls compare clean restoration with sham and matched random perturbations. A separate MedGemma 4B study uses 40 questions per visual dataset, eight paired seeds, and five conditions, producing 3,200 responses with a 64 token limit. Supplementary settings and counts are reported separately from the main cohort (§C.4).

✓ **Reproducibility and analysis units.** We retain responses, representations, residual states, intervention records, seeds, source revisions, and analysis outputs with execution logs and shared identifiers. Calibration supports fitting and methodological choices; evaluation questions remain excluded from probe fitting and primary layer selection. Repeated responses and intervention measurements are distinguished from question and image groups when reporting sample sizes and uncertainty. Full accounting and implementation details appear in §C.5.

6 Experimental Findings

6.1 Geometry and answer-space separation

✓ **Coverage and dispersion.** Figure 3 summarizes geometry across the evaluation cohort, with detailed counts in Appendix Table 5. Only 808 of 6,400 question and model pairs (12.6%) contain at least five parsed responses in each correctness class. In the corrected biomedical full response embeddings, incorrect responses are more dispersed in 545 eligible pairs (67.5%). The corresponding counts are 220/313 for MedMCQA, 204/285 for MedQA, 70/117 for SLAKE, and 51/93 for VQA-RAD. These fractions describe the supported subset; the remaining 5,592 pairs lack sufficient class support. Positive gaps do not necessarily imply significant permutation results, which test an unsigned statistic. This limited coverage means that the observed dispersion pattern cannot be generalized to questions whose sampled responses are uniformly correct or incorrect.

✓ **Compact incorrect responses.** Figure 4 illustrates a reversed dispersion pattern for Lingshu on MedQA question 322. Incorrect responses are more compact despite a nonzero difference between distance distributions ($\Delta_q = -0.178$, $W_q = 0.179$). The Fisher projection uses all labels in the displayed set and illustrates separation within this question; it does not test prediction on unseen questions. This example shows that tightly clustered responses can still be wrong, limiting the interpretation of geometric consistency as evidence of correctness.

✓ **Dependence on answer content.** Table 1 compares Fisher label propagation on 278 matched pairs. Masking explicit answers reduces macro F1 from 0.807 to 0.670 for the biomedical encoder and from 0.818 to 0.672 for the general encoder. The paired decreases are 0.137 and 0.146, with 95% bootstrap intervals over question or image groups of $[0.117, 0.158]$ and $[0.124, 0.167]$, respectively (Appendix Table 15). Rationale scores are 0.666 and 0.665. These results show that explicit answer content contributes to separation within questions; they do not establish prediction on unseen questions.

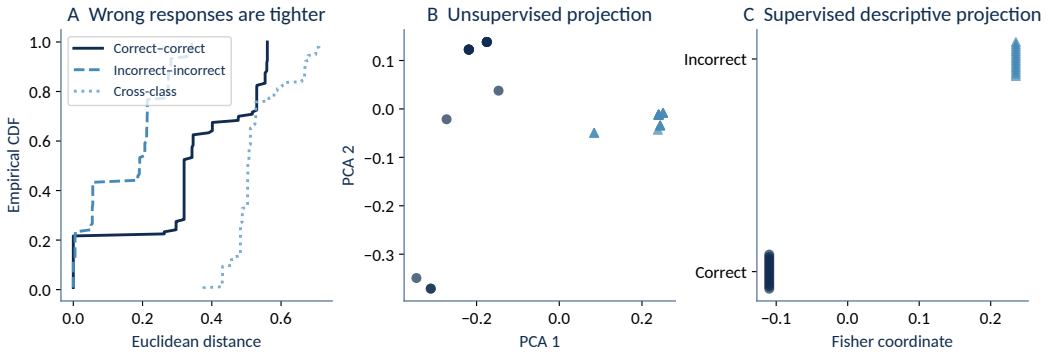


Figure 4: **Geometric analysis-focused close reading of Lingshu / MedQA question 322.** Incorrect responses have smaller mean within-class distance (gap -0.178), despite a detectable distance-law difference (W1 = 0.179). Fisher is fitted to this displayed set and is descriptive; held-out evidence is reported separately.

Table 2: **Prompt-only visual propensity across twenty predetermined image-group-disjoint partitions with eight training groups.** Gain compares type-plus-residual with type only on the same split. Low/high are empirical 2.5/97.5 percentiles across overlapping splits. PS = Positive Splits

Model	Dataset	Splits	Type	Residual	Combined	Gain	Gain_{low}	Gain_{high}	PS
Lingshu-7B	SLAKE	20	0.811	0.842	0.842	0.032	-0.076	0.075	16
Lingshu-7B	VQA-RAD	20	0.811	0.731	0.732	-0.080	-0.247	0.010	3
MedGemma-1.5-4B-it	SLAKE	20	0.753	0.774	0.774	0.021	-0.078	0.077	14
MedGemma-1.5-4B-it	VQA-RAD	20	0.762	0.719	0.719	-0.042	-0.139	0.016	4
MedGemma-4B-it	SLAKE	20	0.692	0.774	0.774	0.082	-0.035	0.136	19
MedGemma-4B-it	VQA-RAD	20	0.770	0.740	0.740	-0.030	-0.225	0.047	8

6.2 Internal states and prediction on unseen questions

✓ **Predictive information in residuals.** Appendix Table 16 reports midpoint probe results for both prediction targets. States before generation predict the majority error label most strongly on visual tasks: AUC ranges from 0.776 to 0.830 on VQA-RAD and from 0.765 to 0.821 on SLAKE. Text task AUC ranges from 0.418 to 0.657. States after a generated prefix predict individual response errors with AUC from 0.398 to 0.759. These settings target different outcomes, and the latter can also encode generated answer content and length.

✓ **Gains beyond answer type.** Table 2 compares answer type with combined answer type and residual predictors across twenty partitions. Adding residuals improves mean AUC on all three SLAKE cells but reduces it on all three VQA-RAD cells. Every empirical interval between the 2.5th and 97.5th percentiles of the gain includes zero. These intervals describe sensitivity across overlapping partitions rather than uncertainty from independent replications. With eight training groups per partition, gains depend on the dataset and fitting sample. Transfer across datasets is also variable (§B.2.3).

✓ **Information from repeated outputs.** Appendix Tables 24 and 25 compare residual probes with output features on the same evaluation partitions. These features summarize parsed answer entropy, variation ratio, parsing failures, and response embedding dispersion without using reference answers. Their predictive value varies across checkpoints and datasets, showing that error information is also accessible from repeated outputs. This comparison uses different information budgets: residual states and answer type are available before generation, whereas output features require 32 sampled responses and predict the majority error label of that response set.

✓ **Medical feature coverage.** Figure 5 summarizes medical descriptors, with definitions and detailed counts in §B.2.1. Lexical rules and source annotations identify negation, laterality, quantities, treatment, anatomy, specialty, and modality. These overlapping descriptors characterize coverage rather than clinical severity, and sparse eligible subsets limit comparisons.

6.3 Evidence dependence and controlled interventions

✓ **Sensitivity to visual input.** Figure 5 reports paired image controls, with answer change rates of 70.5% for blank images, 66.2% for swapped images, and 79.6% for text only inputs. These

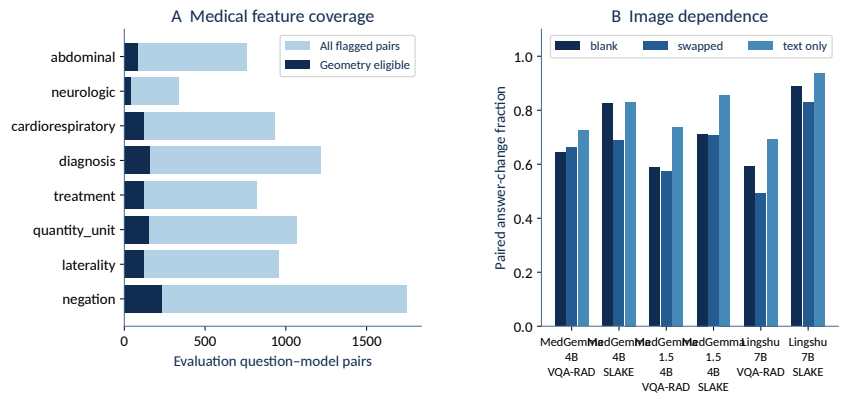


Figure 5: **Medical coverage and evidence sensitivity.** Lexical categories overlap and describe question wording. Image bars use 40 evaluation questions with four paired seeds per cell and condition; answer changes under altered evidence are sensitivity measurements, not altered-image correctness scores.

Table 3: **Sampled restoration in MedGemma 4B.** Changes from the blank image baseline in reference match (ΔMatch; 95% image bootstrap interval) and mean log probability (Δs), using 40 questions per dataset, eight paired seeds, and donors captured before generation.

Selected dataset	Condition applied in this experiment	ΔMatch	95% interval	Mean Log Probability Δs
SLAKE	Restore	−0.003	[−0.009, 0.000]	−0.027
	Random	0.006	[0.000, 0.019]	0.025
	Sham	0.000	[0.000, 0.000]	0.000
VQA-RAD	Restore	0.000	[0.000, 0.000]	0.028
	Random	0.009	[0.000, 0.025]	0.046
	Sham	0.000	[0.000, 0.000]	0.000

comparisons use a separate control subset with paired seeds. An answer change demonstrates input sensitivity but does not establish whether the changed answer is correct or whether the model used clinically relevant image content.

✓ **Residual restoration effects.** Figure 6 compares clean residual restoration with matched random perturbations. Under broad prompt embedding corruption, OpenBioLLM shows the largest mean restoration advantage: 5.529 nats per token on MedMCQA and 5.765 on MedQA. These runs use the recorded raw prompt fallback. Other checkpoints show smaller or mixed effects, including a negative contrast for MedGemma 4B on VQA-RAD. Sham score changes are zero in the reported analysis. These results establish sensitivity of reference answer scores under the tested protocols, without identifying a shared medical mechanism.

✓ **Answer recovery and execution stability.** Table 3 reports the sampled follow up under a fixed MedGemma 4B intervention policy. Restoration changes reference matching by 0.000 on VQA-RAD and -0.003 on SLAKE, with 95% image bootstrap intervals of $[0.000, 0.000]$ and $[-0.009, 0.000]$, respectively. These results show no observed recovery benefit under this policy.

✓ **Answer recovery and execution stability.** The separate sampled follow up and replay checks are detailed in §B.2.3 and §B.2.4. The follow up contains 3,200 responses from 80 questions, eight seeds, and five conditions, but does not establish reliable answer recovery. Cached scoring recovers exact sham behavior, while original and replayed intervention effects differ by up to 0.310 nats in retained failed shards. Successful sham checks therefore do not alone establish reproducibility of intervention effects.

Overall, we can see that geometric differences depend on class support, representation, and answer content. Internal states carry predictive information, but gains beyond answer type vary across visual tasks and fitting partitions. Interventions change outputs and reference answer scores, with effects that depend on the checkpoint and execution protocol. These findings support targeted investigation while leaving a general error representation, reliable answer repair, and a localized medical mechanism unestablished.

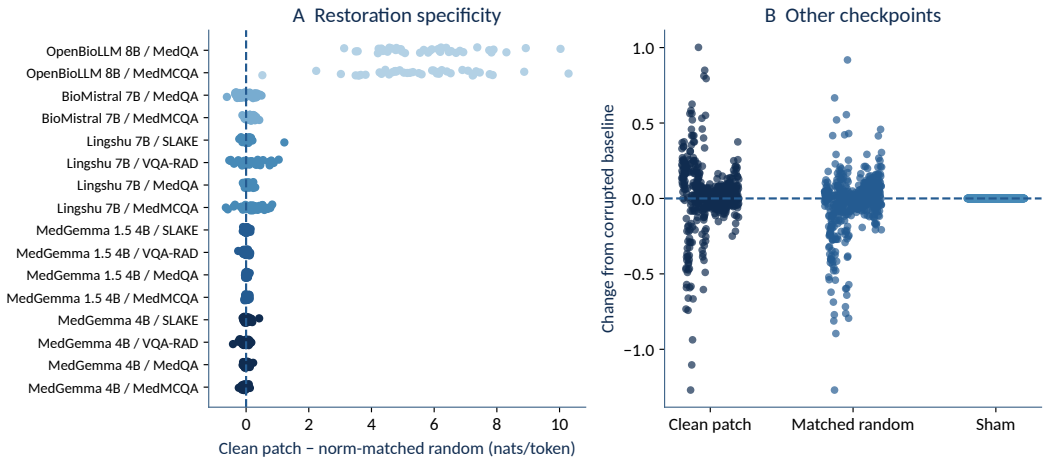


Figure 6: **Controlled residual effects on original-gold mean log probability.** Each point averages three fixed depths for one evaluation question. A contrasts restoration with an equal-norm random perturbation. B resolves other checkpoints. There are 2112 sham measurements; their maximum absolute shift is 0. OpenBioLLM uses the recorded raw-prompt fallback; effects are protocol-specific.

7 Discussion

Our findings show that medical correctness signals depend on what is measured and how the measurement is controlled. Incorrect responses are often more dispersed within eligible questions, yet compact incorrect responses and sensitivity to answer masking show that geometry alone offers an incomplete account of error. Internal states provide complementary predictive information, but gains beyond answer type vary across visual tasks and fitting partitions. This variation limits the interpretation of the tested probes, especially with only eight calibration questions; it does not establish the absence of a more general correctness representation. Together, these results motivate treating geometric structure and predictive accessibility as separate observations whose medical meaning requires further evidence.

Controlled interventions extend this analysis by testing whether changes to evidence or internal states affect model outputs. However, image dependence does not establish use of a clinically relevant feature, and a change in reference answer likelihood does not establish answer repair. The sampled MedGemma 4B study shows no observed recovery benefit under the tested policy, while differences between original and replayed effects highlight the need to verify execution stability alongside sham behavior. These findings make the intervention protocol part of the scientific claim: conclusions must specify the corruption, donor state, intervention position, and measured outcome.

Stronger explanations will therefore require targeted changes to medical evidence, expert assessment, and broader support across independent questions. Contrasts involving laterality, negation, quantities, or presence and absence can test whether a signal tracks a relevant medical distinction. More responses per question improve estimates of response patterns, whereas more independent questions strengthen tests of predictive transfer and fitting stability. By linking these analyses while preserving their distinct claims, MIRAGE provides a basis for developing and testing mechanistic hypotheses without treating benchmark agreement as clinical explanation. Further discussion appears in §K.

8 Conclusion

In this work, we introduce MIRAGE, a framework that separates response geometry, prediction from internal states, and causal sensitivity in medical language model errors. Across five checkpoints and four datasets, geometric patterns depended on class support and answer content, while residual probes provided predictive information without consistent gains beyond answer type on visual tasks. Controlled interventions changed outputs and reference answer scores, but the sampled restoration study showed no recovery benefit under the tested policy. Together, these findings show why separation, prediction, and intervention effects require distinct evidence before supporting a medical explanation. By linking coverage checks, matched baselines, disjoint evaluation, and intervention controls, MIRAGE provides a reproducible basis for auditing correctness signals and developing mechanistic hypotheses while keeping benchmark agreement distinct from clinical validity.

AI USE STATEMENT

AI assistants were used for implementation review, debugging, execution support, statistical checks, and figure ideation. After preparing the first draft, we also used Stanford Agentic Reviewer and CSpaper to obtain preliminary feedback on the methodology, experiments, and presentation and to identify related work for further consideration. All AI-assisted outputs were reviewed and verified by the authors. The authors remain fully responsible for the final content of this work, including all claims, analyses, code, and artifacts.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding Intermediate Layers Using Linear Classifier Probes, 2016. URL https://arxiv.org/abs/1610.01644.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2022. URL https://arxiv.org/abs/2208.02814.

Malaikannan Sankarasubbu Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying, 2023. URL https://arxiv.org/abs/2304.13734.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. ISSN 1530-9312. doi: 10.1162/coli_a_00422. URL http://dx.doi.org/10.1162/coli_a_00422.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Rui Bian and Weihao Cheng. Medscope: A lightweight benchmark of open-source large language models for medical question answering. 2026. doi: 10.64898/2026.03.31.26349827. URL http://dx.doi.org/10.64898/2026.03.31.26349827.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision, 2022. URL https://arxiv.org/abs/2212.03827.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. Open (clinical) llms are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 50–71. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.bionlp-1.5. URL http://dx.doi.org/10.18653/v1/2024.bionlp-1.5.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection, 2024. URL https://arxiv.org/abs/2402.03744.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3563–3599. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.182. URL http://dx.doi.org/10.18653/v1/2025.naacl-long.182.

Xinrong Chen, Xu Chu, Yingmin Qiu, Hengyuan Zhang, Jing Xiong, Shiyu Tang, Shuai Liu, Shaokang Yang, Cheng Yang, Hayden Kwok-Hay So, and Ngai Wong. Residual decoding: Mitigating hallucinations in large vision-language models via history-aware residual guidance, 2026. URL https://arxiv.org/abs/2602.01047v3.

Seonglae Cho, Zekun Wu, Kleyton Da Costa, and Adriano Koshiyama. The confidence manifold: Geometric structure of correctness representations in language models, 2026. URL https://arxiv.org/abs/2602.08159v2.

Rachel L. Draelos, Samina Afreen, Barbara Blasko, Tiffany L. Brazile, Natasha Chase, Dimple Patel Desai, Jessica Evert, Heather L. Gardner, Lauren Herrmann, Aswathy Vaikom House, Stephanie Kass, Marianne Kavan, Kirshma Khemani, Amanda Koire, Lauren M. McDonald, Zahraa Rabeeah, and Amy Shah. Large language models provide unsafe answers to patient-posed medical questions. *npj Digital Medicine*, 9(1), 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-02428-5. URL http://dx.doi.org/10.1038/s41746-026-02428-5.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Springer US, 1993. ISBN 9781489945419. doi: 10.1007/978-1-4899-4541-9. URL http://dx.doi.org/10.1007/978-1-4899-4541-9.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024.

R. A. FISHER. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. ISSN 2637-8051. doi: 10.1145/3458754. URL http://dx.doi.org/10.1145/3458754.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL https://openreview.net/forum?id=Zd0XLr6JKn.

Yuexing Hao, Kumail Alhamoud, Hyewon Jeong, Haoran Zhang, Isha Puri, Philip Torr, Mike Schaekermann, Ariel D. Stern, and Marzyeh Ghassemi. MedPAIR: Measuring Physicians and AI Relevance Alignment in Medical Question Answering, 2025. URL https://arxiv.org/abs/2505.24040.

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. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1275. URL http://dx.doi.org/10.18653/v1/D19-1275.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

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. *Appl. Sci. (Basel)*, 11(14):6421, July 2021.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5848–5864, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.348. URL https://aclanthology.org/2024.findings-acl.348/.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data*, 5(1):180251, November 2018.

Kenneth Li, Oam Patel, Fernanda Viegas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, 2023. URL https://arxiv.org/abs/2306.03341.

Zehui Liao, Shishuai Hu, Ke Zou, Huazhu Fu, Liangli Zhen, and Yong Xia. Vision-amplified semantic entropy for hallucination detection in medical visual question answering, 2025. URL https://arxiv.org/abs/2503.20504v1. Version 1; subsequent versions use a different title.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229. URL http://dx.doi.org/10.18653/v1/2022.acl-long.229.

Anglin Liu, Ruichao Chen, Yi Lu, Hongxia Xu, and Jintai Chen. Med-Scout: Curing MLLMs' Geometric Blindness in Medical Perception via Geometry-Aware RL Post-Training, 2026. URL https://arxiv.org/abs/2601.23220.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Jialin Liu, Siru Liu, and Adam Wright. Benchmark integrity and reasoning-trace errors in medical question answering with large language models: Mixed methods study with sparse autoencoders (preprint). *Journal of Medical Internet Research*, December 2025. ISSN 1438-8871. doi: 10.2196/90061. URL http://dx.doi.org/10.2196/90061.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, 2022. URL https://arxiv.org/abs/2202.05262.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In Jing Jiang, David Reitter, and Shumin Deng (eds.), *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 314–334, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.21. URL https://aclanthology.org/2023.conll-1.21/.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2858–2873, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.143. URL https://aclanthology.org/2025.emnlp-main.143/.

Gabriel Peyre and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. URL https://arxiv.org/abs/1803.00567.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010. ISSN 1544-6115. doi: 10.2202/1544-6115.1585. URL http://dx.doi.org/10.2202/1544-6115.1585.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of EMNLP-IJCNLP*, 2019. URL https://arxiv.org/abs/1908.10084.

Emanuele Ricco, Elia Onofri, Lorenzo Cima, Stefano Cresci, and Roberto Di Pietro. A geometric analysis of small-sized language model hallucinations, 2026. URL https://openreview.net/forum?id=5RQPAHUtCO.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1), 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01283-6. URL http://dx.doi.org/10.1038/s41746-024-01283-6.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, Harshad Purandare, Abhishek Bijay Mishra, Sam Schmidgall, Tao Tu, Anil Palepu, Chunjong Park, Tim Strother, Rahul Thapa, Yong Cheng, Preeti Singh, Kat Black, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Joelle Barral, Tris Warkentin, Shravya Shetty, Dale Webster, Sunny Virmani, David F. Steiner, Can Kirmizibayrak, and Daniel Golden. Medgemma 1.5 technical report, 2026a. URL https://arxiv.org/abs/2604.05081.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2026b. URL https://arxiv.org/abs/2507.05201.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL http://dx.doi.org/10.1038/s41586-023-06291-2.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning, 2025. URL https://arxiv.org/abs/2506.07044.

Can Wang and Yiqun Chen. Evaluating large language models for evidence-based clinical question answering. *Patterns*, 7(5):101519, 2026. ISSN 2666-3899. doi: 10.1016/j.patter.2026.101519. URL http://dx.doi.org/10.1016/j.patter.2026.101519.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Rui Xu, Yi Chen, Sihong Xie, and Hui Xiong. Geometry-calibrated conformal abstention for language models, 2026. URL https://arxiv.org/abs/2604.27914.

Zaifu Zhan, Mengyuan Cui, and Rui Zhang. Can Large Language Models Self-Correct in Medical Question Answering? An Exploratory Study, 2026. URL https://arxiv.org/abs/2604.00261.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods, 2023. URL https://arxiv.org/abs/2309.16042.

Appendix

Table of Content

A Related Work **17**

A.1 Medical language models and the meaning of benchmark correctness 17

A.2 Repeated-response uncertainty and geometric analysis 17

A.3 Internal representations, probes, and causal interventions 18

A.4 Statistical foundations and our position 18

A.5 Explanation quality, clinical error structure, and evidence relevance 19

B More Details on Methodology **19**

B.1 Preliminary definitions and edge cases 19

B.2 Controlled analyses and robustness checks 20

B.2.1 Medical views of the output space 20

B.2.2 Residual capture and question-disjoint probes 20

B.2.3 Transfer, evidence controls, and residual restoration 21

B.2.4 Review-stage stability and selective evidence controls 21

C More Details on Experimental Design **21**

C.1 Checkpoints and compatible task matrix 21

C.2 Dataset selection and frozen cohort construction 22

C.3 Generation, prompting, and hardware 22

C.4 Parsing, representations, and implementation settings 23

C.5 Accounting, statistical units, and reproducibility 23

C.6 Revision computation and provenance 24

D Study design and reproducibility **24**

D.1 Prompt requests and generation policy 24

D.2 Core parameters 24

D.3 Supplementary control cohorts 25

E Evaluation coverage and medical descriptors **26**

E.1 Question support and response behavior 26

E.2 Supplementary cohort size sensitivity 27

F Response geometry and answer content **29**

F.1 Projection and label propagation definitions 29

F.2 Text views and paired comparisons 29

F.3 Response counts and projection controls 30

F.4 Encoder and distance sensitivity 33

G Internal state prediction and output baselines **35**
G.1 Answer type baseline 35
G.2 Layer profiles and transfer 36
G.3 Fitting stability and surface controls 40
G.4 Uncertainty from repeated outputs 43
G.5 Calibration support and abstention 45

H Supplementary evidence and residual interventions **46**
H.1 Intervention configurations 46
H.2 Paired image controls 47
H.3 Broad residual restoration 47
H.4 Medical contrasts and intervention sites 48

I Sampled restoration and downstream states **50**
I.1 Generation policy and latent support diagnostics 50

J Numerical integrity and execution audits **51**
J.1 Encoder correction and eligibility 52
J.2 Scope of the encoder discrepancy 54
J.3 Historical geometry audit and examples 54
J.4 Intervention replay checks 55

K Discussion **55**
K.1 Geometric patterns depend on sampling and representation 55
K.2 Answer content and task structure shape correctness signals 56
K.3 Predictive accessibility depends on fitting support 56
K.4 Intervention effects require precise behavioral interpretation 56

L Open Questions and Future Research Directions **56**

A Related Work

A.1 Medical language models and the meaning of benchmark correctness

Medical model development has progressed along complementary paths: adapting General language models to biomedical text, instruction tuning for medical interaction, and integrating image understanding. BioMistral illustrates domain adaptation in an accessible text-model family (Labrak et al., 2024); OpenBioLLM provides a medical instruction-oriented checkpoint and a documented interaction format (Ankit Pal, 2024). MedGemma and its 1.5 release extend an open medical model collection across text and image tasks (Sellergren et al., 2026b;a). Lingshu targets unified multimodal medical understanding and reasoning (Team et al., 2025). These systems motivate our choice of small checkpoints because direct residual access and repeated interventions are feasible without changing model weights. Our comparisons concern their behavior under a recorded sampling protocol, rather than reproducing the heterogeneous evaluations in their release reports.

Medical benchmarks also encode different notions of success. MedQA draws on medical examinations, and MedMCQA provides a broad multiple-choice setting (Jin et al., 2021; Pal et al., 2022). In both, an answer key can establish choice agreement while leaving the explanation unassessed. VQA-RAD contains clinically generated questions about radiology images, whereas SLAKE combines medical visual questions with semantic and knowledge-oriented annotations (Lau et al., 2018; Liu et al., 2021). Their short-answer formats introduce a distinction between exact answer matching and clinical equivalence. We use the released reference answers consistently, but do not turn every mismatch into an expert-adjudicated hallucination label. This decision determines the scope of every geometric and activation result in the paper.

The broader medical evaluation literature argues for assessing more than examination accuracy. The MultiMedQA work connects medical question answering with human evaluation of factuality, possible harm, and other properties (Singhal et al., 2023). Med-HALT constructs medical hallucination-oriented tests, and MedHallu studies detection of medically incorrect generated content (Pal et al., 2023; Pandit et al., 2025). These benchmarks address questions adjacent to ours, but their labels and task constructions are not interchangeable with choice correctness in a repeated-response experiment. We position MIRAGE as an analysis of benchmark correctness with medical structure, providing the response-level evidence needed for a later clinical annotation layer.

Prompting is another part of the evaluation instrument. Clinical language models can be sensitive to natural changes in instruction phrasing (Ceballos-Arroyo et al., 2024). This matters for geometric analysis because a prompt affects not only accuracy but also the number of distinct responses, explanation length, and whether both answer classes appear. Our use of one documented answer-first request reduces uncontrolled prompt variation, while retaining the actual template or fallback used by each checkpoint. The resulting experiments support within-protocol analyses; they do not establish a template-independent ranking of medical models.

A.2 Repeated-response uncertainty and geometric analysis

Hallucination research distinguishes several forms of inconsistency between generated text, evidence, and external facts (Ji et al., 2023). TruthfulQA specifically tests whether models reproduce false beliefs and misconceptions, illustrating that fluent, repeatable answers can remain false (Lin et al., 2022). This differs from measuring agreement with a medical examination key, but it motivates testing concentrated errors alongside dispersed responses. Repeated sampling provides one way to expose instability without requiring a new external source for every response. SelfCheckGPT uses agreement among sampled outputs as a signal for factuality assessment (Manakul et al., 2023). Semantic uncertainty groups language outputs by meaning, and subsequent semantic-entropy work connects this approach to detecting unreliable generations (Kuhn et al., 2023; Farquhar et al., 2024). The common motivation is that lexical variation and semantic variation should not be conflated. In medicine, that distinction is amplified by abbreviations, near-synonyms, and short answers whose interpretation depends on the question.

APORIA approaches related questions through geometric organization: response embeddings, within- and cross-class distances, supervised projections, and annotation-efficient label propagation (Ricco et al., 2026). We directly use its pinned projection and propagation implementations rather than substituting a loosely similar nearest-neighbor method. We also report a separate nearest-five baseline

to make the algorithmic distinction visible. Our medical extension changes the experimental scope and interpretation: it records eligibility over all questions, compares biomedical and General embeddings, removes answer lines, groups duplicate texts, and connects the resulting measurements to internal states and evidence interventions.

Embedding choice deserves explicit treatment. Sentence-BERT established a practical framework for sentence-level representations (Reimers & Gurevych, 2019), and MiniLM offers an efficient transformer compression approach underlying a widely used General sentence encoder (Wang et al., 2020). Biomedical pretraining addresses vocabulary and distribution differences in medical text (Gu et al., 2021). The exact biomedical sentence checkpoint used here is a derived embedding model, not simply the original PubMedBERT pretraining checkpoint. We therefore pin the encoder repositories and pooling procedure as part of the measurement definition. A change in encoder can change which semantic distinctions are geometrically prominent; agreement between encoders is an empirical control rather than an assumption.

A.3 Internal representations, probes, and causal interventions

Linear probes provide a simple way to ask whether a property is accessible from intermediate representations (Alain & Bengio, 2016). However, probe accuracy depends on training support, control tasks, and the relationship between a decoded feature and the generator's actual computation (Hewitt & Liang, 2019; Belinkov, 2022). Work on latent knowledge and internal truthfulness has shown that activations can contain signals relevant to correct or truthful statements (Burns et al., 2022; Azaria & Mitchell, 2023). INSIDE studies internal-state diversity for hallucination detection, while semantic entropy probes seek a cheaper internal approximation to a repeated-generation uncertainty signal (Chen et al., 2024; Han et al., 2024). These results motivate internal analysis but do not remove the need to specify the prediction target and the moment at which information becomes available.

MIRAGE separates prompt-only question propensity from post-answer response correctness. A prompt-only state is shared by all stochastic draws of the same question and therefore cannot identify which of those draws will later be wrong without additional randomness-dependent information. A post-answer state may already encode the emitted answer. Holding out entire questions prevents direct question leakage, but it does not erase this temporal distinction. We consequently present the two sites as different tasks and add length and answer-format controls rather than treating their AUCs as interchangeable detector scores.

Causal methods require stronger tests. Causal abstraction formalizes the relationship between interpretable variables and internal interventions (Geiger et al., 2021). ROME uses causal tracing and weight editing to investigate factual associations (Meng et al., 2022), and inference-time intervention modifies activations to affect truthfulness (Li et al., 2023). Activation-patching methodology shows that corruption and metric choices can substantially alter conclusions (Zhang & Nanda, 2023). Our contribution is not a newly identified medical circuit. It is a controlled residual-restoration study integrated with medical output geometry, using fixed locations, sham checks, random controls, and explicit baselines. This supplies causal sensitivity evidence at a specified site while keeping circuit-level claims separate.

A.4 Statistical foundations and our position

Wasserstein distance provides a principled comparison between empirical distance laws (Peyre & Cuturi, 2019), while Fisher projection provides supervised separation through class means and within-class covariance (FISHER, 1936). Their use in a repeated-response setting introduces dependence: pairwise distances share responses, and responses share questions. We permute response labels within a question rather than treating distances as independent observations, retain finite permutation corrections (Phipson & Smyth, 2010), and report the exact multiple-testing family for Benjamini–Hochberg adjustment (Benjamini & Hochberg, 1995). Question-cluster bootstrap summaries follow the resampling principle of preserving the sampling unit (Efron & Tibshirani, 1993). These choices make the scope of evidence visible. Relative to prior work, MIRAGE is a medical triangulation study: it connects established tools through a common protocol and uses their disagreements to refine the interpretation of model error.

A.5 Explanation quality, clinical error structure, and evidence relevance

The distinction between selected-answer agreement and explanation validity has direct medical precedents. Chen et al. (2025) introduce challenging clinical question sets with expert-written explanations and evaluate generated explanations separately from answer selection. Their setting clarifies what an answer-label mask cannot establish: a rationale can retain medically meaningful or misleading content after its explicit choice token is removed. Clinical evidence assessment adds another axis. Wang & Chen (2026) evaluate questions grounded in guidelines and systematic reviews, where ambiguity and evidence strength differ from examination-key matching. Our masks and residual probes measure accessibility under the benchmark rule; extending their targets to these explanation and evidence judgments requires corresponding labels.

Structured errors also deserve tests beyond aggregate accuracy. BiasMedQA examines how clinically motivated bias prompts alter medical model behavior (Schmidgall et al., 2024), while repeated self-correction can preserve or introduce errors rather than reliably repair them (Zhan et al., 2026). These studies complement our observation of repeated, compact nonreference answer modes: repetition does not establish sound medical reasoning. MedScope supplies a lightweight behavioral benchmark covering efficiency, consistency and subject-wise variation (Bian & Cheng, 2026); our additional contribution is aligning those auditing concerns with response geometry, timed residual measurements and intervention controls. For patient-facing use, physician-led safety assessment identifies concerns that examination correctness does not capture (Draelos et al., 2026). None of our answer-key or synthetic-contrast labels is a substitute for such safety evaluation.

Mechanistic medical work increasingly connects internal features with domain-specific judgments. Liu et al. (2025) combine benchmark-integrity review, a medical reasoning-error taxonomy and sparse-autoencoder interventions. This is a complementary route toward interpretable error features, whereas our residual replacement tests operate at whole-block states and do not identify individual medical circuits. MedPAIR compares model-assigned evidence relevance with physician judgments (Hao et al., 2025); it motivates testing whether an intervention changes information that medical readers consider relevant, not merely whether an answer score moves. Our minimal pairs supply a controlled intermediate test by changing one stated relation or quantity with rule-derived contrast labels, while leaving physician-validated relevance as a distinct next step. Finally, Med-Scout studies geometric perception in medical images (Liu et al., 2026). Its anatomical and spatial notion of geometry differs from the distance laws of generated text in MIRAGE, and image-ablation effects alone do not establish that perceptual capability.

B More Details on Methodology

B.1 Preliminary definitions and edge cases

Correctness labels and parsing. Correctness is assigned only to the extracted answer under the benchmark's stated evaluation rule and does not validate the accompanying explanation. A response may therefore match the reference while containing unsupported reasoning, or mismatch the reference while containing medically relevant content. For open-ended visual answers, normalization removes superficial formatting variation but does not adjudicate synonyms, equivalent anatomical descriptions, or other clinically equivalent formulations. We use "incorrect" throughout to mean benchmark mismatch. Unparsed responses receive no correctness label for geometric or response-level prediction analyses but remain in the denominator when reporting overall reference-match rates. Questions with no parsed responses are excluded from propensity prediction but retained in coverage reporting. A question with responses from only one correctness class is likewise retained for coverage but does not yield a two-class geometry statistic.

Response views and geometry eligibility. The encoder is applied to either the complete response or a view in which the explicitly formatted answer line is removed while retaining the rationale. Embeddings are normalized before distance analysis. For each question,

$$G_q = \{x_i : y_i = 0\}, \qquad H_q = \{x_i : y_i = 1\},$$

and geometry is computed only when $|G_q|, |H_q| \geq 5$. This threshold is a sampling requirement that ensures both within-class distance laws contain sufficient pairs; it is not a clinical or statistical

significance criterion. Missing source fields and lexical feature matches are not interpreted as evidence of clinical absence.

Distance-law interpretation. For completeness, we define the cross-class distance multiset $D(G_q, H_q)$ and its empirical law P_{GH}, although the primary dispersion analysis compares only the within-class laws P_G and P_H. The Wasserstein statistic $W_q = W_1(P_G, P_H)$ measures transport between distance distributions rather than between the two response point clouds. It is unsigned and therefore does not identify which class is more dispersed. The signed gap Δ_q supplies this direction, with $\Delta_q > 0$ indicating greater mean within-class dispersion for incorrect responses. HII provides a scale-normalized contrast but can become unstable when both class medians approach zero; we therefore interpret it jointly with the raw distance statistics, duplicate structure, and the underlying empirical laws.

Propensity target. The prompt-only residual is shared across stochastic generations of the same question–evidence condition and therefore cannot encode the realized correctness of an individual sampled response. Its target z_q is consequently defined at the question level from the observed sampled responses. A tie in the incorrect-response fraction is assigned to the non-majority-incorrect class. This target describes the finite sampled cohort and should not be interpreted as an oracle probability of model error.

Intervention score. The intervention score is the mean teacher-forced log probability of a fixed reference answer sequence. Restoration effects compare the score after replacing a corrupted residual with the corresponding clean residual against the corrupted-input baseline. Matched equal-norm random perturbations provide a specificity control. These quantities measure sensitivity of a reference-answer score to a specified internal intervention; they are not measures of sampled-answer accuracy. In particular, when the evidence is deliberately corrupted or replaced, the original benchmark reference is not assumed to remain clinically correct.

Interpretation of benchmark correctness. All analyses in this work use the clinician-verified/source-validated benchmark annotations supplied with the selected datasets as the operational target. They provide a defensible reference for this benchmark audit, while still not establishing factual completeness, explanation quality, patient-level safety, or deployment performance. No independent re-annotation is required for the present study; controlled semantic contrasts remain useful for attributing a signal to a particular clinical feature.

B.2 Controlled analyses and robustness checks

B.2.1 Medical views of the output space

The medical analysis combines source fields with transparent lexical features. Available source fields include specialty, anatomy, modality, task, and answer type; missing values remain unknown. Versioned regular expressions identify negation, laterality, quantity with unit, treatment, diagnosis, cardiorespiratory, neurologic, and abdominal terms in question text. Features may overlap, and a missing match does not imply clinical absence. We ask which categories appear in the cohort and support mixed-class geometry, how distance contrasts vary across supported categories, and whether activation or intervention effects concentrate where answer formats or evidence requirements differ. Because categories are neither randomized nor independent of dataset or checkpoint, we report counts before means and treat these analyses as descriptive. An expert-review queue retains unfilled annotation fields rather than assigning clinical severity or error subtype automatically.

B.2.2 Residual capture and question-disjoint probes

We capture decoder-block residuals at two sites: a prompt-only state from a separate pass with no generated token, computed once per question–evidence condition and reused across draws, and a post-answer state from the prompt plus the first eight generated tokens, or the available shorter prefix. We store residuals in FP32 despite BF16 inference to avoid overflow from narrower FP16 storage, and retain site, token count, and layer index with each response. At each layer, states are unit-normalized, standardized on calibration data, and evaluated with class-balanced logistic regression ($C = 1$, at most 1,000 iterations): prompt-only fitting uses eight calibration questions, post-answer fitting uses parsed responses from the same questions, and evaluation uses 400 disjoint questions. We report AUC, macro-F1, accuracy, and question/image-cluster bootstrap intervals, with

a log-token-count baseline for post-answer states under the same split. Prompt-only nulls shuffle evaluation labels across questions, whereas post-answer nulls shuffle labels within each evaluation question while preserving its sampled error rate; thus the latter null need not have AUC $= 0.5$. We use 199 fixed-score permutations and 200 cluster-bootstrap replicates with layerwise adjustment within each cell, report complete curves alongside fixed midpoint summaries, and never select layers by evaluation performance.

B.2.3 Transfer, evidence controls, and residual restoration

For cross-dataset transfer, we fit post-answer probes on eight calibration questions from one dataset and evaluate them on 400 questions from another using the same generator, at depths of $1/4$, $1/2$, and $3/4$ of the decoder stack (rounded down over zero-based indices), yielding 120 directed comparisons. Image controls compare original, blank, different frozen, and text-only inputs with four paired seeds per question; the primary image-control subset uses 40 evaluation questions per visual cell, with four calibration controls retained. We separately measure parsed-answer changes and original-reference agreement, treating changes as evidence of input dependence rather than correctness. Restoration uses the fixed 40-question evaluation control subset per compatible cell plus eight calibration questions. Visual corruption uses a blank image; text corruption zeros non-special prompt-token embeddings except the final prompt position. At each depth, we restore the clean final-prompt residual, reinsert the corrupted residual as a sham, or add an equal-norm random vector to corrupted or clean input. We score the reference answer without its formatting prefix and report restoration over the corrupted baseline and specificity over matched random perturbation.

B.2.4 Review-stage stability and selective evidence controls

Review-stage controls are sensitivity analyses rather than a new confirmatory cohort. At the midpoint layer, 20 fixed partitions fit on eight independent question/image groups and evaluate on the other 400, without stratification or class-based selection; degenerate targets are retained. We compare answer-type-only, residual-only, and concatenated propensity models, and a 256-dimensional fixed unigram/bigram hash plus log response length against residuals for post-answer controls. All probes use unit residual normalization, train-only standardization, balanced logistic loss, and $C = 1$; review fits use deterministic LIBLINEAR with at most 2,000 iterations, while the original atlas keeps its original solver. Direction stability uses normalized coefficients after undoing standardization. Ninety-nine full-refit nulls shuffle pre-answer labels across partitions or post-answer labels within questions and refit the scaler and classifier. Split quantiles capture fitting-support sensitivity, and question-cluster bootstrap intervals quantify conditional uncertainty; overlapping splits are not independent replications. The 18-pair medical-language suite covers laterality, presence/absence, and dose arithmetic, with three wordings and both answer assignments. Labels follow the stated text or arithmetic rule. We score both candidates and their source-versus-counterfactual log-probability margin, inserting source residuals into counterpart prompts at three depths. Controls reinsert the target state, add an equal-norm random delta, or use a residual from a prompt differing only by an irrelevant record identifier; states are captured under matched teacher-forced contexts. We test final-prompt placement first and, after weak effects, additionally patch the last token of the edited evidence span using token alignment; both locations and all depths are retained.

C More Details on Experimental Design

C.1 Checkpoints and compatible task matrix

The study includes MedGemma-4b-it, MedGemma-1.5-4b-it, Lingshu-7B, BioMistral-7B, and OpenBioLLM-8B. The first three accept images and run on both text and visual datasets. BioMistral and OpenBioLLM run on the two text datasets. The resulting 16 cells allow two complementary comparisons: shared text questions across five checkpoints, and shared visual questions across three checkpoints. We do not construct text-only surrogates for image questions for the text models, because that would change the evidence available in the primary task. Text-only visual inputs instead appear as a controlled condition for the VLMs.

The five checkpoints represent several model families, but they are not five independent architectures. The two MedGemma releases share a family, and model specialization, instruction format, and

pretraining history differ across the collection. Checkpoint identity is therefore the unit of model comparison. Every repository is pinned to a recorded revision before generation. We retain the native Transformers implementation and do not rely on an unversioned remote model adapter. Gappendix D lists model identifiers and revisions, allowing the complete task matrix to be reconstructed without guessing from display names.

These are small models relative to many contemporary medical systems, yet they remain large enough for nontrivial internal-state analysis. The design prioritizes repeated response collection and intervention access over maximizing benchmark accuracy. It does not compare a medical checkpoint against its matched General base model, and it does not isolate the causal effect of medical fine-tuning. Accordingly, we describe results as properties of the selected medical checkpoints under the experiment, rather than properties caused by medical specialization itself.

C.2 Dataset selection and frozen cohort construction

The text datasets are the validation split of MedMCQA and the test split of the four-option English MedQA-USMLE release. The visual datasets are the test split of VQA-RAD and the English subset of the SLAKE test split. Dataset repositories, revisions, source indices, normalized question hashes, and image hashes are recorded in a cohort lock. Each dataset contributes 408 questions, with eight designated for calibration and 400 for evaluation. This yields 1,632 unique questions overall and 1,600 unique evaluation questions, shared across compatible checkpoints.

Question hashes prevent duplicate normalized questions from crossing the study partitions. For image tasks, sampling uses one question per selected image and keeps image groups disjoint between calibration and evaluation. The image hashes identify the actual frozen evidence files rather than relying only on dataset row numbers. The MedMCQA cohort excludes the questions used in the earlier pilot. These safeguards protect the experimental split within this study; they do not establish that a released checkpoint never encountered benchmark material during training.

The source datasets differ in answer structure. All text questions are four-option MCQs. The visual evaluation cohort retains the source answer-type fields; the 40-question image-control subset is separately reported because it was selected for paired intervention capacity. This variation is useful for medical coverage but also creates an evaluation confound: normalized exact match is much easier to satisfy for a binary answer than for a free-text anatomical description. We therefore report answer-type support and add a post hoc type-only propensity baseline when interpreting the activation results. The baseline uses only the same calibration labels; its introduction during manuscript review is explicitly distinguished from the frozen primary protocol.

C.3 Generation, prompting, and hardware

Every primary question–checkpoint pair receives 32 sampled responses, using temperature 1.0, top-p 0.95, and a maximum of 192 new tokens. The prompt requests an answer first and a short explanation afterward. This format supports conservative answer extraction while retaining a rationale view when the model emits a separate explanation. Sampling uses recorded seeds derived from the frozen study configuration. No best-of-n selection or correctness-based resampling replaces the collected primary responses. Failed parsing and truncation remain observable outcomes of the same protocol.

Generation uses BF16 model inference with Eager attention on AMD MI210 GPUs with 64 GB memory, one GPU per cell and up to four cells concurrently. Images are loaded from frozen files and thumbnailed to fit within a 448-pixel bound before the native processor. The processor may apply its own subsequent image transformation. This is a controlled image-input protocol rather than an evaluation of every model's largest supported image resolution. The prompt-only pass, generated-prefix replay, and intervention forwards run with the model in evaluation mode, with internal states saved as FP32 arrays.

When an available tokenizer or processor supplies a chat template, the backend uses it. Otherwise, it falls back to the raw prompt with an `Answer:` continuation. The recorded OpenBioLLM run used this fallback, although the model card recommends an explicit Llama-3 instruction template (Ankit Pal, 2024). Its results therefore describe the collected fallback protocol and should not be read as a release-quality benchmark score. The same provenance is relevant to its strong restoration

response under broad text corruption. We retain the completed run as a protocol-specific observation rather than silently changing the prompt during manuscript preparation.

C.4 Parsing, representations, and implementation settings

MCQ parsing accepts an explicit answer marker followed by a letter A–D, with a conservative bare-letter continuation fallback. VQA parsing uses an explicit answer line when present or otherwise the first output line. Leading yes/no forms are reduced to the corresponding binary answer without consulting the reference. Other answers undergo Unicode normalization, lowercasing, whitespace normalization, and limited surrounding punctuation removal. Decimal points, minus signs, and units are preserved rather than aggressively stripped. Unparsed MCQ responses remain failures for the all-draw reference-match rate and are excluded from binary-label geometry and probe fitting.

The full-response encoder views use all parsed response text. The rationale view removes the first line and any remaining explicit answer lines, then excludes empty results. It therefore isolates *separately formatted rationale text*; it does not capture every explanation that a model places on the same line as its answer. Nor does it remove every reference to the chosen option or medical conclusion. This partial control is still informative because it tests whether the prominent answer line accounts for a substantial part of observed separability, but it must be interpreted with its changed text support.

The biomedical encoder is `NeuML/pubmedbert-base-embeddings`, and the General encoder is `sentence-transformers/all-MiniLM-L6-v2`. Both use attention-mask-weighted mean pooling and unit normalization, with an encoder input cap of 512 tokens. Their dimensions are 768 and 384, respectively. The generators and encoders have separate tokenizers and token caps. Thus a 192-token generated response is not guaranteed to occupy the same number of encoder tokens. Revisions and pooling settings are fixed and recorded so that geometric measurements can be reproduced with the same representation definition.

Geometry uses 499 permutations, activation geometry uses 99 permutations at the three designated depths, and probe inference uses the settings in Section 4. Response-count controls use only observed prefixes of 8, 16, and 32 responses. Larger nominal counts in the analysis interface remain unavailable; no 100- or 150-response main result is reported. The earlier 100-response pilot is separate and is not pooled with the frozen main cohort.

C.5 Accounting, statistical units, and reproducibility

The primary collection contains $16 \times 408 \times 32 = 208{,}896$ responses. The evaluation-only subset contains 204,800 responses from 6,400 question–model pairs; calibration contributes 4,096 responses. The additional paired image controls are a deliberately smaller 40-question evaluation subset per visual cell with four paired seeds per condition; their image-group denominators are reported in the control tables. The separate sampled intervention follow-up contains 3,200 responses across 80 questions. None of these short control studies is pooled into the primary 400-question geometry or probe denominator.

Intervention accounting is distinct from response generation. Sixteen cells, eight questions per cell, three fixed depths, and four intervention conditions yield 1,536 score measurements. Half are on evaluation questions. The 120 transfer comparisons are fitted-probe evaluations, not additional generated responses or independent medical cases. We preserve these distinctions because collapsing all counts into a single sample size would overstate the amount of independent evidence. For medical generalization, the number and diversity of questions matter alongside the much larger number of stochastic draws.

All setup, test, generation, analysis, and manuscript build commands retain persistent supervised logs with timestamps, command metadata, stdout, stderr, exit status, source hashes, and resource samples. The primary array completed all 16 cells successfully. Completion markers and 1,079 analysis output hashes were checked before manuscript construction. A timestamped research journal records plans, failures, corrections, and interpretation. Manuscript-stage computations create new artifact directories and retain the original results, so sensitivity analyses can be compared against the primary analysis rather than overwriting it.

C.6 Revision computation and provenance

The revision reuses the frozen generated responses and decoder states for sixteen-cell re-encoding, threshold sensitivity, masked-answer Fisher tests and repeated fitting. It does not change their question or response counts. Corrected full, rationale and explicit-answer-masked views use the same pinned biomedical and General encoders. The added mask replaces an explicit parsed answer label with a shared placeholder while preserving same-line explanation text; it uses the prediction, never the reference answer. Missing separate rationales and collapsed masked-text groups are reported as unsupported views, not silently replaced with other text.

New generator execution covers MedGemma-4b-it, BioMistral and OpenBioLLM on eighteen synthetic contrast pairs. Two directions, three depths and four patch conditions produce 432 margin measurements per model and site, or 2,592 across three models and two sites. The sites share pairs and are not independent datasets. OpenBioLLM additionally reruns the original sixteen broad-intervention questions under raw serialization and Llama-3 user-message serialization, keeping user instructions, seeds, four sampled draws and maximum 192 tokens fixed. This produces 128 template-control responses and 384 broad-intervention measurements. The serialization control does not add the model-card example's system instruction and is therefore a controlled format comparison rather than a replication of its full deployment recipe. Original, failed, corrected and exploratory runs occupy distinct folders with completion markers, commands, errors, source hashes and timestamps.

D Study design and reproducibility

The main study uses 400 evaluation questions and eight calibration questions per dataset, with 32 responses per compatible question and checkpoint pair. Supplementary control studies use smaller cohorts with separately reported settings and denominators. This appendix organizes implementation, coverage, geometry, prediction, interventions, and numerical audits by topic. Corrected embeddings support the primary geometry findings; historical outputs are identified explicitly.

D.1 Prompt requests and generation policy

For MCQs, the question is followed by options in their original A-D order and the instruction:

> Give the answer first as "Answer: A" (or B, C, D), then explain briefly in at most three sentences.

For visual questions, the instruction is:

> Give a short answer on the first line as "Answer: <short answer>", then explain briefly in at most three sentences.

Native chat templates are applied when available. A raw-prompt continuation ending in `Answer:` is used otherwise. The full rendered prompt for every question and condition is retained. The prompt asks for a separate short explanation, but the parser and rationale-view code do not assume that the model complied. Raw text is never replaced with a cleaned paraphrase before storage.

Each response stores its actual seed, and the generation driver derives response seeds deterministically from the frozen question order and draw index. Paired image conditions reuse the corresponding seed values. Temperature is 1.0, top-p is 0.95, and maximum new tokens is 192. The model is placed in evaluation mode. The primary response count is 32 for every one of 408 questions in every compatible cell. Four draws per altered image condition are collected for the predeclared 40-question evaluation control subset of each visual cell, with calibration controls retained separately. No question is selected for the control subset after examining its correctness.

D.2 Core parameters

Component	Recorded setting
Generator inference	BF16; Eager attention; native Transformers; one MI210 64 GB per cell
Image preparation	RGB; aspect-preserving thumbnail within 448 by 448; native processor afterward
Primary sampling	32 draws; temperature 1.0; top-p 0.95; at most 192 generated tokens
Encoder inference	Mean pooling; unit normalization; at most 512 tokens; corrected runs use FP32 Eager attention, fixed padding, and unique text encoding
Geometry eligibility	At least five parsed responses in each correctness class
Geometry null	499 response-label permutations; fixed seed; finite $(1 + k)/(1 + B)$ correction
Fisher covariance	Sum of biased class covariances; trace-scaled ridge; primary $\lambda = 10^{-3}$
LP splitting	Up to three response splits; three labeled-subset draws per split; budgets 5, 10, 20
Grouped splitting	Whitespace-normalized representation-specific text; up to 100 candidate group splits
LP annotation support	Two labeled responses per class reserved, then remaining budget sampled
Probe states	Unit-normalized FP32 stored residuals; calibration-fitted standardization
Probe estimator	Class-balanced logistic regression; $C = 1$; maximum 1,000 iterations
Probe uncertainty	200 question/image-cluster bootstrap draws; 199 fixed-score permutations
Prompt-only target	One row per question; majority nonmatching among parsed responses
Post-answer site	Last position after replaying at most the first eight generated tokens
Intervention depths	$\lfloor .25(L-1) \rfloor$, $\lfloor .5(L-1) \rfloor$, $\lfloor .75(L-1) \rfloor$
Intervention site	Final prompt token, at a decoder-block output
Intervention outcome	Mean teacher-forced original-gold log probability, excluding formatting prefix
Text corruption	Zero non-special prompt embeddings except final prompt position
Visual corruption	Blank image of the same original dimensions, followed by the same preprocessing
Controls	Corrupted-state sham; equal-norm random perturbation; random perturbation on clean input

D.3 Supplementary control cohorts

The following supplementary studies use smaller cohorts for paired controls and computational tests:

- Image dependence uses 40 evaluation questions per visual model-dataset cell, four paired seeds, and image-group intervals. The full 400-question visual cohort is used for primary generation, geometry, and probes.
- Broad residual restoration uses the fixed 40-question evaluation subset per compatible cell, three depths, and sham/random controls. Sham measurements are intervention records rather than independent medical cases.
- The sampled restoration follow-up uses 80 questions total (40 VQA-RAD and 40 SLAKE), eight seeds and five conditions, for 3,200 responses. It is a separate behavioral control study.
- The medical-language minimal-pair suite contains 18 synthetic prompt instances with three wordings per feature. It tests rule-derived counterfactual margins and is not clinical validation.

Table 5: **Evaluation coverage and biomedical full-response geometry.** Eligibility requires at least five parsed draws in each correctness class. Positive gap counts eligible pairs with greater incorrect-response dispersion.

Model	Data	Questions	Groups	Match (%)	Unparsed	Eligible	Positive gap	BH < .05
MedGemma 4B	MedMCQA	400	400	51.992	0	11	4	1
MedGemma 4B	MedQA	400	400	51.539	0	21	9	0
MedGemma 4B	VQA-RAD	400	195	54.586	0	23	16	1
MedGemma 4B	SLAKE	400	88	56.836	0	35	19	0
MedGemma 1.5 4B	MedMCQA	400	400	37.711	738	19	13	0
MedGemma 1.5 4B	MedQA	400	400	38.672	1214	26	14	2
MedGemma 1.5 4B	VQA-RAD	400	195	47.531	0	61	32	7
MedGemma 1.5 4B	SLAKE	400	88	40.758	0	79	50	7
Lingshu 7B	MedMCQA	400	400	54.453	0	12	4	7
Lingshu 7B	MedQA	400	400	59.594	0	17	11	13
Lingshu 7B	VQA-RAD	400	195	44.992	0	9	3	2
Lingshu 7B	SLAKE	400	88	38.070	0	3	1	0
BioMistral 7B	MedMCQA	400	400	37.969	205	230	164	38
BioMistral 7B	MedQA	400	400	39.211	44	164	124	28
OpenBioLLM 8B	MedMCQA	400	400	30.555	742	41	35	19
OpenBioLLM 8B	MedQA	400	400	34.227	426	57	46	24

E Evaluation coverage and medical descriptors

E.1 Question support and response behavior

Table 6 distinguishes unique questions, question-model pairs, and generated responses. Table 7 shows how repetition and truncation differ across cells. These summaries should be read before comparing a conditional geometry average between models. In particular, a single-class question can be either consistently matching or consistently nonmatching, and a text model's higher mixed-class coverage can reflect greater response variability rather than superior error detection.

Table 6: Evaluation accounting by dataset. Shared questions across checkpoints and questions on the same image are not counted as independent medical cases.

dataset	Questions	Groups	Pairs	Responses	Correct	Unparsed	Truncated	Eligible
MedMCQA	400	400	2000	64000	27223	1685	2333	313
MedQA	400	400	2000	64000	28575	1684	9572	285
SLAKE	400	88	1200	38400	17365	0	28	117
VQA-RAD	400	195	1200	38400	18830	0	69	93

Table 7: Exact-text repetition, single-class questions and truncation by evaluation cell.

Model	Data	Mean duplicate fraction	Single class	Truncated
MedGemma 4B	MedMCQA	0.002	376	70
MedGemma 4B	MedQA	0.000	356	479
MedGemma 4B	VQA-RAD	0.000	359	0
MedGemma 4B	SLAKE	0.000	343	0
MedGemma 1.5 4B	MedMCQA	0.001	364	933
MedGemma 1.5 4B	MedQA	0.000	342	5283
MedGemma 1.5 4B	VQA-RAD	0.001	291	37
MedGemma 1.5 4B	SLAKE	0.001	267	13
Lingshu 7B	MedMCQA	0.904	388	134
Lingshu 7B	MedQA	0.879	383	312
Lingshu 7B	VQA-RAD	0.948	389	32
Lingshu 7B	SLAKE	0.939	397	15
BioMistral 7B	MedMCQA	0.133	57	863
BioMistral 7B	MedQA	0.141	135	1494
OpenBioLLM 8B	MedMCQA	0.299	300	333
OpenBioLLM 8B	MedQA	0.230	289	2004

The lexical feature set is deliberately simple and versioned. Negation includes terms such as no, not, without, absent, negative, except, denies, and neither. Laterality includes left, right, bilateral,

unilateral, ipsilateral, and contralateral. Quantity-with-unit matches numeric expressions followed by the implemented unit alternatives, including mg, mcg, g, kg, ml, mmol, mmhg, cm, mm, units, and a percent pattern. Because the implementation includes word-boundary conditions, lexical coverage should be reproduced from code rather than inferred from this prose description. Treatment uses stems for treatment, therapy, administration, prescribing, and surgery plus drug/dose terms. Diagnosis uses diagnosis, disease, disorder, syndrome, finding, and abnormality stems. Anatomy-related slices use the explicit organ and system terms in the versioned source file.

The indicators operate on question text, not generated response text. They therefore define the context being answered rather than automatically identifying the content of an error. They are case-insensitive lexical matches, not a clinical entity linker or ontology reasoner. Multiple indicators can be true for one question. Unknown source specialty, anatomy, or modality stays unknown, even when a lexical match supplies a plausible contextual clue. Table 8 records the resulting support and geometry summaries, and Table 9 isolates the substantial open/closed format difference.

Table 8: Medical lexical strata on evaluation questions. Features overlap; means describe this sampled cohort and do not estimate causal feature effects.

Feature	Unique questions	Question-model pairs	Eligible	Mean accuracy	Median W1	Positive gap
negation	351	1751	232	0.403	0.138	165
laterality	218	958	123	0.407	0.128	94
quantity_unit	213	1065	152	0.444	0.152	110
treatment	165	821	122	0.455	0.155	89
diagnosis	281	1215	156	0.416	0.143	119
cardiorespiratory	230	930	121	0.470	0.107	85
neurologic	83	335	39	0.413	0.143	23
abdominal	185	759	83	0.469	0.130	58

Table 9: Answer-type coverage and eligibility. Open VQA answers use conservative normalized exact match.

Data	Answer type	Unique questions	Pairs	Accuracy	Eligible
MedMCQA	mcq	400	2000	0.425	313
MedQA	mcq	400	2000	0.446	285
VQA-RAD	closed	222	666	0.741	44
VQA-RAD	open	178	534	0.178	49
SLAKE	closed	130	390	0.821	20
SLAKE	open	270	810	0.275	97

E.2 Supplementary cohort size sensitivity

This supplementary analysis uses the original 40 question cohort, not the main 400 question cohort. We vary the number of distinct evaluation questions while keeping each retained response set fixed. For each dataset and size 10, 20, 30 or 40, we draw 500 subsets without replacement using a fixed seed. The same selected question/image groups are retained across checkpoints. We report geometry eligibility among all selected question-model pairs and positive dispersion among eligible pairs.

These curves measure sensitivity within the observed cohort. Their range necessarily collapses at forty because every subset is the full observed set; this is a finite-population identity, not evidence that sampling uncertainty has vanished. They cannot establish what happens at eighty new evaluation questions, and repeating the existing forty questions or draws would not answer that question.

The dataset breakdown also matters. At the full retained cohort, MedMCQA contributes 39 eligible pairs (29 positive), MedQA 31 (23 positive), VQA-RAD 11 (8 positive), and SLAKE only 7 (3 positive). Thus 70 of the 88 eligible pairs are from text datasets. The pooled majority result should not be read as a majority result in every visual dataset: SLAKE's observed direction is positive in fewer than half its very small eligible set. Among the 500 ten-question SLAKE subsets, 97 contain no eligible pair at all. We retain those undefined fractions rather than replacing them with zero or omitting their frequency from the support table.

Table 10: Five hundred subsets of independent question/image groups per size from the frozen evaluation cohort, retaining all selected-image questions and reusing selection across checkpoints. Percentiles describe finite-cohort subsampling variation. At full support every subset is the same cohort; its zero range is not a population confidence interval. Undefined counts record subsets with no eligible pair.

Dataset	Groups	Eligibility	Eligibility_low	Eligibility_high	Positive	Positive_low	Positive_high	Undefined
MedMCQA	10	0.157	0.080	0.240	0.709	0.375	1.000	0
MedMCQA	20	0.155	0.100	0.220	0.702	0.460	0.923	0
MedMCQA	30	0.155	0.107	0.203	0.702	0.523	0.881	0
MedMCQA	40	0.158	0.115	0.200	0.704	0.552	0.840	0
MedMCQA	80	0.156	0.130	0.181	0.704	0.601	0.808	0
MedMCQA	160	0.157	0.139	0.172	0.705	0.645	0.769	0
MedMCQA	400	0.157	0.157	0.157	0.703	0.703	0.703	0
MedQA	10	0.140	0.060	0.240	0.708	0.250	1.000	0
MedQA	20	0.142	0.080	0.200	0.720	0.500	0.923	0
MedQA	30	0.143	0.093	0.193	0.703	0.508	0.886	0
MedQA	40	0.143	0.102	0.188	0.711	0.538	0.863	0
MedQA	80	0.144	0.113	0.175	0.718	0.616	0.822	0
MedQA	160	0.143	0.126	0.163	0.718	0.656	0.781	0
MedQA	400	0.142	0.142	0.142	0.716	0.716	0.716	0
SLAKE	10	0.097	0.050	0.147	0.601	0.364	0.857	0
SLAKE	20	0.097	0.067	0.130	0.601	0.453	0.759	0
SLAKE	30	0.097	0.072	0.122	0.601	0.487	0.733	0
SLAKE	40	0.098	0.080	0.117	0.598	0.509	0.689	0
SLAKE	80	0.098	0.092	0.102	0.599	0.573	0.627	0
SLAKE	88	0.098	0.098	0.098	0.598	0.598	0.598	0
VQA-RAD	10	0.078	0.015	0.158	0.551	0.000	1.000	7
VQA-RAD	20	0.078	0.032	0.128	0.540	0.250	0.857	0
VQA-RAD	30	0.079	0.039	0.122	0.553	0.312	0.778	0
VQA-RAD	40	0.077	0.044	0.112	0.547	0.341	0.739	0
VQA-RAD	80	0.078	0.058	0.099	0.549	0.432	0.667	0
VQA-RAD	160	0.077	0.069	0.085	0.548	0.500	0.595	0
VQA-RAD	195	0.077	0.077	0.077	0.548	0.548	0.548	0

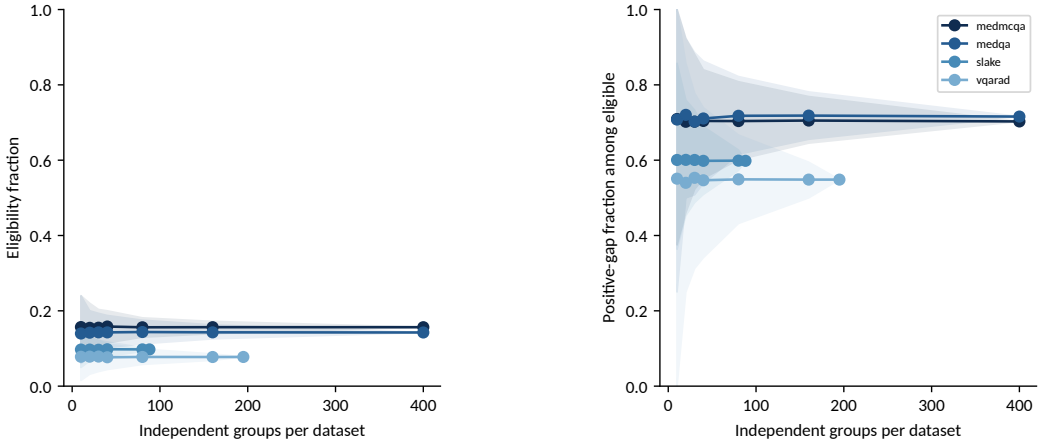


Figure 7: Finite-cohort sensitivity with 32 responses per question. Subsets select independent groups and include all questions for selected images; this is not a new population sample.

A valid extension freezes independent question/image groups, excludes overlap with all existing calibration/evaluation groups, preserves checkpoints, decoding, parser, corrected encoding and threshold families, and reports supplementary and main cohorts separately before any pooled analysis. The present 400-question extension satisfies the frozen-cohort and provenance requirements; it evaluates the same fixed probes and retains a separate calibration-only fitting design. The cohort curves therefore quantify finite-cohort sensitivity, while the primary 400-question results provide the main support. Category-specific clinical conclusions remain constrained by the observed lexical and image-group coverage.

F Response geometry and answer content

The main analysis requires at least five parsed correct and five parsed incorrect responses for a question-model pair. Of 6,400 main evaluation pairs, 808 satisfy this criterion. In the corrected PubMed full-response view, 545/808 eligible pairs have positive incorrect-minus-correct dispersion gaps. The dataset counts are 220/313 for MedMCQA, 204/285 for MedQA, 70/117 for SLAKE, and 51/93 for VQA-RAD. Pairs with one parsed class or insufficient support remain in coverage tables and are not assigned zero effects.

Generation behavior affects support. The main dataset totals are 64,000 primary evaluation draws for each text dataset and 38,400 for each visual dataset, before compatibility weighting is summarized in the cell-level table. Unparsed and truncated responses are retained as outcomes; they are excluded only where a parsed correctness label is mathematically required.

Threshold and representation sensitivity. The threshold family, corrected encoder comparison, and all four representation views are shown in Figures 18 and the linked tables. Positive-gap fractions remain conditional on eligibility and vary with metric, encoder, and minimum class count. The encoder incident is preserved: Eager attention, fixed padding and FP32 re-encoding are used for the reported estimates, while the original SDPA outputs and replay diagnostics remain available for audit.

APORIA-style exemplar. The retained Lingshu/MedQA question 322 is reused without selecting a new case after correction. Its distance laws and Fisher projection are descriptive; the generated figure is a close reading of one response cloud, not an estimate of a population effect.

The prefix and label-budget plots use nested subsets of the main responses. They do not create independent cohorts, and the maximum observed primary prefix remains 32 responses. Fisher and label-propagation outputs are reported with their available support rather than extrapolated to unsupported response counts.

F.1 Projection and label propagation definitions

For a labeled training set, the APORIA Fisher direction is

$$v = \frac{\tilde{v}}{\|\tilde{v}\|_2}, \qquad \tilde{v} = \left(S_G + S_H + \lambda \frac{\mathrm{tr}(S_G + S_H)}{d} I \right)^{-1} (\mu_H - \mu_G), \tag{1}$$

with biased empirical covariance matrices S_G, S_H. If the trace per dimension is below 10^{-12}, the upstream implementation raises a degeneracy error rather than manufacturing a direction. Such skipped fits are retained. Projected coordinates are $z_i = x_i^\top v$.

For an incoming point z, the Wasserstein propagator compares the distribution of distances from z to the training points in each class with that class's training within-class distance distribution. If the two resulting scores are $s_G(z)$ and $s_H(z)$, prediction is incorrect when $s_H(z) < s_G(z)$. The centroid propagator instead compares distance to projected class centroids. The separate MIRAGE-nearest-five baseline averages up to five nearest training distances in each class in the original encoder space. It is not labeled as the APORIA algorithm. The majority baseline uses the annotated subset's majority label with the implementation's deterministic tie handling.

F.2 Text views and paired comparisons

The main Fisher table's repeated encoder cells are typographically elided, rather than missing observations. Table 11 supplies a fully populated version. Its intervals concern the average trial

score after resampling question/image groups; they are not the minimum and maximum trial scores. "General" denotes the pinned all-MiniLM-L6-v2 checkpoint, while "PubMed" denotes the pinned PubMedBERT embedding checkpoint. The full, masked and rationale rows use the same supported question-model pairs, but each text view defines its own duplicate groups.

Table 11: Fully populated companion to the main masked-answer table: no blank repeated encoder cells. Confidence intervals resample independent groups after averaging trials within a question–model pair. Full, masked and rationale labels describe encoder input views.

Encoder	View	Mean F1	95% low	95% high	Groups	Pairs
PubMed	full	0.807	0.787	0.828	210	278
PubMed	masked	0.670	0.644	0.696	210	278
PubMed	rationale	0.666	0.639	0.693	210	278
General	full	0.818	0.799	0.838	210	278
General	masked	0.672	0.646	0.701	210	278
General	rationale	0.665	0.639	0.693	210	278

For clarity, the transport integral and directional mean in the main definitions are

$$W_q = \int_{\mathbb{R}} |F_G(t) - F_H(t)| \, \mathrm{d}t, \qquad \Delta_q = \mathbb{E}_{d\sim P_H}[d] - \mathbb{E}_{d\sim P_G}[d]. \tag{2}$$

The empirical units in these formulas are within-class distances, but the permutation unit is a response label and the bootstrap unit is a question or image group. Pairwise distances that share a response are not independent observations. Neither the number of distances nor the number of sampled completions substitutes for the number of independent medical cases.

The explicit-answer mask replaces the parsed answer marker while preserving same-line explanation text. It is a partial content control: diagnoses, option-associated words, hedging, and answer references can remain. The 278 matched pairs and both encoders are reported in the main Fisher tables. Full-minus-masked macro-F1 differences are 0.137 for PubMed and 0.146 for the General encoder; intervals and all representation views are retained in the generated tables.

Table 12: Paired corrected Fisher contrasts with question/image-cluster intervals. Masking changes duplicate groups; these are matched-question, not identical-response-fold contrasts.

encoder	contrast	mean	lo	hi	clusters	rows
PubMed	full minus masked	0.137	0.117	0.158	210	278
PubMed	full minus rationale	0.141	0.120	0.163	210	278
General	full minus masked	0.146	0.124	0.167	210	278
General	full minus rationale	0.153	0.131	0.174	210	278

F.3 Response counts and projection controls

The historical raw-distance atlas uses a deterministic example-selection rule: the first eligible evaluation question in sorted identifier order for each cell. It displays within-correct, within-incorrect, and cross-class empirical distributions without choosing the largest effect. This complements the analytically selected main-text example. Cross-class separation can be large even when the two within-class laws are similar, as the proposition in Section 3 explains.

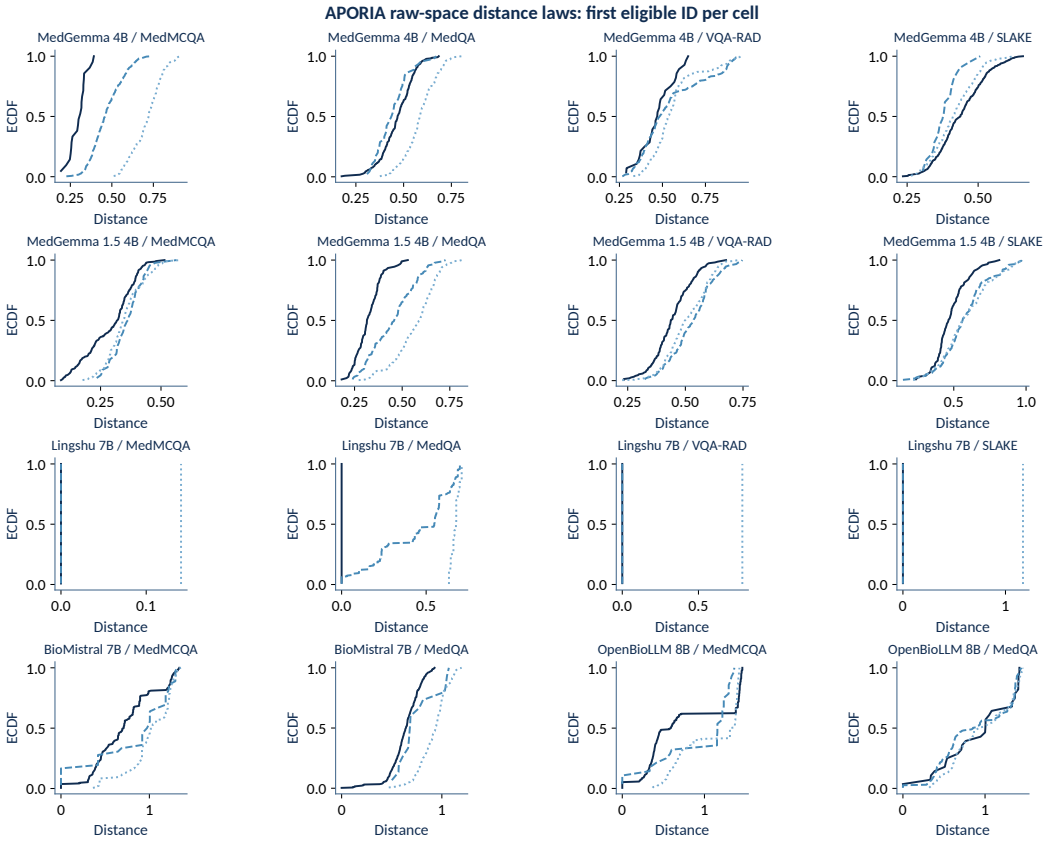


Figure 8: Raw correct–correct (navy solid), incorrect–incorrect (blue dashed), and cross-class (light-blue dotted) distance laws. Exemplars use the first eligible evaluation question by sorted ID in each cell, independently of effect size.

Table 13 reports support for the principal label-propagation comparison. Full-response and rationale-only estimates have representation-specific support; the paired main table uses the common 278 question-model pairs across the four representations. These are eligible fitted comparisons, not all 6,400 evaluation pairs. The saved question-level means permit other explicitly labelled aggregations, including equal cell weighting; those would answer a different averaging question.

Table 13: Fisher–W macro-F1 at budget 20 with all available question–model pairs, weighted equally. Different representation supports motivate the paired main-text comparison.

representation	split	mean	count
General_full	duplicate_grouped	0.799	773
General_full	response_random	0.838	790
General_rationale	duplicate_grouped	0.665	278
General_rationale	response_random	0.703	290
PubMed_full	duplicate_grouped	0.793	773
PubMed_full	response_random	0.836	790
PubMed_rationale	duplicate_grouped	0.666	278
PubMed_rationale	response_random	0.696	290

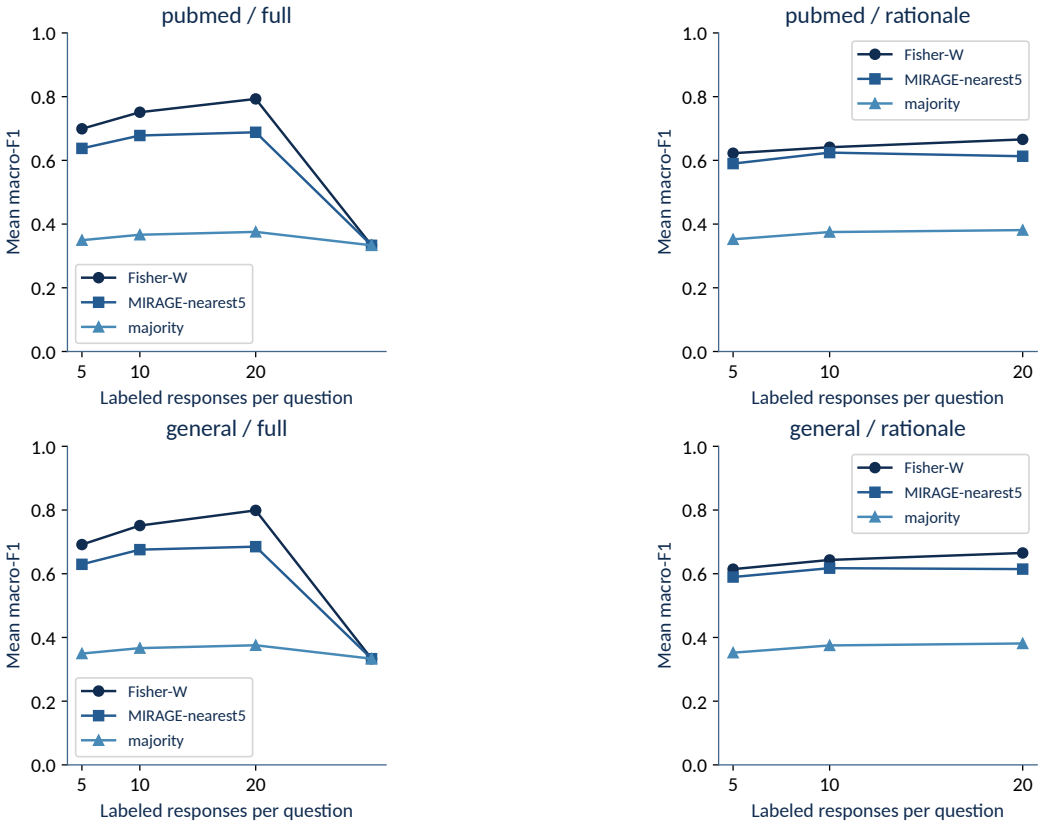


Figure 9: Duplicate-grouped label efficiency at budgets 5, 10 and 20. Points equally weight available question–model pairs; support changes with budget and representation. These are within-question held-out responses, not unseen-question transfer.

The prefix analysis uses the first 8, 16, or 32 observed responses in each eligible primary question set. A prefix is plotted only when it contains at least two examples per class, which is a descriptive sensitivity condition distinct from the primary five-per-class rule. Missing points therefore carry information about early class coverage. The plot does not extrapolate to unobserved sample counts.

It also does not claim that the first draws are an independent replication of the full response set: they are nested subsets.

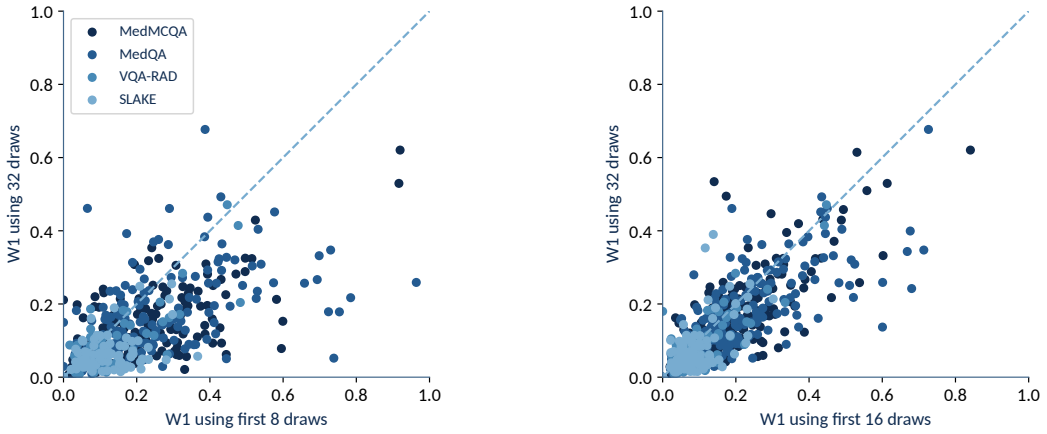


Figure 10: Prefix sensitivity at 8, 16 and 32 draws. Prefixes require two responses per class for this descriptive control. The main geometry eligibility rule remains five per class. No 64/100/150-response expanded result is implied.

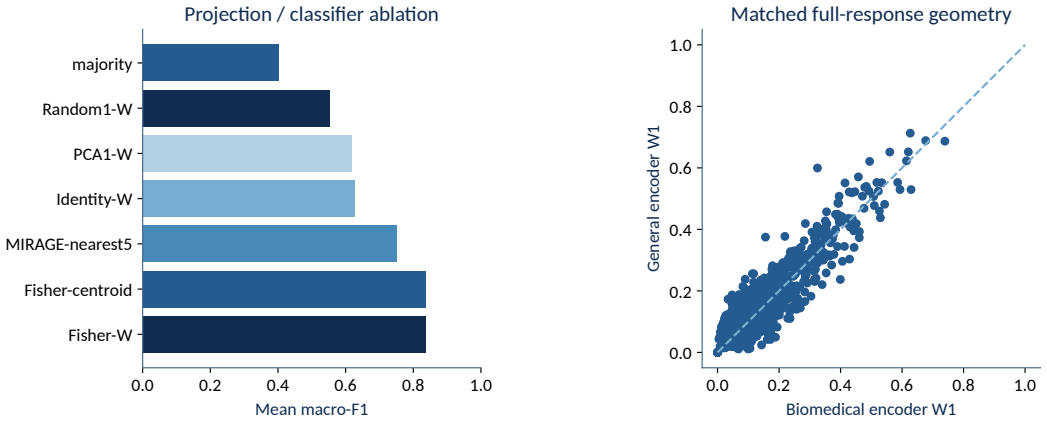


Figure 11: Projection ablations at budget 20 and matched encoder geometry. The classifier bars use available supports per method; the source data retain exact support counts. Fisher–W and Fisher-centroid have identical macro-F1 in the recorded trials; their AUCs can differ.

The projection ablation uses the implemented APORIA estimators on the same type of fixed response split. Fisher-W and Fisher-centroid have identical macro-F1 in all 13,136 matched recorded trials, while their AUCs differ in 373 trials because the continuous score orderings need not agree. This distinction avoids inferring identical decision scores from identical classification summaries. The main analysis therefore attributes the shared classification behavior to the fitted separation and recorded decision outcomes, without claiming equivalence of the two algorithms in general. Calibration-only regularization trials remain in the complete propagation table and are not selected using evaluation scores.

F.4 Encoder and distance sensitivity

We retain both biomedical and General sentence encoders and extend the corrected Eager analysis to three distances: Euclidean distance on unit embeddings, cosine dissimilarity, and cosine dissimilarity after regularized rank-32 PCA whitening. The whitener is fitted separately for each cell, encoder and text view using only its eight original calibration questions' parsed responses. We center with the

calibration mean, project onto the first 32 fitted components, and divide coordinate j by $\sqrt{\lambda_j + 0.1\bar{\lambda}}$, where $\bar{\lambda}$ averages the retained eigenvalues. Cosine dissimilarity then normalizes the projected vectors. The component limit is reduced if calibration support requires it; fitted components, scales, means and training response IDs are retained. Evaluation labels select the same mixed-class eligible questions as in the main analysis and never fit the transformation.

Table 14: Distance sensitivity on matched eligible evaluation pairs. Independent question/image support is reported explicitly; intervals bootstrap groups. White32 uses calibration-only regularized PCA whitening. W1 values from different metrics have different scales; no new multiple-testing claim is attached to this scan.

Encoder	View	Metric	Positive fraction	CI-low	CI-high	Groups	Pairs	Mean W1
General	full	cosine	0.678	0.645	0.711	610	808	0.119
General	full	euclidean	0.692	0.660	0.724	610	808	0.161
General	full	white32_cosine	0.663	0.631	0.696	610	808	0.142
General	masked	cosine	0.681	0.648	0.713	610	808	0.126
General	masked	euclidean	0.687	0.654	0.719	610	808	0.168
General	masked	white32_cosine	0.631	0.597	0.665	610	808	0.148
PubMed	full	cosine	0.667	0.634	0.701	610	808	0.104
PubMed	full	euclidean	0.675	0.641	0.707	610	808	0.154
PubMed	full	white32_cosine	0.630	0.595	0.665	610	808	0.146
PubMed	masked	cosine	0.666	0.633	0.699	610	808	0.106
PubMed	masked	euclidean	0.676	0.643	0.708	610	808	0.157
PubMed	masked	white32_cosine	0.644	0.610	0.676	610	808	0.152

Table 15: Paired Fisher macro-F1 contrasts on common pairs, with independent groups reported. Positive differences mean that the full answer performs better. These are encoder comparisons using the existing Fisher rule, not new metric-specific classifiers.

Encoder	Contrast	F1 Difference	CI-low	CI-high	Groups	Pairs
PubMed	full-masked	0.137	0.117	0.158	210	278
PubMed	full-rationale	0.141	0.120	0.163	210	278
General	full-masked	0.146	0.124	0.167	210	278
General	full-rationale	0.153	0.131	0.174	210	278

The distance sensitivity tables report dispersion direction for matched eligible pairs across encoders and metrics. The paired full-minus-masked Fisher macro F1 differences are 0.137 (95% group interval 0.117-0.158) for PubMed and 0.146 (0.124-0.167) for MiniLM. These contrasts concern the fixed Fisher rule; the metric scan separately examines distance distributions.

Dispersion direction and Fisher classification measure different properties. Answer identity can aid class separation without determining within class spread. Whitening is a sensitivity analysis, and its unadjusted descriptive summaries do not support a new discovery claim.

Proposition 1 (Monotone distances need not preserve dispersion direction)**.** *For unit vectors x, y, cosine dissimilarity is $d_C(x, y) = \frac{1}{2} d_E(x, y)^2$. Consequently, individual pair-distance rankings agree, but the sign of a difference between mean distances need not agree. Moreover, for distance laws supported on $[0, 2]$, their transformed laws satisfy*

$$W_1(g_{\#}P, g_{\#}Q) \leq 2W_1(P, Q), \qquad g(t) = t^2/2. \tag{3}$$

Proof. The identity follows by expanding $\|x - y\|_2^2$. A mean does not commute with nonlinear squaring: for example, a law concentrated at 1 has larger mean than a law placing equal mass at 0 and 1.8, but has smaller mean square. Finally $|g(s) - g(t)| \leq 2|s - t|$ on $[0, 2]$; apply this inequality to any coupling and minimize its expected transport cost. □

Thus Euclidean and cosine scans are related controls, not independent replications. The whitened scan changes the underlying geometry, while the original separation-versus-dispersion proposition remains applicable under all three metrics.

Table 16: **Fixed midpoint decoder-residual probes.** Pre-answer states predict the question-level majority-error propensity; post-answer states predict sampled response correctness. Each fit uses eight calibration questions and evaluates 400 disjoint questions (or their available response rows). AUCs target different outcomes and are not a head-to-head detector comparison. Layer indices are zero based.

Model	Data	Layer	Pre AUC	Post AUC	Length AUC
MedGemma 4B	MedMCQA	16	0.486	0.445	0.522
MedGemma 4B	MedQA	16	0.562	0.525	0.423
MedGemma 4B	VQA-RAD	16	0.790	0.759	0.648
MedGemma 4B	SLAKE	16	0.794	0.739	0.600
MedGemma 1.5 4B	MedMCQA	16	0.489	0.506	0.457
MedGemma 1.5 4B	MedQA	16	0.530	0.492	0.559
MedGemma 1.5 4B	VQA-RAD	16	0.776	0.686	0.664
MedGemma 1.5 4B	SLAKE	16	0.821	0.757	0.625
Lingshu 7B	MedMCQA	13	0.441	0.536	0.443
Lingshu 7B	MedQA	13	0.517	0.533	0.549
Lingshu 7B	VQA-RAD	13	0.830	0.679	0.639
Lingshu 7B	SLAKE	13	0.765	0.398	0.328
BioMistral 7B	MedMCQA	15	0.430	0.512	0.514
BioMistral 7B	MedQA	15	0.465	0.507	0.509
OpenBioLLM 8B	MedMCQA	15	0.418	0.491	0.542
OpenBioLLM 8B	MedQA	15	0.657	0.511	0.504

G Internal state prediction and output baselines

Probe fits use eight calibration questions and evaluate 400 disjoint questions. Prompt-only targets are question-level majority error propensity; post-answer targets are response-level correctness and can contain answer-prefix information. Twenty repeated calibration partitions, surface baselines, full-refit nulls, and fixed midpoint summaries are retained. Low direction cosine across fits is a stability observation under this small calibration design, not evidence that no multidimensional representation exists.

Cross-dataset transfer evaluates 400 target questions per directed route. The stronger visual-to-visual routes and weaker text/cross-format routes are reported together; answer type and task format remain plausible shared predictors. No transfer result is interpreted as a domain general medical correctness direction.

G.1 Answer type baseline

Table 17: Post hoc answer-type-only propensity baseline. One-hot answer type and class-balanced logistic regression use the same eight calibration questions and majority-error target. No evaluation labels are used for fitting. A high activation AUC does not by itself exceed this format baseline.

model	dataset	n_train	n_test	Midpoint pre AUC	Answer-type AUC
Lingshu-7B	SLAKE	8	400	0.765	0.808
Lingshu-7B	VQA-RAD	8	400	0.830	0.812
MedGemma-1.5-4B-it	SLAKE	8	400	0.821	0.751
MedGemma-1.5-4B-it	VQA-RAD	8	400	0.776	0.758
MedGemma-4B-it	SLAKE	8	400	0.794	0.689
MedGemma-4B-it	VQA-RAD	8	400	0.790	0.765

The answer-type-only control was added during manuscript review after inspecting the primary propensity results. It one-hot encodes the source answer type and fits class-balanced logistic regression with $C = 1$ using the same eight calibration questions. Its target is the same majority-error label used by prompt-only probes. All text questions have one answer type, so their baseline AUC is 0.5. The six visual cells provide the informative comparison in Table 17. This control does not use evaluation labels to fit the predictor, but the decision to add it was post hoc. A prospective replication should predefine this baseline and stratify or balance answer type.

G.2 LAYER PROFILES AND TRANSFER

Table 18 provides the fixed midpoint results with pointwise question-cluster intervals and null summaries. The fitted readout is held fixed during the evaluation-label permutation test. It is not refitted to randomized calibration labels. Thus the test assesses association of the fitted scores with evaluation outcomes under the specified permutation scheme, not the complete training procedure's null distribution. The BH family includes tested layers from both capture sites within each model-dataset cell; the length baseline has no corresponding permutation row.

Table 18: Midpoint AUCs with question/image-cluster bootstrap intervals and fixed-score permutation nulls. Site 0 permutes one predetermined question per independent image; site 1 shuffles response labels within questions, preserving question-specific correctness rates. Null diagnostics therefore have explicitly different support from full-cohort AUCs.

Model	Data	Site	Layer	AUC	CI-low	CI-high	Null mean	BH q
MedGemma 4B	MedMCQA	0	16	0.486	0.433	0.545	0.506	0.920
MedGemma 4B	MedMCQA	1	16	0.445	0.395	0.499	0.445	0.453
MedGemma 4B	MedQA	0	16	0.562	0.514	0.621	0.501	0.146
MedGemma 4B	MedQA	1	16	0.525	0.476	0.574	0.524	0.349
MedGemma 4B	VQA-RAD	0	16	0.790	0.737	0.834	0.500	0.005
MedGemma 4B	VQA-RAD	1	16	0.759	0.715	0.796	0.753	0.005
MedGemma 4B	SLAKE	0	16	0.794	0.756	0.832	0.492	0.005
MedGemma 4B	SLAKE	1	16	0.739	0.702	0.781	0.727	0.005
MedGemma 1.5 4B	MedMCQA	0	16	0.489	0.440	0.551	0.500	0.828
MedGemma 1.5 4B	MedMCQA	1	16	0.506	0.457	0.554	0.504	0.043
MedGemma 1.5 4B	MedQA	0	16	0.530	0.474	0.579	0.500	0.392
MedGemma 1.5 4B	MedQA	1	16	0.492	0.445	0.531	0.493	0.733
MedGemma 1.5 4B	VQA-RAD	0	16	0.776	0.711	0.826	0.505	0.010
MedGemma 1.5 4B	VQA-RAD	1	16	0.686	0.637	0.734	0.694	1.000
MedGemma 1.5 4B	SLAKE	0	16	0.821	0.784	0.859	0.493	0.005
MedGemma 1.5 4B	SLAKE	1	16	0.757	0.731	0.787	0.737	0.005
Lingshu 7B	MedMCQA	0	13	0.441	0.387	0.496	0.501	1.000
Lingshu 7B	MedMCQA	1	13	0.536	0.476	0.592	0.537	1.000
Lingshu 7B	MedQA	0	13	0.517	0.449	0.564	0.501	1.000
Lingshu 7B	MedQA	1	13	0.533	0.480	0.592	0.534	1.000
Lingshu 7B	VQA-RAD	0	13	0.830	0.782	0.873	0.507	0.009
Lingshu 7B	VQA-RAD	1	13	0.679	0.604	0.733	0.678	0.253
Lingshu 7B	SLAKE	0	13	0.765	0.717	0.814	0.499	0.005
Lingshu 7B	SLAKE	1	13	0.398	0.347	0.454	0.397	0.005
BioMistral 7B	MedMCQA	0	15	0.430	0.376	0.488	0.503	1.000
BioMistral 7B	MedMCQA	1	15	0.512	0.480	0.543	0.503	0.091
BioMistral 7B	MedQA	0	15	0.465	0.402	0.513	0.499	1.000
BioMistral 7B	MedQA	1	15	0.507	0.461	0.546	0.507	1.000
OpenBioLLM 8B	MedMCQA	0	15	0.418	0.357	0.474	0.505	1.000
OpenBioLLM 8B	MedMCQA	1	15	0.491	0.450	0.529	0.489	0.292
OpenBioLLM 8B	MedQA	0	15	0.657	0.603	0.718	0.499	0.011
OpenBioLLM 8B	MedQA	1	15	0.511	0.474	0.548	0.508	0.120

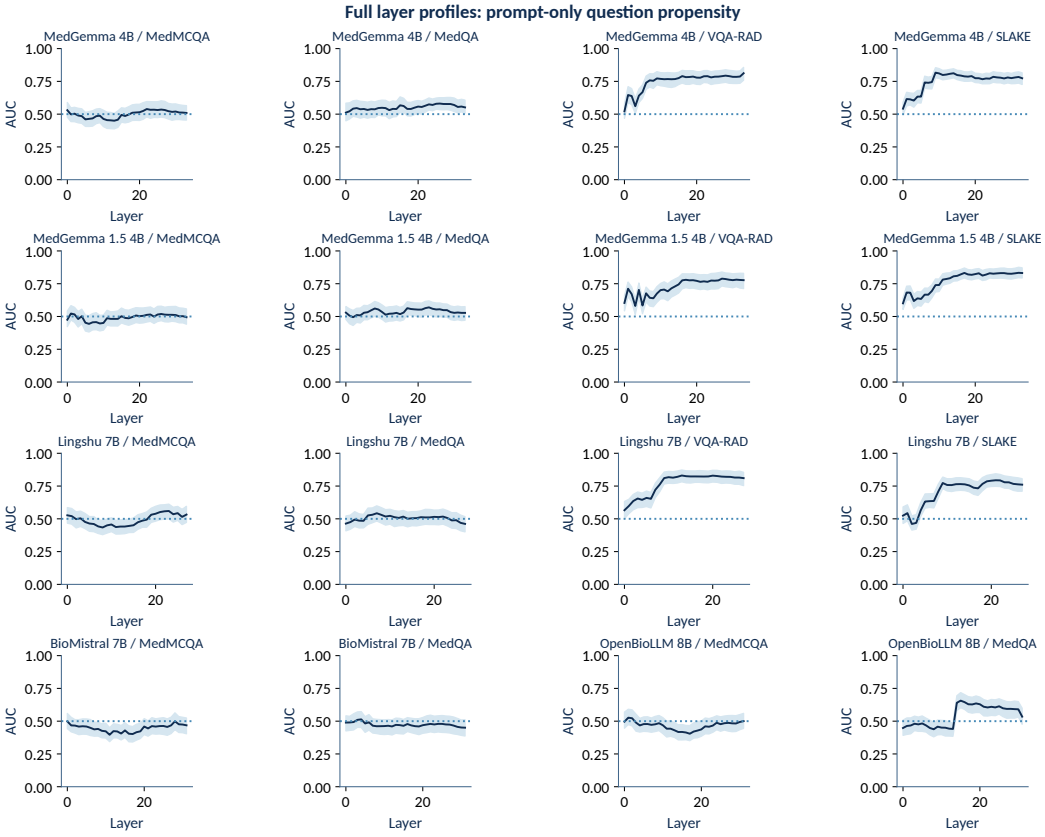


Figure 12: All prompt-only layer probes. Bands are pointwise 95% cluster bootstrap intervals, not simultaneous bands; layers are shown without evaluation-based selection.

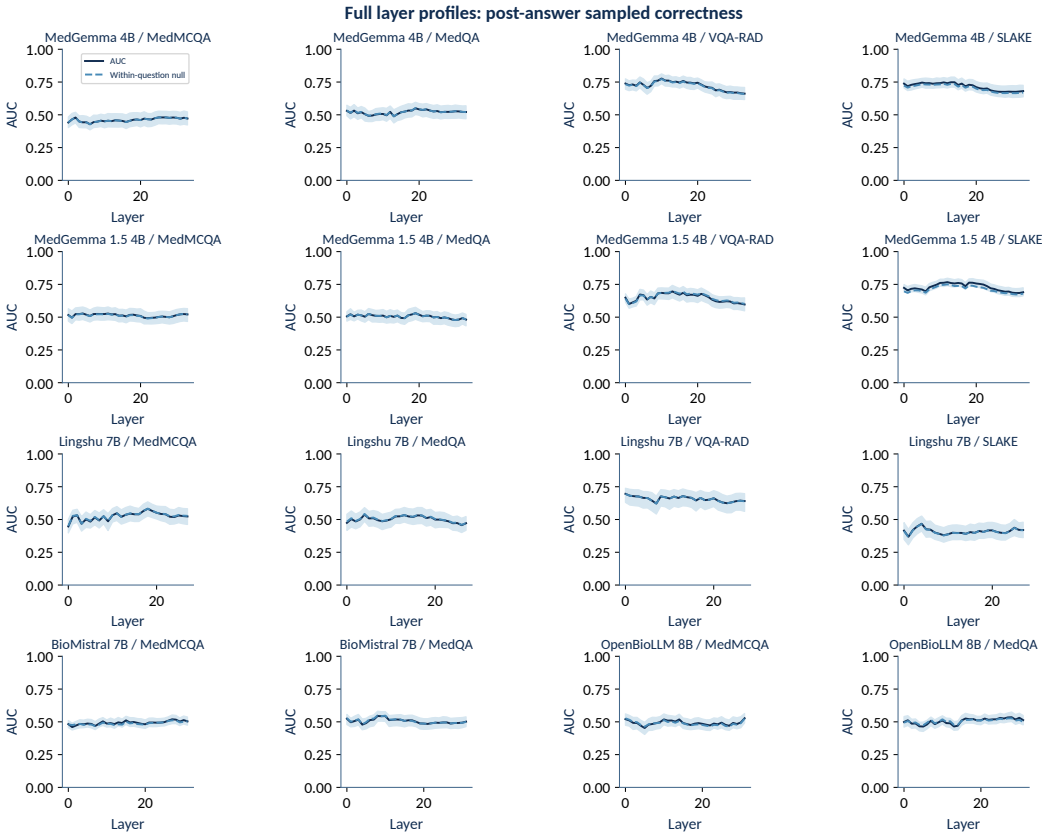


Figure 13: Post-answer AUC and the within-question permutation-null mean. Significant conditional association does not imply high absolute AUC; question-level label imbalance is preserved by the null.

All layer profiles are shown, including weak and reversed associations. The shaded intervals are pointwise bootstrap summaries from 200 resamples of evaluation question/image groups, not simultaneous confidence bands over the decoder stack. With only eight calibration questions, fitted directions can be sensitive to which questions occupy that partition. The primary profiles use one fixed partition. The supplementary fitting analysis in Section G.3 examines twenty predetermined partitions without selecting a favorable fit. The post-answer profile additionally depends on the short generated prefix, which can contain answer markers, option identities, or a completed short answer.

The post-answer permutation scheme preserves each evaluation question's sampled correctness rate. If a score largely ranks questions rather than separating responses within them, that question-level association persists under the null. Consequently, the plotted null curve can be above or below 0.5, and the observed curve should be compared with it before interpreting conditional significance. AUC itself remains a global ranking metric over the included evaluation responses. Reporting both quantities prevents a conditional permutation threshold from being misread as strong absolute detection.

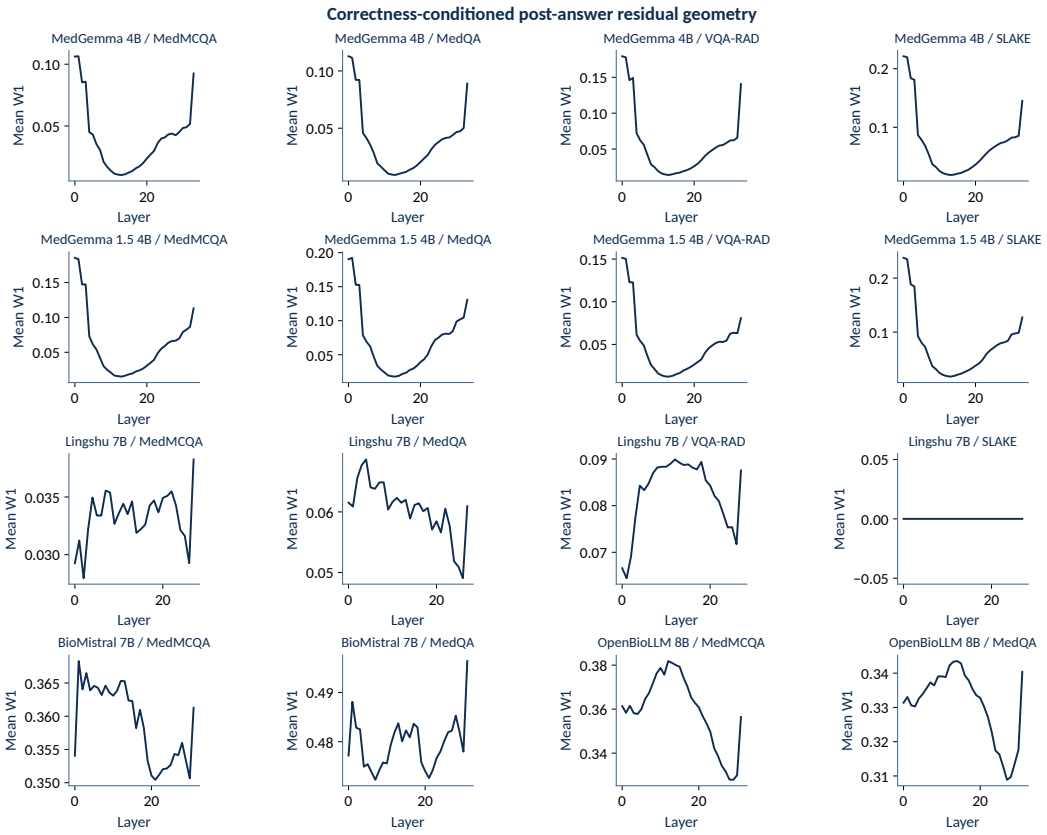


Figure 14: Mean within-question distance-law contrast in normalized post-answer residual states. Averages use eligible evaluation question–model pairs at each layer. This descriptive geometry is separate from probe generalization.

The activation-distance atlas computes within-question correctness-conditioned geometry on normalized post-answer residuals. Its points are internal representations from the generator rather than external sentence embeddings, but the same eligibility and dependence concerns apply. Mean distance-law contrast is descriptive; it does not identify a causal pathway. Fixed-depth permutation arrays are retained separately. Comparing this atlas with probe curves asks whether representational geometry and linear accessibility change together, while the intervention results test a different behavioral outcome at a prompt position.

Table 19: Cross-dataset post-answer probe AUCs at the fixed midpoint, summarized over compatible models. Calibration uses eight source questions; evaluation uses 400 target questions with image-cluster uncertainty. Full depth-specific results remain in the source CSV.

source	target	mean	min	max	count
MedMCQA	MedQA	0.489	0.465	0.514	5
MedMCQA	SLAKE	0.516	0.373	0.675	3
MedMCQA	VQA-RAD	0.516	0.391	0.651	3
MedQA	MedMCQA	0.523	0.495	0.552	5
MedQA	SLAKE	0.475	0.421	0.576	3
MedQA	VQA-RAD	0.496	0.469	0.548	3
SLAKE	MedMCQA	0.519	0.505	0.541	3
SLAKE	MedQA	0.549	0.527	0.583	3
SLAKE	VQA-RAD	0.613	0.359	0.758	3
VQA-RAD	MedMCQA	0.499	0.429	0.535	3
VQA-RAD	MedQA	0.511	0.490	0.532	3
VQA-RAD	SLAKE	0.679	0.645	0.712	3

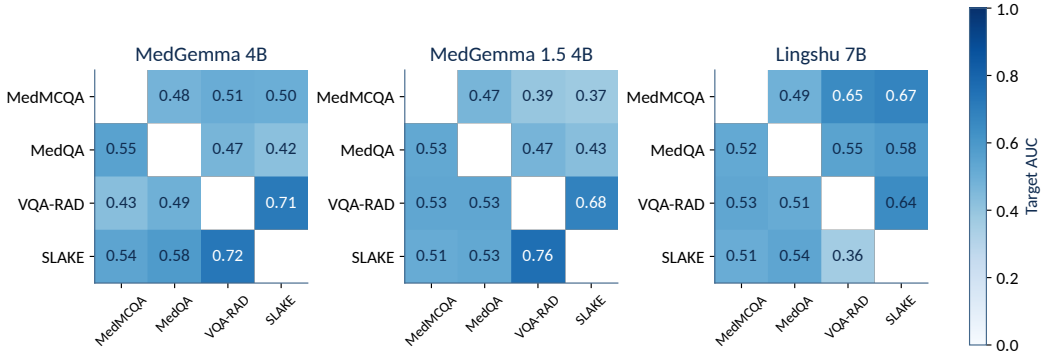


Figure 15: Midpoint post-answer probe transfer within each VLM. Rows are source calibration datasets and columns are target evaluation datasets. Diagonal is intentionally blank: within-dataset probes are a separate analysis.

Transfer uses the same model across source and target datasets. The source fit contains eight calibration questions and their parsed post-answer responses. The target evaluation contains 400 disjoint questions. The main midpoint summary is one of three fixed depths, with the full 120 comparisons available in the transfer CSV. Cross-model transfer would require an additional representation-alignment procedure and is not implied by these matrices. The image-to-image pattern is compatible with shared answer-format effects, so it does not independently establish a domain general medical correctness feature.

G.3 Fitting stability and surface controls

Repeated calibration fits use eight calibration questions and 20 predetermined question/image-group partitions, with 400 evaluation questions in each primary partition. Overlapping partitions are dependent sensitivity checks, not independent replications. Full-refit nulls preserve the relevant

Table 20: Calibration class support in the twenty unstratified group-disjoint splits. Degenerate fits are counted rather than replaced by favorable splits.

model	dataset	splits	pre_degenerate	post_degenerate	train_positive_min	train_positive_max
BioMistral-7B	MedMCQA	20	1	0	3	8
BioMistral-7B	MedQA	20	0	0	1	7
Lingshu-7B	MedMCQA	20	0	0	1	6
Lingshu-7B	MedQA	20	1	0	0	7
Lingshu-7B	SLAKE	20	0	0	13	32
Lingshu-7B	VQA-RAD	20	0	0	4	12
MedGemma-1.5-4B-it	MedMCQA	20	1	1	3	8
MedGemma-1.5-4B-it	MedQA	20	0	0	2	6
MedGemma-1.5-4B-it	SLAKE	20	0	0	15	26
MedGemma-1.5-4B-it	VQA-RAD	20	0	0	2	13
MedGemma-4B-it	MedMCQA	20	0	0	2	6
MedGemma-4B-it	MedQA	20	0	0	2	5
MedGemma-4B-it	SLAKE	20	0	0	9	20
MedGemma-4B-it	VQA-RAD	20	0	0	3	11
OpenBioLLM-8B	MedMCQA	20	5	1	3	8
OpenBioLLM-8B	MedQA	20	0	0	2	7

Table 21: Cosine similarity of fitted residual-probe directions across alternative splits, after mapping standardized coefficients back to unit-normalized residual coordinates. Pairwise split similarities are descriptive and dependent.

model	dataset	site	mean	min	max
BioMistral-7B	MedMCQA	post	0.017	-0.094	0.275
BioMistral-7B	MedMCQA	pre	0.037	-0.147	0.328
BioMistral-7B	MedQA	post	0.024	-0.107	0.265
BioMistral-7B	MedQA	pre	0.024	-0.247	0.326
Lingshu-7B	MedMCQA	post	0.016	-0.165	0.199
Lingshu-7B	MedMCQA	pre	0.021	-0.214	0.234
Lingshu-7B	MedQA	post	0.008	-0.183	0.245
Lingshu-7B	MedQA	pre	0.019	-0.254	0.252
Lingshu-7B	SLAKE	post	0.181	-0.028	0.554
Lingshu-7B	SLAKE	pre	0.219	-0.055	0.723
Lingshu-7B	VQA-RAD	post	0.073	-0.152	0.414
Lingshu-7B	VQA-RAD	pre	0.107	-0.173	0.618
MedGemma-1.5-4B-it	MedMCQA	post	0.034	-0.119	0.342
MedGemma-1.5-4B-it	MedMCQA	pre	0.064	-0.207	0.323
MedGemma-1.5-4B-it	MedQA	post	0.049	-0.290	0.693
MedGemma-1.5-4B-it	MedQA	pre	0.020	-0.235	0.354
MedGemma-1.5-4B-it	SLAKE	post	0.188	-0.015	0.522
MedGemma-1.5-4B-it	SLAKE	pre	0.125	-0.142	0.522
MedGemma-1.5-4B-it	VQA-RAD	post	0.090	-0.168	0.477
MedGemma-1.5-4B-it	VQA-RAD	pre	0.072	-0.161	0.427
MedGemma-4B-it	MedMCQA	post	0.020	-0.166	0.174
MedGemma-4B-it	MedMCQA	pre	0.015	-0.300	0.401
MedGemma-4B-it	MedQA	post	0.036	-0.187	0.635
MedGemma-4B-it	MedQA	pre	0.020	-0.273	0.355
MedGemma-4B-it	SLAKE	post	0.213	-0.026	0.718
MedGemma-4B-it	SLAKE	pre	0.192	-0.142	0.455
MedGemma-4B-it	VQA-RAD	post	0.076	-0.126	0.534
MedGemma-4B-it	VQA-RAD	pre	0.089	-0.241	0.590
OpenBioLLM-8B	MedMCQA	post	0.022	-0.217	0.294
OpenBioLLM-8B	MedMCQA	pre	0.029	-0.161	0.335
OpenBioLLM-8B	MedQA	post	0.027	-0.227	0.349
OpenBioLLM-8B	MedQA	pre	0.056	-0.216	0.396

Table 22: Within-answer-type propensity AUC stability. Unsupported one-class evaluation strata have undefined AUC and are excluded from valid split counts.

model	dataset	site	method	stratum	valid	mean	q025	q975
Lingshu-7B	SLAKE	pre	residual	closed	20	0.588	0.452	0.684
Lingshu-7B	SLAKE	pre	residual	open	20	0.698	0.465	0.774
Lingshu-7B	SLAKE	pre	type	closed	20	0.500	0.500	0.500
Lingshu-7B	SLAKE	pre	type	open	20	0.500	0.500	0.500
Lingshu-7B	SLAKE	pre	type_residual	closed	20	0.588	0.452	0.684
Lingshu-7B	SLAKE	pre	type_residual	open	20	0.698	0.465	0.774
Lingshu-7B	VQA-RAD	pre	residual	closed	20	0.502	0.380	0.583
Lingshu-7B	VQA-RAD	pre	residual	open	20	0.532	0.378	0.680
Lingshu-7B	VQA-RAD	pre	type	closed	20	0.500	0.500	0.500
Lingshu-7B	VQA-RAD	pre	type	open	20	0.500	0.500	0.500
Lingshu-7B	VQA-RAD	pre	type_residual	closed	20	0.502	0.380	0.583
Lingshu-7B	VQA-RAD	pre	type_residual	open	20	0.532	0.379	0.680
MedGemma-1.5-4B-it	SLAKE	pre	residual	closed	20	0.540	0.436	0.637
MedGemma-1.5-4B-it	SLAKE	pre	residual	open	20	0.658	0.488	0.721
MedGemma-1.5-4B-it	SLAKE	pre	type	closed	20	0.500	0.500	0.500
MedGemma-1.5-4B-it	SLAKE	pre	type	open	20	0.500	0.500	0.500
MedGemma-1.5-4B-it	SLAKE	pre	type_residual	closed	20	0.540	0.436	0.637
MedGemma-1.5-4B-it	SLAKE	pre	type_residual	open	20	0.658	0.488	0.721
MedGemma-1.5-4B-it	VQA-RAD	pre	residual	closed	20	0.508	0.429	0.576
MedGemma-1.5-4B-it	VQA-RAD	pre	residual	open	20	0.620	0.411	0.768
MedGemma-1.5-4B-it	VQA-RAD	pre	type	closed	20	0.500	0.500	0.500
MedGemma-1.5-4B-it	VQA-RAD	pre	type	open	20	0.500	0.500	0.500
MedGemma-1.5-4B-it	VQA-RAD	pre	type_residual	closed	20	0.508	0.429	0.576
MedGemma-1.5-4B-it	VQA-RAD	pre	type_residual	open	20	0.620	0.411	0.768
MedGemma-4B-it	SLAKE	pre	residual	closed	20	0.578	0.468	0.706
MedGemma-4B-it	SLAKE	pre	residual	open	20	0.731	0.560	0.808
MedGemma-4B-it	SLAKE	pre	type	closed	20	0.500	0.500	0.500
MedGemma-4B-it	SLAKE	pre	type	open	20	0.500	0.500	0.500
MedGemma-4B-it	SLAKE	pre	type_residual	closed	20	0.578	0.468	0.706
MedGemma-4B-it	SLAKE	pre	type_residual	open	20	0.731	0.560	0.808
MedGemma-4B-it	VQA-RAD	pre	residual	closed	20	0.507	0.442	0.552
MedGemma-4B-it	VQA-RAD	pre	residual	open	20	0.631	0.348	0.787
MedGemma-4B-it	VQA-RAD	pre	type	closed	20	0.500	0.500	0.500
MedGemma-4B-it	VQA-RAD	pre	type	open	20	0.500	0.500	0.500
MedGemma-4B-it	VQA-RAD	pre	type_residual	closed	20	0.507	0.442	0.552
MedGemma-4B-it	VQA-RAD	pre	type_residual	open	20	0.632	0.348	0.787

Table 23: Original-split permutation tests with full refitting after shuffling calibration labels and independently shuffling evaluation labels. Pre-answer shuffles within partitions; post-answer shuffles within each question. BH correction here spans the fixed-midpoint site/cell family. Resolution is 0.01 for 99 permutations.

site	layer	auc	full_cohort_auc	null_mean	p	n	n_train	n_test	scheme	model	dataset	q_bh
pre	16	0.487	0.487	0.499	0.720	99	8	400	one_fixed_question_per_image	MedGemma-4B-it	MedMCQA	0.880
post	16	0.442	0.442	0.442	0.070	99	256	12800	within_question	MedGemma-4B-it	MedMCQA	0.140
pre	16	0.559	0.559	0.498	0.030	99	8	400	one_fixed_question_per_image	MedGemma-4B-it	MedQA	0.069
post	16	0.528	0.528	0.533	0.760	99	256	12800	within_question	MedGemma-4B-it	MedQA	0.880
pre	13	0.810	0.830	0.503	0.010	99	8	195	one_fixed_question_per_image	Lingshu-7B	VQA-RAD	0.025
post	13	0.688	0.688	0.689	0.660	99	256	12800	within_question	Lingshu-7B	VQA-RAD	0.858
pre	13	0.831	0.765	0.494	0.010	99	8	88	one_fixed_question_per_image	Lingshu-7B	SLAKE	0.025
post	13	0.395	0.395	0.395	0.010	99	256	12800	within_question	Lingshu-7B	SLAKE	0.025
pre	15	0.421	0.421	0.503	1.000	99	8	400	one_fixed_question_per_image	BioMistral-7B	MedMCQA	1.000
post	15	0.520	0.520	0.492	0.010	99	256	12595	within_question	BioMistral-7B	MedMCQA	0.025
pre	15	0.471	0.471	0.495	0.770	99	8	400	one_fixed_question_per_image	BioMistral-7B	MedQA	0.880
post	15	0.504	0.504	0.481	0.040	99	255	12756	within_question	BioMistral-7B	MedQA	0.085
pre	15	0.414	0.414	0.501	0.990	99	8	400	one_fixed_question_per_image	OpenBioLLM-8B	MedMCQA	1.000
post	15	0.504	0.504	0.499	0.100	99	254	12058	within_question	OpenBioLLM-8B	MedMCQA	0.188
pre	15	0.663	0.663	0.499	0.010	99	8	400	one_fixed_question_per_image	OpenBioLLM-8B	MedQA	0.025
post	15	0.528	0.528	0.528	0.520	99	250	12374	within_question	OpenBioLLM-8B	MedQA	0.723
pre	16	0.776	0.790	0.503	0.010	99	8	195	one_fixed_question_per_image	MedGemma-4B-it	VQA-RAD	0.025
post	16	0.762	0.762	0.740	0.160	99	256	12800	within_question	MedGemma-4B-it	VQA-RAD	0.256
pre	16	0.753	0.796	0.506	0.010	99	8	88	one_fixed_question_per_image	MedGemma-4B-it	SLAKE	0.025
post	16	0.743	0.743	0.732	0.010	99	256	12800	within_question	MedGemma-4B-it	SLAKE	0.025
pre	16	0.487	0.487	0.496	0.670	99	8	400	one_fixed_question_per_image	MedGemma-1.5-4B-it	MedMCQA	0.858
post	16	0.506	0.506	0.503	0.010	99	241	12062	within_question	MedGemma-1.5-4B-it	MedMCQA	0.025
pre	16	0.527	0.527	0.496	0.150	99	8	400	one_fixed_question_per_image	MedGemma-1.5-4B-it	MedQA	0.253
post	16	0.491	0.491	0.488	0.150	99	217	11586	within_question	MedGemma-1.5-4B-it	MedQA	0.253
pre	16	0.768	0.782	0.499	0.010	99	8	195	one_fixed_question_per_image	MedGemma-1.5-4B-it	VQA-RAD	0.025
post	16	0.700	0.700	0.687	0.340	99	256	12800	within_question	MedGemma-1.5-4B-it	VQA-RAD	0.495
pre	16	0.715	0.822	0.499	0.010	99	8	88	one_fixed_question_per_image	MedGemma-1.5-4B-it	SLAKE	0.025
post	16	0.754	0.754	0.734	0.010	99	256	12800	within_question	MedGemma-1.5-4B-it	SLAKE	0.025
pre	13	0.441	0.441	0.496	0.990	99	8	400	one_fixed_question_per_image	Lingshu-7B	MedMCQA	1.000
post	13	0.536	0.536	0.537	0.960	99	256	12800	within_question	Lingshu-7B	MedMCQA	1.000
pre	13	0.517	0.517	0.497	0.240	99	8	400	one_fixed_question_per_image	Lingshu-7B	MedQA	0.366
post	13	0.530	0.530	0.522	0.010	99	256	12800	within_question	Lingshu-7B	MedQA	0.025

question-level label composition, so a conditional null AUC need not equal 0.5. Surface baselines use emitted-token features and response length; they do not establish that residual information is clinically meaningful.

G.4 Uncertainty from repeated outputs

We add inexpensive output-level comparators using exactly the saved original split and twenty repeated 8/400 question/image-disjoint partitions. Each question contributes four features: entropy of its parsed-answer frequency distribution, one minus the modal parsed-answer frequency, the fraction of unparsed draws, and mean pairwise cosine dissimilarity of corrected PubMed full-response embeddings. For parsed-answer proportions $\hat{p}_q(a)$, the first feature is

$$\widehat{H}_{\text{answer}}(q) = -\sum_a \hat{p}_q(a) \log \hat{p}_q(a). \tag{4}$$

All thirty-two primary draws contribute to the available response set. Features use neither gold answers nor correctness labels. Their low-dimensional logistic predictor uses the same calibration-only standardization, balanced loss, regularization and solver as the repeated residual probes. Type-plus-output concatenates the metadata and four output features. We independently reproduce the type-only AUC on every supported split before comparing methods. All methods omit the same single degenerate training partition; no split is redrawn to improve the result.

The comparison is informative but uses different information budgets. Prompt-only states and answer type are available before generation; output uncertainty requires repeated completions and predicts the majority-error label defined from that response set. It is therefore a post-generation question-level audit, not a substitute for a prompt-time predictor or a prospective response-error detector. In particular, answer-frequency entropy is not predictive token entropy and does not merge paraphrases through entailment. The original records store sampled token IDs, not full vocabulary distributions; predictive token entropy cannot be recovered from those IDs alone.

The generated tables report the matched partition AUCs and paired gains. These comparisons assess whether repeated outputs add information beyond answer type under the stated sampling budget. Split percentiles describe fitting sensitivity and are not confidence intervals from independent cohorts.

Relation to medical semantic entropy and decoding. VASE uses weak visual transformations and a distorted-image contrast to make semantic uncertainty more sensitive to image evidence (Liao

Table 24: Mean question-propensity AUC over twenty matched group-disjoint partitions with eight training groups. Type, residual and type-plus-residual are prompt-available comparators; output and type-plus-output use the generated response set. Degenerate cell/splits are omitted consistently across methods. No detector-superiority claim across information budgets is implied.

Model	Dataset	Utput	Esidual	Ype	Ype_output	Ype_residual
BioMistral-7B	MedMCQA	0.665	0.585	0.500	0.665	0.585
BioMistral-7B	MedQA	0.616	0.517	0.500	0.616	0.517
Lingshu-7B	MedMCQA	0.496	0.518	0.500	0.496	0.518
Lingshu-7B	MedQA	0.501	0.512	0.500	0.501	0.512
Lingshu-7B	SLAKE	0.703	0.842	0.811	0.873	0.842
Lingshu-7B	VQA-RAD	0.576	0.731	0.811	0.800	0.732
MedGemma-1.5-4B-it	MedMCQA	0.544	0.565	0.500	0.544	0.565
MedGemma-1.5-4B-it	MedQA	0.587	0.505	0.500	0.587	0.505
MedGemma-1.5-4B-it	SLAKE	0.857	0.774	0.753	0.872	0.774
MedGemma-1.5-4B-it	VQA-RAD	0.842	0.719	0.762	0.845	0.719
MedGemma-4B-it	MedMCQA	0.521	0.521	0.500	0.521	0.521
MedGemma-4B-it	MedQA	0.545	0.518	0.500	0.545	0.518
MedGemma-4B-it	SLAKE	0.779	0.774	0.692	0.826	0.774
MedGemma-4B-it	VQA-RAD	0.767	0.740	0.770	0.812	0.740
OpenBioLLM-8B	MedMCQA	0.560	0.566	0.500	0.560	0.566
OpenBioLLM-8B	MedQA	0.616	0.594	0.500	0.616	0.594

Table 25: Paired AUC gain of type-plus-output over type-only, with 2.5th and 97.5th percentiles across overlapping fitting partitions. These ranges measure split sensitivity and are not confidence intervals for a population mean.

Model	Dataset	Ean	Ow	Igh
BioMistral-7B	MedMCQA	0.165	-0.179	0.271
BioMistral-7B	MedQA	0.116	-0.206	0.212
Lingshu-7B	MedMCQA	-0.004	-0.037	0.018
Lingshu-7B	MedQA	0.001	-0.065	0.064
Lingshu-7B	SLAKE	0.063	0.046	0.072
Lingshu-7B	VQA-RAD	-0.012	-0.152	0.036
MedGemma-1.5-4B-it	MedMCQA	0.044	-0.084	0.108
MedGemma-1.5-4B-it	MedQA	0.087	-0.023	0.146
MedGemma-1.5-4B-it	SLAKE	0.119	0.067	0.131
MedGemma-1.5-4B-it	VQA-RAD	0.083	0.027	0.106
MedGemma-4B-it	MedMCQA	0.021	-0.110	0.128
MedGemma-4B-it	MedQA	0.045	-0.148	0.139
MedGemma-4B-it	SLAKE	0.134	0.097	0.156
MedGemma-4B-it	VQA-RAD	0.042	-0.058	0.078
OpenBioLLM-8B	MedMCQA	0.060	-0.126	0.155
OpenBioLLM-8B	MedQA	0.116	-0.053	0.189

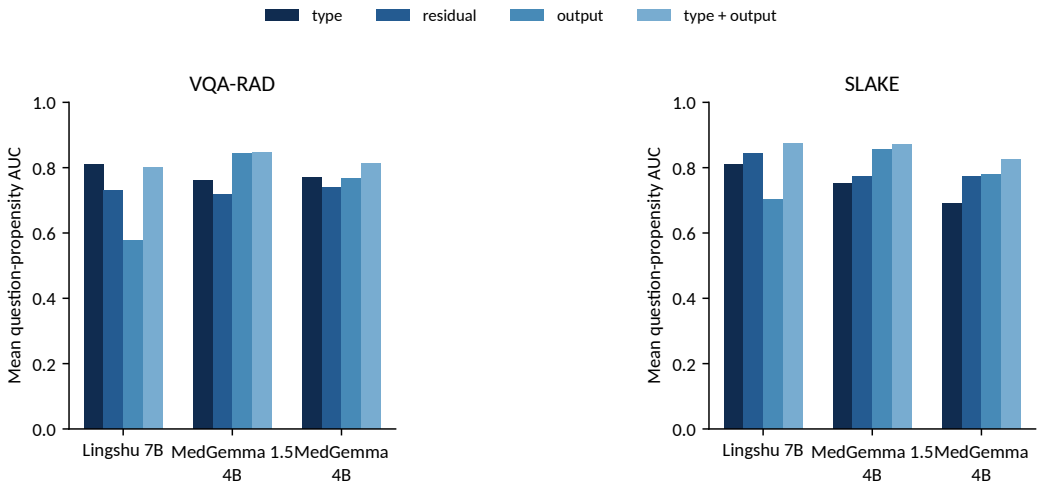


Figure 16: Matched-split visual propensity comparisons. Output uncertainty uses answer entropy, variation ratio, parser failures and response-embedding dispersion after generation; it is not available at the prompt-only capture site.

et al., 2025). Its visual sampling and semantic predictive distribution differ from our fixed-image answer-frequency features. The cited version is explicitly the original VASE paper; later versions of the same arXiv record have a different title. A matched VASE comparison would require its image transformations, semantic grouping and generation budget on the same frozen evaluation questions. Our results do not claim to outperform it. Residual Decoding uses historical decoding information to modify generation (Chen et al., 2026); it is a complementary mitigation method rather than a static probe, and its general vision-language evaluation should not be presented as an already matched medical baseline. Neither method is silently relabeled as implemented here.

G.5 Calibration support and abstention

Recent correctness-subspace evidence is more conditional than a universal-direction interpretation suggests. The revised Confidence Manifold study reports compact correctness representations, regime-dependent advantages over output uncertainty, and weak single-dataset transfer improved by joint training (Cho et al., 2026). This supports examining domain, task, fitting support and training mixture when comparing results. MIRAGE's medical majority-error target before generation, response-error target after a short prefix, eight calibration questions and open/closed visual formats differ from a factual-statement correctness task. Our findings therefore qualify a broad extrapolation; they are not a direct failed replication of that study.

Proposition 2 (Finite calibration data do not identify a unique representation direction)**.** *Let $H \in \mathbb{R}^{n\times d}$ be centered calibration representations with $d > n-1$. If $v \in \ker H$, then the linear scores of w and $w+v$ agree on the calibration observations. The null space has dimension at least $d-n+1$.*

Proof. Centering gives $\operatorname{rank}(H) \le n-1$. Rank-nullity gives the dimension bound, and $H(w+v) = Hw$ follows from $Hv = 0$. □

For the prompt-only design, eight calibration questions give a centered rank of at most seven. Post-answer fitting has more response rows but only eight independent question groups; its algebraic rank bound uses the response-row count, not eight. Regularization can choose a unique fitted coefficient, but the observed data still constrain only a small part of the ambient space. Different samples, correlations and feature scaling can rotate the fitted vector. Low cosine agreement between fitted vectors consequently does not disprove a stable multidimensional correctness subspace. A dedicated subspace test would fit ranks using training data, compare held-out projection performance and principal angles, and repeat across independent cohorts. The present repeated-split test establishes estimator sensitivity under the specified small calibration design.

Geometry-informed abstention is a separate decision problem. Geometry-Calibrated Conformal Abstention explicitly connects representation information to an abstention rule (Xu et al., 2026); conformal risk control calibrates suitable loss criteria under exchangeability assumptions (Angelopoulos et al., 2022). Neither a high AUC nor a BH-adjusted geometry comparison is itself a calibrated abstention guarantee. For a fixed uncertainty score S and threshold τ, participation and selective error are

$$C(\tau) = \Pr(S \le \tau), \qquad R(\tau) = \frac{\mathbb{E}[\mathbf{1}\{Y=1\}\mathbf{1}\{S \le \tau\}]}{C(\tau)}, \quad C(\tau) > 0. \tag{5}$$

A participation bound does not determine the numerator. Standard split-conformal prediction-set coverage also differs from a bound on the error fraction among accepted single answers. Calibration must target the claimed quantity and must use data held out from fitting or an appropriate cross-fitting construction.

The existing eight fitting questions cannot additionally serve as an independent certification set. Even with eight fresh exchangeable scores, the usual split-conformal order statistic at nominal 90% coverage has index $\lceil 9 \times 0.9 \rceil = 9$; its conservative finite-sample construction is trivial when only eight finite calibration scores are available. For a different illustrative calculation, zero errors on eight independent accepted cases gives a one-sided exact 95% binomial upper error bound $1 - 0.05^{1/8} \approx 0.312$ for a fixed, previously chosen rule. Neither calculation is a reported MIRAGE deployment guarantee. A clinically useful abstention study needs substantially more independent calibration cases and a prespecified risk target.

H Supplementary evidence and residual interventions

H.1 Intervention configurations

Table 26: Restoration configurations and the questions they answer. Teacher-forced candidate-context donors in the original experiments and prompt-only donors in the new generation follow-up are distinct policies.

Configuration	Intervention and comparator	Interpretation
Broad text restoration	Zero non-special prompt embeddings; replace final-prompt residual with its clean donor	Tests recovery of reference-answer score after broad prompt disruption; does not isolate one medical fact.
Broad visual restoration	Blank image; clean final-prompt donor at fixed depth	Tests sensitivity to image-conditioned prompt information; original reference is a recovery target after corruption.
Sham	Corrupted donor into the same corrupted context	Numerical and implementation control; an exact sham should have zero score change under matched contexts.
Matched random	Random displacement with the donor displacement's norm	Controls perturbation magnitude, without guaranteeing equal semantic content or naturalness.
Medical minimal pairs	Swap donor from a changed laterality, presence/absence or arithmetic statement	Tests counterfactual answer preference for a controlled linguistic contrast.
Edited-token site	Whole residual at the changed evidence token	Tests site dependence; adaptive follow-up, not identification of a unique medical circuit.
Serialization control	Same question with raw versus Llama-3 user headers	Tests dependence on prompt formatting; parser incompatibility is recorded separately from clinical error.
Sampled follow-up	Midpoint prompt-only donor; patch generation prefill, then use cached decoding	Relates one fixed intervention policy to sampled reference matching and downstream support diagnostics.

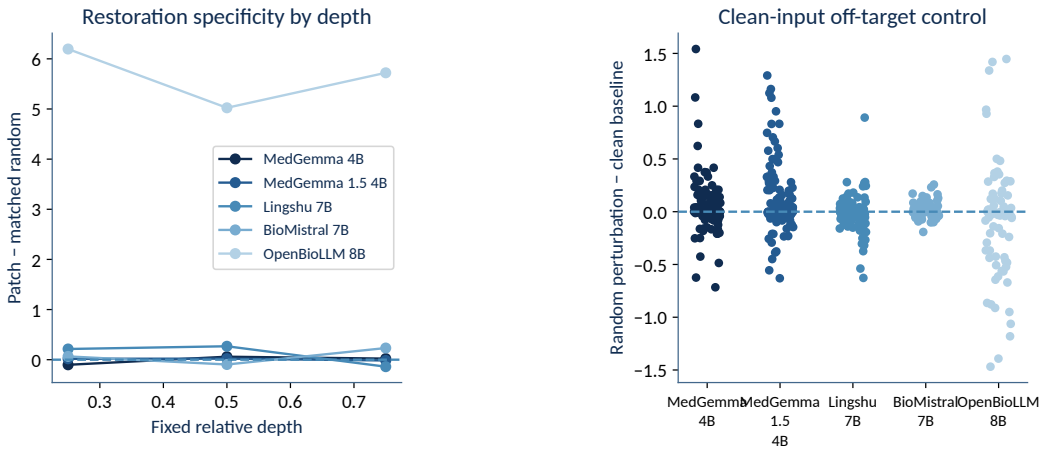


Figure 17: Fixed-depth restoration contrast and clean-input random control. Off-target effects are measured against clean input, avoiding the incompatible corrupted-input baseline. Depth means pool the compatible datasets for each checkpoint.

H.2 Paired image controls

Image dependence is tested on a supplementary evaluation subset with four paired seeds and image group bootstrap intervals. These intervals do not describe the full 400 question cohort. The image control table records the paired counts and answer changes; these measure evidence dependence rather than clinical correctness after an evidence change.

Medical lexical strata are transparent automated features for negation, laterality, quantity/unit, treatment, diagnosis, anatomy, specialty, and modality. They overlap and are not clinician severity labels. Image controls are a smaller, predeclared 40-question evaluation subset per visual model-dataset cell with four paired seeds; their image-group bootstrap intervals therefore do not describe all 400 questions. Across the 1,056 paired question-seed observations per condition, answer changes are 744/1,056 for blank, 699/1,056 for swapped, and 841/1,056 for text-only inputs. These demonstrate evidence dependence, not clinical correctness of the changed response.

Table 27: Image-control rates averaged first over paired seeds within question/model, then across questions and checkpoints. Intervals resample image groups shared across checkpoints; the clusters column gives independent image support.

dataset	condition	mean	lo	hi	clusters	rows
SLAKE	blank	0.808	0.744	0.875	40	120
SLAKE	swapped	0.742	0.635	0.842	40	120
SLAKE	text_only	0.875	0.815	0.929	40	120
VQA-RAD	blank	0.608	0.490	0.721	40	120
VQA-RAD	swapped	0.577	0.446	0.706	40	120
VQA-RAD	text_only	0.719	0.615	0.815	40	120
pooled	blank	0.708	0.640	0.775	80	240
pooled	swapped	0.659	0.575	0.744	80	240
pooled	text_only	0.797	0.738	0.853	80	240

H.3 Broad residual restoration

Each cell contributes the fixed 40-question evaluation control subset, three depths, and four intervention conditions. Table 28 reports mean score changes for restoration, matched random, and sham conditions. These are short control-study measurements and are not pooled with the 400-question primary cohort. Figure 17 separates depth-dependent specificity from random perturbations applied to clean input. The clean-input control is scored relative to the clean baseline, avoiding an artificial apparent effect caused by subtracting the corrupted baseline from a clean-input intervention.

Table 28: Mean original-gold log-probability change from corrupted input across the frozen evaluation controls and three depths per cell. Off-target random uses clean input and is analyzed with delta_from_clean in the accompanying data.

model	dataset	clean_patch	matched_random	sham
BioMistral-7B	MedMCQA	0.179	0.031	0.000
BioMistral-7B	MedQA	-0.251	-0.238	0.000
Lingshu-7B	MedMCQA	0.039	-0.193	0.000
Lingshu-7B	MedQA	0.062	0.001	0.000
Lingshu-7B	SLAKE	0.062	0.009	0.000
Lingshu-7B	VQA-RAD	-0.014	-0.124	0.000
MedGemma-1.5-4B-it	MedMCQA	0.005	-0.021	0.000
MedGemma-1.5-4B-it	MedQA	0.001	-0.021	0.000
MedGemma-1.5-4B-it	SLAKE	-0.004	0.002	0.000
MedGemma-1.5-4B-it	VQA-RAD	-0.001	0.009	0.000
MedGemma-4B-it	MedMCQA	-0.005	0.048	0.000
MedGemma-4B-it	MedQA	0.048	0.026	0.000
MedGemma-4B-it	SLAKE	0.021	-0.011	0.000
MedGemma-4B-it	VQA-RAD	0.051	0.079	0.000
OpenBioLLM-8B	MedMCQA	5.114	-0.414	0.000
OpenBioLLM-8B	MedQA	5.664	-0.101	0.000

The random vector is sampled from a Gaussian direction and rescaled to the norm of the clean-minus-corrupted residual difference at that question and layer. Norm matching controls perturbation magnitude in residual coordinates; it does not make a random direction equivalent to a semantically meaningful intervention. A positive restoration-minus-random contrast is therefore evidence of specificity relative to that control, not proof of a unique mechanism. The selected final-prompt position is held fixed across conditions, and the intervention modifies a decoder-block output rather than isolating an attention head or MLP subspace.

The outcome averages log probabilities of the reference-answer tokens under teacher forcing. Formatting-prefix tokens are excluded using token offsets, while the reference answer is still conditioned on any preceding answer tokens. For MCQs, this can be a very short token sequence, making large log-probability changes possible after severe prompt corruption. For visual questions, the original reference is a counterfactual scoring target after altered evidence. The outcome is useful for a controlled comparison but is not the same as sampled answer accuracy, explanation quality, or medical harm.

H.4 Medical contrasts and intervention sites

Broad residual restoration uses the fixed 40-question control subset per compatible cell, three fixed depths, and sham and norm-matched random controls. It is separate from the 400-question primary cohort. The strongest main evaluation contrasts are OpenBioLLM 5.529 nats/token on MedMCQA and 5.765 on MedQA; other checkpoints are smaller or mixed. Medical minimal pairs use 18 synthetic clinical-language prompts with three wordings per feature, both answer assignments and both directions. They test a specified counterfactual margin, not diagnosis or treatment validity.

The accepted intervention records have maximum absolute sham shift zero. Two generation shards nevertheless produced nonzero sham checks before prospective replay. The replay protocol passed all repeated controls, but original/replay intervention effects differed by up to 0.310 nats. We retain both versions and scope causal statements to the tested execution path, rather than treating the replay as proof that the original discrepancy had a known hardware cause.

The exact question IDs, image groups, seeds, prompt text, token spans, states, and source hashes for these controls are retained in their artifact directories. Their smaller denominators are part of the interpretation and are repeated in figure captions and tables.

Table 29: Broad-corruption restoration-minus-random contrasts, averaging fixed depths within question and bootstrapping the control questions. Control selection uses distinct images; clusters reports independent support per cell.

model	dataset	mean	lo	hi	clusters	rows
BioMistral-7B	MedMCQA	0.148	0.113	0.181	40	40
BioMistral-7B	MedQA	-0.013	-0.089	0.065	40	40
Lingshu-7B	MedMCQA	0.232	0.119	0.338	40	40
Lingshu-7B	MedQA	0.061	0.033	0.089	40	40
Lingshu-7B	SLAKE	0.053	0.001	0.128	40	40
Lingshu-7B	VQA-RAD	0.110	-0.002	0.232	40	40
MedGemma-1.5-4B-it	MedMCQA	0.026	0.011	0.040	40	40
MedGemma-1.5-4B-it	MedQA	0.022	0.013	0.032	40	40
MedGemma-1.5-4B-it	SLAKE	-0.006	-0.021	0.010	40	40
MedGemma-1.5-4B-it	VQA-RAD	-0.010	-0.034	0.012	40	40
MedGemma-4B-it	MedMCQA	-0.053	-0.073	-0.034	40	40
MedGemma-4B-it	MedQA	0.022	0.001	0.042	40	40
MedGemma-4B-it	SLAKE	0.032	0.003	0.066	40	40
MedGemma-4B-it	VQA-RAD	-0.029	-0.064	0.007	40	40
OpenBioLLM-8B	MedMCQA	5.529	4.984	6.080	40	40
OpenBioLLM-8B	MedQA	5.765	5.314	6.249	40	40

Table 30: Medical-language minimal-pair controls. Both-correct is the fraction of paired option-order/direction tests whose source and counterfactual margins favor their respective rule-derived labels. Effects are changes in source-versus-counterfactual answer margin, averaged over direction, option order and fixed depths. Only three independent wordings support each feature.

model	feature	templates	both_correct	source_patch	matched_random	nuisance_patch	gain_random	gain_nuisance
BioMistral-7B	dosage	3	0.000	-0.009	-0.000	-0.010	-0.008	0.002
BioMistral-7B	laterality	3	1.000	0.017	-0.006	0.002	0.023	0.016
BioMistral-7B	negation	3	0.000	0.005	-0.017	-0.014	0.022	0.019
MedGemma-4B-it	dosage	3	0.333	-0.014	0.006	0.014	-0.020	-0.028
MedGemma-4B-it	laterality	3	1.000	0.042	0.014	-0.014	0.028	0.056
MedGemma-4B-it	negation	3	1.000	0.010	0.010	0.017	0.000	-0.007
OpenBioLLM-8B	dosage	3	0.000	0.021	0.010	0.010	0.011	0.010
OpenBioLLM-8B	laterality	3	1.000	0.007	0.002	0.014	0.005	-0.007
OpenBioLLM-8B	negation	3	0.333	0.007	0.015	0.023	-0.008	-0.016

Table 31: Adaptive follow-up at the last changed evidence token after weak final-prompt effects. Same pairs, directions, option orders and three fixed depths; sham remains exactly zero. Recovery is preference for the source answer after a donor patch, not correctness for the unedited target input.

model	feature	source_patch	matched_random	nuisance_patch	gain_random	gain_nuisance	recover_fraction
BioMistral-7B	dosage	0.023	0.003	-0.012	0.020	0.035	0.444
BioMistral-7B	laterality	0.821	0.019	0.002	0.802	0.819	0.056
BioMistral-7B	negation	0.184	-0.017	-0.012	0.201	0.196	0.444
MedGemma-4B-it	dosage	0.201	0.177	0.021	0.024	0.181	0.389
MedGemma-4B-it	laterality	9.056	0.503	0.035	8.552	9.021	0.333
MedGemma-4B-it	negation	3.653	0.191	0.021	3.462	3.632	0.167
OpenBioLLM-8B	dosage	0.031	0.005	0.010	0.026	0.021	0.500
OpenBioLLM-8B	laterality	1.413	0.113	0.017	1.300	1.396	0.167
OpenBioLLM-8B	negation	0.587	0.013	0.019	0.573	0.568	0.389

Table 32: Matched OpenBio prompt-format control on four evaluation questions per text dataset and four paired draws per question. Gain is broad-corruption restoration minus norm-matched random. Llama3-user changes serialization while preserving user instructions; it does not add the model card example system prompt.

dataset	format	draws	parsed	correct	mean	lo	hi	clusters	rows
MedMCQA	llama3_user	160	6	2	-0.378	-0.468	-0.290	40	40
MedMCQA	raw	160	145	45	5.529	4.984	6.080	40	40
MedQA	llama3_user	160	4	1	0.168	0.007	0.316	40	40
MedQA	raw	160	155	44	5.765	5.314	6.249	40	40

I SAMPLED RESTORATION AND DOWNSTREAM STATES

The separately logged MedGemma 4B follow up contains 3,200 responses across 80 questions (40 VQA-RAD and 40 SLAKE), eight paired seeds and five conditions. It is a supplementary behavioral and latent support study. Cached reference scoring gives exact sham equality, while the sampled comparison does not establish reliable answer recovery under the tested policy.

Output-uncertainty and cohort-size curves reuse the main responses and question/image-group partitions. They are sensitivity analyses of the frozen cohort, not fresh 80-question or 400-question validation sets. The benchmark annotations are the correctness reference throughout; they do not constitute prospective clinical deployment validation.

I.1 GENERATION POLICY AND LATENT SUPPORT DIAGNOSTICS

The new follow-up uses MedGemma-4B and 40 evaluation questions in each visual dataset (80 total), selected before observing restoration. Each condition receives eight paired seeds at temperature 1 and top-p 0.95, with a 64-token cap. The midpoint is fixed. The five conditions are clean image, blank image, clean-donor restoration into the blank input, equal-norm random displacement, and corrupted-donor sham. A patch applies only during the generation prefill; cached decoding then continues from that intervention. This gives 3,200 additional completions on 80 questions and their image groups. The seed-level text equality of blank and sham is checked exactly.

Unlike the original candidate-context scoring experiments, this follow-up captures donors from a prompt-only forward pass. It scores the original reference under that same donor policy and separately samples completions. This explicit match avoids interpreting two different donor protocols as one score-to-behavior test. Reference matching counts every draw, including parser failures; after blanking it is a counterfactual reference-recovery diagnostic, not a claim that the altered image still supports the original answer.

Table 33: MedGemma-4B sampled-behavior follow-up with eight paired seeds per condition and a 64-token cap. Independent group support is explicit. Match is original-reference match under the frozen parser; after blanking, this is a reference-recovery diagnostic. All draws, including parse failures, remain in the denominator. Cap reports the fraction reaching 64 tokens.

Dataset	Condition	Rows	Groups	Batch	Parsed	Gap
SLAKE	blank	320	40	0.197	1.000	0.184
SLAKE	clean	320	40	0.550	1.000	0.225
SLAKE	random	320	40	0.203	1.000	0.178
SLAKE	restore	320	40	0.194	1.000	0.178
SLAKE	sham	320	40	0.197	1.000	0.184
VQA-RAD	blank	320	40	0.297	1.000	0.328
VQA-RAD	clean	320	40	0.544	1.000	0.300
VQA-RAD	random	320	40	0.306	1.000	0.294
VQA-RAD	restore	320	40	0.297	1.000	0.338
VQA-RAD	sham	320	40	0.297	1.000	0.328

The sampled behavior table reports reference matching for each condition. This supplementary study does not establish reliable answer recovery under the tested policy. Some explanations reach the token cap, so the study does not validate completed explanations. Where observed paired differences are uniformly zero, a zero width empirical bootstrap interval describes the retained sample rather than certainty about the population effect.

The follow up also exposed a scoring discrepancy between full replay and cached execution. Reference scores must use the same donor and execution policy as the reported behavioral comparison. Cached scoring forces the reference trajectory one token at a time and evaluates raw logits before the forcing processor, with the patch applied only during prefill. Original and corrected scoring records remain separately archived; this scoring check introduces no additional sampled completions.

For a one-token reference, increased log probability increases its untruncated sampling probability, but need not change greedy choice: a reference probability can rise from 0.20 to 0.25 while another answer remains more probable. Top-p truncation, multi-token answers and parser equivalence classes

introduce additional differences between the measured mean log score and sampled answer matching. Therefore a restoration score must be reported together with the relevant behavioral endpoint rather than translated into an accuracy change.

The support diagnostics evaluate the *downstream final-block* state at the final prompt position, after the midpoint intervention. Measuring only the replacement vector would be uninformative because a clean donor is natural at its source by construction. Let $\bar{h}$ be the unit-normalized downstream state and $\{\bar{h}_j^{\mathrm{cal}}\}_{j=1}^{8}$ the same checkpoint/dataset's clean calibration-prompt states. We report

$$D_{\mathrm{NN}}(\bar{h}) = \min_j (1 - \bar{h}^\top \bar{h}_j^{\mathrm{cal}}), \tag{6}$$

$$N_{\mathrm{range}}(\bar{h}) = d^{-1} \sum_{k=1}^{d} \mathbf{1}\left\{ \bar{h}_k \notin [\min_j \bar{h}_{jk}^{\mathrm{cal}}, \max_j \bar{h}_{jk}^{\mathrm{cal}}] \right\}. \tag{7}$$

We also report cosine displacement and relative Euclidean displacement from the matched clean downstream state. Eight reference prompts provide sparse support, so even an unmodified evaluation state can exceed calibration coordinate ranges. No threshold is tuned to evaluation behavior.

Table 34: Downstream final-block prompt-state diagnostics averaged over the frozen evaluation control subset. Nearest-neighbor cosine distance and coordinate-range exceedance use eight clean calibration prompts. These are empirical support indicators, not manifold-membership tests or activated-neuron counts.

Dataset	Condition	N_cosine	Utside_calibration_range	Osine_from_clean	Elative_l2_from_clean
SLAKE	blank	0.014	0.194	0.009	0.157
SLAKE	clean	0.014	0.222	-0.000	0.000
SLAKE	random	0.014	0.212	0.010	0.163
SLAKE	restore	0.013	0.190	0.008	0.144
SLAKE	sham	0.014	0.194	0.009	0.157
VQA-RAD	blank	0.011	0.148	0.009	0.179
VQA-RAD	clean	0.014	0.146	0.000	0.000
VQA-RAD	random	0.012	0.162	0.009	0.185
VQA-RAD	restore	0.012	0.135	0.007	0.158
VQA-RAD	sham	0.011	0.148	0.009	0.179

Restoration moves the downstream state somewhat closer to its matched clean counterpart (mean relative Euclidean displacement falls from 0.181 to 0.171 on VQA-RAD, and 0.196 to 0.160 on SLAKE), without improving sampled matching. However, clean evaluation states themselves have larger mean nearest-calibration distance than blank states on both datasets (VQA-RAD 0.0238 versus 0.0187; SLAKE 0.0148 versus 0.0142). Removing distinctive image information can make a state more typical of a sparse reference set. Thus, lower nearest-neighbor distance is not sufficient evidence of preserved visual information, natural computation or answer correctness.

These quantities are diagnostics of empirical support, not tests proving on-manifold computation. Residual coordinates are signed dense features; range exceedance is not a count of newly activated MLP units. A complete hidden-pathway audit would additionally measure the relevant nonlinear unit activations and establish their natural support. Work on divergent counterfactual representations distinguishes harmless deviations from deviations that recruit different computational pathways (Grant et al., 2026). A donor drawn from a natural run can still form an unnatural joint configuration with the target's other tokens, heads and cached state. Our downstream measurements expose this issue without asserting that proximity certifies faithful causal mediation.

J NUMERICAL INTEGRITY AND EXECUTION AUDITS

This section records implementation checks separately from the main 400 question analysis. The corrected primary estimates use Eager attention, fixed padding, FP32 encoder evaluation, and the validated 16-cell cohort. Historical SDPA outputs, failed attempts, and recovery artifacts remain retained with their command logs.

J.1 Encoder correction and eligibility

The original dynamic-padding SDPA path changed vectors for identical text under different batch compositions. Batch replay reproduced the historical discrepancy; Eager attention with fixed padding and unique-text encoding removed the tested duplicate discrepancy. This is a local implementation diagnosis, not a prevalence claim for every backend or hardware stack. The main geometry tables use corrected vectors, while the historical first-stored-vector analysis is preserved as an audit artifact.

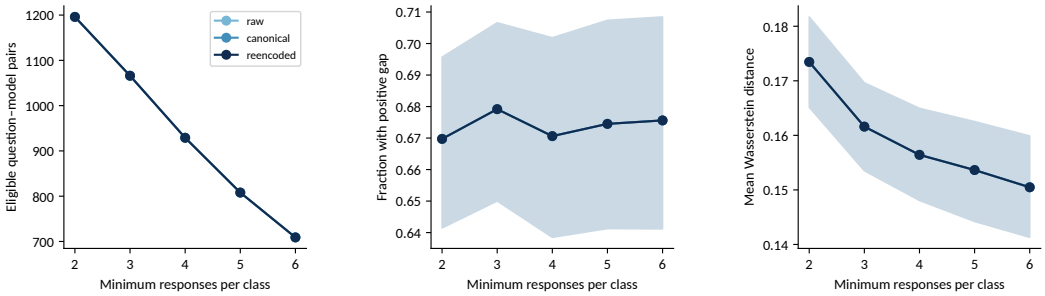


Figure 18: Geometry eligibility and encoder correction. Shading denotes 95% question/image-cluster intervals. Re-encoded uses Eager attention, fixed padding and one computed vector per exact text. Canonical only reuses the first stored vector; it does not repair that vector. All tendencies are conditional on eligibility.

Table 35: Exact original batch replay on the largest duplicate discrepancy in OpenBio/MedMCQA for each encoder. Distances compare the same text; Eager attention removes the large batch-dependent discrepancy. This reproduces an implementation-specific SDPA path, not a claim about every SDPA backend.

encoder	attention	condition	singleton_l2	old_i_l2	old_j_l2
PubMed	SDPA	singleton	0.000	1.323	0.207
PubMed	SDPA	duplicate_batch	0.000	1.323	0.207
PubMed	SDPA	long_neighbor	0.000	1.323	0.207
PubMed	SDPA	fixed512	0.000	1.323	0.207
PubMed	SDPA	original_batch_i	1.323	0.000	1.320
PubMed	SDPA	original_batch_j	0.207	1.320	0.000
PubMed	Eager	singleton	0.000	1.323	0.207
PubMed	Eager	duplicate_batch	0.000	1.323	0.207
PubMed	Eager	long_neighbor	0.000	1.323	0.207
PubMed	Eager	fixed512	0.000	1.323	0.207
PubMed	Eager	original_batch_i	0.000	1.323	0.207
PubMed	Eager	original_batch_j	0.000	1.323	0.207
General	SDPA	singleton	0.000	1.177	0.519
General	SDPA	duplicate_batch	0.000	1.177	0.519
General	SDPA	long_neighbor	0.000	1.177	0.519
General	SDPA	fixed512	0.000	1.177	0.519
General	SDPA	original_batch_i	1.177	0.000	1.202
General	SDPA	original_batch_j	0.519	1.202	0.000
General	Eager	singleton	0.000	1.177	0.519
General	Eager	duplicate_batch	0.000	1.177	0.519
General	Eager	long_neighbor	0.000	1.177	0.519
General	Eager	fixed512	0.000	1.177	0.519
General	Eager	original_batch_i	0.000	1.177	0.519
General	Eager	original_batch_j	0.000	1.177	0.519

Eligibility requires five parsed responses in each correctness class. The main coverage table retains all 6,400 evaluation pairs, including single-class and unparsed cases. Threshold scans are sensitivity analyses and do not select a favorable threshold. No missing pair is converted to a zero effect.

Table 36: Eligibility sensitivity for biomedical full responses. Intervals bootstrap independent question/image groups, retaining shared checkpoints. BH families comprise all available representations and both original partitions within each cell at each threshold; thresholds are separate sensitivity analyses.

policy	threshold	pairs	positive	fraction	lo	hi	clusters	w1	w1_lo	w1_hi	bh
canonical	2	1196	801	0.670	0.641	0.696	782	0.173	0.165	0.182	130
canonical	3	1066	724	0.679	0.650	0.707	727	0.162	0.154	0.170	141
canonical	4	929	623	0.671	0.638	0.702	669	0.156	0.148	0.165	145
canonical	5	808	545	0.675	0.641	0.707	610	0.154	0.144	0.163	149
canonical	6	709	479	0.676	0.641	0.708	553	0.150	0.141	0.160	140
raw	2	1196	801	0.670	0.641	0.696	782	0.173	0.165	0.182	130
raw	3	1066	724	0.679	0.650	0.707	727	0.162	0.154	0.170	141
raw	4	929	623	0.671	0.638	0.702	669	0.156	0.148	0.165	145
raw	5	808	545	0.675	0.641	0.707	610	0.154	0.144	0.163	149
raw	6	709	479	0.676	0.641	0.708	553	0.150	0.141	0.160	140
reencoded	2	1196	801	0.670	0.641	0.696	782	0.173	0.165	0.182	130
reencoded	3	1066	724	0.679	0.650	0.707	727	0.162	0.154	0.170	141
reencoded	4	929	623	0.671	0.638	0.702	669	0.156	0.148	0.165	145
reencoded	5	808	545	0.675	0.641	0.707	610	0.154	0.144	0.163	149
reencoded	6	709	479	0.676	0.641	0.708	553	0.150	0.141	0.160	140

Table 37: Parsed-response support for the propensity target, including both original partitions. Zero-parsed questions are excluded from propensity analysis; all retained questions otherwise receive equal weight.

model	dataset	questions	minimum	median	maximum	zero	below8	below16
BioMistral-7B	MedMCQA	408	23	32.000	32	0	0	0
BioMistral-7B	MedQA	408	24	32.000	32	0	0	0
Lingshu-7B	MedMCQA	408	32	32.000	32	0	0	0
Lingshu-7B	MedQA	408	32	32.000	32	0	0	0
Lingshu-7B	SLAKE	408	32	32.000	32	0	0	0
Lingshu-7B	VQA-RAD	408	32	32.000	32	0	0	0
MedGemma-1.5-4B-it	MedMCQA	408	4	31.000	32	0	2	5
MedGemma-1.5-4B-it	MedQA	408	10	30.000	32	0	0	7
MedGemma-1.5-4B-it	SLAKE	408	32	32.000	32	0	0	0
MedGemma-1.5-4B-it	VQA-RAD	408	32	32.000	32	0	0	0
MedGemma-4B-it	MedMCQA	408	32	32.000	32	0	0	0
MedGemma-4B-it	MedQA	408	32	32.000	32	0	0	0
MedGemma-4B-it	SLAKE	408	32	32.000	32	0	0	0
MedGemma-4B-it	VQA-RAD	408	32	32.000	32	0	0	0
OpenBioLLM-8B	MedMCQA	408	12	31.000	32	0	0	2
OpenBioLLM-8B	MedQA	408	21	32.000	32	0	0	0

J.2 SCOPE OF THE ENCODER DISCREPANCY

The recorded incident concerns two pinned sentence encoders. The biomedical checkpoint is NeuML/pubmedbert-base-embeddings. The General checkpoint is sentence-transformers/all-MiniLM-L6-v2. The retained diagnostic environment used PyTorch 2.6.0+rocm6.2.4, Transformers 4.57.6 and the study's MI210 execution stack. Inputs used mean pooling, unit normalization and a 512-token truncation limit. The original SDPA path used dynamic batch padding. The incident is established by duplicate-text discrepancies and exact replay of original batch compositions; it is not a cross-hardware prevalence estimate or an identification of one universally defective kernel.

The largest selected duplicate discrepancy was approximately 1.320 in PubMed vectors and 1.202 in MiniLM vectors. Replaying the original batches reproduces the historical vectors, while Eager attention makes the tested singleton and batch contexts agree within approximately 2.2×10^{-6}. This includes tests with repeated texts and long neighboring sequences. The main results were recomputed with Eager FP32 inference, fixed padding, and unique-text encoding followed by scattering back to response order. The correction preserves response multiplicities, and the integrity audit verifies 19,498 duplicate vector instances exactly within corrected runs. Deduplication alone would make copies agree without establishing that their first stored vector was sound, which is why batch replay and re-encoding are both necessary.

J.3 HISTORICAL GEOMETRY AUDIT AND EXAMPLES

Table 38 examines historical evaluation rows crossing the within-cell BH threshold in the primary raw biomedical representation. The audit retains the original statistic and recomputes it after assigning every exact duplicate string the first stored embedding for that string. It preserves response counts and correctness labels and uses the same 499 label permutations. The operation does not merge duplicate responses into a single observation; it makes the representation consistent for identical text while preserving the empirical frequency of each text.

Table 38: Embedding-consistency audit of within-cell threshold crossings and retained qualitative cases. Identical text is assigned its first stored embedding. Maximum Wasserstein changes summarize numerical duplicate consistency; complete question-level records remain in the audit CSV. This is not a new corrected hypothesis family.

model	dataset	audited_pairs	max_w1_change	mean_unique_texts
BioMistral-7B	MedMCQA	38	0.000	25.632
BioMistral-7B	MedQA	28	0.000	27.464
Lingshu-7B	MedMCQA	7	0.000	8.857
Lingshu-7B	MedQA	13	0.000	9.308
Lingshu-7B	VQA-RAD	2	0.000	5.500
MedGemma-1.5-4B-it	MedQA	2	0.000	31.500
MedGemma-1.5-4B-it	SLAKE	7	0.000	32.000
MedGemma-1.5-4B-it	VQA-RAD	7	0.000	32.000
MedGemma-4B-it	MedMCQA	1	0.000	32.000
MedGemma-4B-it	VQA-RAD	1	0.000	32.000
OpenBioLLM-8B	MedMCQA	19	0.000	21.684
OpenBioLLM-8B	MedQA	24	0.000	23.083

The Lingshu/VQA-RAD cardiac-silhouette question contains only two exact strings, each aligned with one answer class. Under canonical embeddings, both within-class distance laws are point masses at zero. The original W_1 value of 3.172004×10^{-7} is therefore not interpreted as a medical effect. Other audited cases retain nonzero distance-law differences. For BioMistral/MedQA question 797, the canonical W_1 is approximately 0.3114 compared with 0.3108 originally, but the signed mean gap changes from approximately 0.2765 to 0.1229. This larger change in the signed summary demonstrates that duplicate-embedding consistency can affect more than a tiny threshold crossing. This first stored vector audit alone does not identify the cause. The subsequent batch replay and re-encoding checks are reported in Section J.1.

The global multiple-testing sensitivity pools the main eligible geometry rows across representations, cells, and partitions, using their recorded permutation values. Within-cell and pooled BH counts are both retained because thresholds depend on the declared family. The full pooled-adjustment CSV is retained, but the manuscript does not switch families to obtain a preferred number of discoveries. Neither correction supplies arbitrary-dependence guarantees, and all substantive examples are interpreted with their effect sizes and response structure.

Table 39 supplies short, traceable excerpts. The main-text movement-disorder example is chosen to demonstrate a reversed dispersion direction. The two-string cardiac-silhouette example is chosen to demonstrate a numerical pathology. The AP-view and BioMistral immunology examples illustrate the interaction between answer modes and explanatory content. These are analytically selected cases, not a random clinical error sample. Their full texts, source questions, option mappings, and response IDs remain in the evidence JSON. Reference matching certifies only the extracted answer under the benchmark rule.

Table 39: Traceable qualitative examples from measured response sets. Excerpts are limited to the first 90 characters; full responses and IDs are retained in examples.json. Matching denotes the benchmark answer rule, not validation of the explanation.

Model / gold	Question	Matching response excerpt	Nonmatching response excerpt
Lingshu-7B / no	VQA-RAD-21-daea93bd8ad1	No. 1. The clavicles are not equidistant from the spinous processes of the thoracic verteb	Gap. The heart is normal in size and contour. There : The lungs are clear bilatereral witho
BioMistral-7B / D	MedQA-797-7ae08ed8f2ea	Answer: D. IL-4	Answer: C. IFN-α. The development of invasive aspergillosis is one of the common complicat
Lingshu-7B / C	MedQA-322-87549258a961	Answer: C The patient's symptoms and clinical presentation suggest Parkinson's disease, w	Answer: A A The patient's symptoms and physical examination findings suggest Parkinson's

J.4 Intervention replay checks

The accepted main intervention records have exact sham equality. Two generation shards failed their sham assertion during initial execution; both original directories remain archived. Prospective replay of the affected controls passed three repetitions, but the original and replayed intervention effects differed by as much as 0.310 nats. We therefore disclose both versions and treat intervention estimates as conditional on the tested execution path. A passing sham is an implementation check, not proof that between-run causal effects are stable.

K Discussion

K.1 Geometric patterns depend on sampling and representation

Our results do not establish a universal geometric signature of medical error. Incorrect responses can be dispersed, form compact modes, or repeat a small set of answers. Separation and dispersion also capture different properties: correct and incorrect responses can occupy distinct regions while having similar within class distance distributions. The observed geometry therefore depends on response sampling, answer content, and representation choice.

Eligibility further limits interpretation. Comparing correctness classes requires sufficient responses from both, excluding questions with uniformly correct or incorrect sampled answers. The observed dispersion tendency describes this selected subset and cannot be extended directly to all evaluation questions. Consistently incorrect questions are particularly important because agreement among repeated answers can coexist with benchmark error.

Within these limits, geometry can guide further experiments. A compact incorrect mode motivates testing a recurring alternative answer; sensitivity to masking motivates separating answer identity from

explanatory content. Such patterns generate hypotheses whose predictive and causal implications require separate evaluation.

K.2 Answer content and task structure shape correctness signals

Masking explicit answers reduces Fisher classification performance across both encoders, showing that answer content contributes to separation. This control does not fully remove answer information: rationales can still contain diagnoses, option references, or related terms. Remaining separation therefore cannot be attributed directly to medical reasoning.

Answer type also provides a strong baseline for visual majority error prediction. This association may reflect task difficulty, response constraints, or benchmark scoring, and the present analysis does not isolate their contributions. Residual gains beyond answer type vary across datasets and fitting partitions, making the baseline essential to interpreting probe performance.

Medical descriptors require similar care. Lexical indicators identify whether a question mentions laterality, negation, quantities, or treatment; they do not establish whether the model used that information correctly. Dataset references define the correctness target for this audit, while claims about clinical equivalence or evidence use require additional assessment.

K.3 Predictive accessibility depends on fitting support

Successful probing shows that information is recoverable under a specified fitting and evaluation procedure. It does not establish that the generator uses that information, and the two probe targets answer different questions: states before generation predict majority error across sampled responses, whereas states after a generated prefix predict individual response errors.

The small calibration set also constrains interpretation. Eight calibration questions provide limited support for fitting a direction in a large representation space. Variation across partitions may reflect estimator sensitivity rather than the absence of a stable representation. Likewise, low agreement between fitted directions does not rule out a shared multidimensional subspace. Stronger conclusions require broader fitting support, independent evaluation groups, and explicit tests of transfer and subspace stability.

Output uncertainty offers a complementary comparison, but its information budget differs. It uses repeated generations to predict the majority error label of that response set, whereas prompt residuals are available before generation. Comparisons should therefore state both the prediction target and when the required information becomes available.

K.4 Intervention effects require precise behavioral interpretation

Residual restoration tests whether inserting a clean state into a corrupted computation changes a specified output. The large OpenBioLLM contrasts show increased reference answer scores relative to matched random perturbations under broad corruption. They do not identify a particular medical fact or show that relevant information survived in the corrupted state. Dependence on prompt serialization further makes the interaction protocol part of the result.

The sampled MedGemma study provides a separate behavioral test under a fixed donor and generation policy. It shows no observed recovery benefit in the tested subset. Because this policy differs from the original scoring experiments, it does not establish a general relationship between their score changes and sampled accuracy. Such a relationship requires scoring and generation under matched interventions.

Execution checks impose another qualification. Exact sham behavior verifies one implementation property, while differences between original and replayed effects reveal a separate reproducibility concern. Similarly, proximity to a sparse set of clean residual states does not certify natural computation or preserved medical information. These controls help delimit an intervention claim rather than independently establish a mechanism.

L Open Questions and Future Research Directions

MIRAGE separates geometric patterns, predictive information, and intervention effects, but leaves open how these signals can support a more general understanding of medical errors. The following questions connect our findings to experiments that could clarify their meaning and practical value.

✓ **Can correctness information be separated from answer identity?** Masking explicit answers reduces separation, while answer type explains substantial predictive variation in visual tasks. Yet masking can leave answer clues in explanations and may also remove useful semantic content. A stronger test would vary correctness while matching answer format, length, and terminology, then evaluate transfer to different answer identities and question types. This could clarify whether a representation tracks evidence consistency, task structure, or recurring answer patterns.

✓ **Is predictive instability a property of the representation or the estimator?** Our probes use only eight calibration questions, so variation across fitted directions may reflect limited support rather than an unstable underlying signal. Future studies could increase independent calibration groups and compare individual directions with subspaces whose dimensions are selected using training data. Evaluating both predictive transfer and subspace agreement would help distinguish insufficient fitting data from information that is specific to a task or checkpoint.

✓ **When does changing an internal score improve an answer?** Residual interventions can change reference answer likelihood without producing an observed recovery benefit. The connection may depend on the donor state, intervention position, decoding policy, and the distribution of competing answers. Experiments that measure scores and sampled behavior under the same intervention could identify when probability changes translate into reliable answer changes. They should also measure damage to initially correct responses, since recovery alone does not establish a useful intervention.

✓ **What would establish a mechanism for using medical evidence?** Sensitivity to an image or a broad prompt corruption does not identify the clinical information responsible for an answer. A stronger experiment would change one relevant feature, such as laterality or presence of a finding, while preserving other evidence. Interventions could then test whether a candidate state transfers the corresponding answer preference across varied cases and wording, while leaving unrelated decisions stable. Expert assessment and reproducible execution would be needed to connect such selectivity to a medically meaningful interpretation.

✓ **Can these signals support reliable abstention?** Geometric separation and probe AUC do not directly determine when a model should defer. An abstention study would need a defined error target, independent calibration data, and evaluation of both retained coverage and errors among accepted answers. It should also compare the information and compute required by internal probes and repeated output methods. Whether these signals remain useful under changes in question type, image acquisition, or clinical setting is an open empirical question.