Why Large Language Models' Clinical Reasoning Fails: Insights from Explainable Deep Learning

Mirage Modi[1], Jordan E. Krull[1,2,$], Donte Johnson[1], Xiaoying Wang[1,2], Timothy D. Gauntner[2,3], Mingjia Li[3], Hao Cheng[1], Anjun Ma[1,2], Ping Zhang[1,4], Daniel G. Stover[3], Zihai Li[2,3], Qin Ma[1,2,$]

[1]Department of Biomedical Informatics, The Ohio State University College of Medicine, Columbus, OH, USA
[2]Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center – James Cancer Hospital and Solove Research Institute, Columbus, OH, USA
[3]Division of Medical Oncology, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center – James Cancer Hospital and Solove Research Institute, Columbus, OH, USA
[4]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA
[$]Corresponding author: Jordan.Krull@osumc.edu, Qin.Ma@osumc.edu

Abstract

BACKGROUND Medical large language models (LLMs) achieving high benchmark accuracy exhibit unexplained variability in clinical tasks, producing errors that clinicians cannot safeguard against. Sparse autoencoders offer a mechanistic interpretability approach to reveal how models represent medical knowledge and why they fail.

METHODS We evaluated clinical reasoning stability in GPT-5, MedGemma-27B-Text-IT, and OpenBioLLM-Llama3-70B using 355 systematic perturbations in physician-validated oncology cases, comparing staging and treatment against National Comprehensive Cancer Network (NCCN) and American Joint Committee on Cancer (AJCC) 8/9th edition guidelines. We trained sparse autoencoders on 1 billion tokens from 50,000 Medical Information Mart for Intensive Care (MIMIC-IV) clinical notes to analyze how each architecture encodes polysemous medical terms, performed 850 ablation experiments, and tested a two-stage retrieval intervention for sense disambiguation.

RESULTS Models exhibited dramatic reasoning instability. OpenBioLLM staging accuracy shifted from 45.9% to 99.1% based solely on prompt format. For cases with intentionally insufficient information, MedGemma generated definitive staging 100% of the time while GPT-5 appropriately indicated uncertainty, yet GPT-5 was only 7% accurate in staging IVB tumors. Overall, AJCC concordance was 97.5% for GPT-5 versus 26.8% for MedGemma and 31.0% for OpenBioLLM. Sparse autoencoder analysis revealed MedGemma's top-10 features showed 77.8% overlap across word senses (e.g., "cardiac arrest" versus "respiratory arrest") versus 13.6% for OpenBioLLM, indicating hierarchical versus distributed encoding. Ablating context-encoding features degraded prediction by 16.4 nats in MedGemma versus 6.7 nats in OpenBioLLM. A retrieval intervention improved MedGemma disambiguation by 10.2% ($p=0.022$, paired t-test) but harmed OpenBioLLM by 2.0% ($p=0.021$).

CONCLUSIONS Medical AI systems exhibit clinical reasoning fragility not captured by benchmark performance. We find that architecturally distinct models encode medical concepts differently, and interventions effective for one architecture may harm another. We recommend that safety validation must be architecture-specific, as benchmark equivalence does not imply functional equivalence.

Introduction

Large language models (LLMs) are rapidly entering clinical practice. Medical question-answering systems now report over eight million queries monthly, and domain-specific models achieve near-expert performance on standardized medical examinations.[1–3] Med-PaLM 2 attains 86.5% accuracy on United States Medical Licensing Examination–style questions, with physician evaluators preferring its responses over those written by clinicians on most quality dimensions.[2] More recent open-source medical models, including MedGemma and OpenBioLLM, demonstrate comparable benchmark performance while offering greater accessibility for research and deployment.[3,4] These advances have accelerated enthusiasm for integrating LLMs, as clinical decision support tools, into clinical documentation, diagnostic support, and treatment planning.

However, benchmark performance provides an incomplete and potentially misleading picture of clinical readiness. Recent evaluations reveal that models achieving superhuman scores on multiple-choice examinations exhibit striking deficits in clinical reasoning under realistic conditions. Beyond accuracy metrics, LLMs demonstrate fragile clinical reasoning, showing systematic overconfidence on diagnostic belief revision tasks, sensitivity to minor prompt wording changes (up to 10% accuracy shifts), and susceptibility to propagating planted false information in up to 83% of adversarial cases.[5–7]

Current evaluation paradigms rely on accuracy metrics from curated benchmarks, which fail to distinguish whether models encode clinical knowledge through robust conceptual representations or through superficial statistical regularities. This distinction matters because models achieving similar benchmark performance may differ fundamentally in their capacity for reliable generalization. We hypothesize that mechanistic interpretability methods, specifically sparse autoencoders applied to model internals, can reveal whether medical LLMs develop stable, clinically meaningful representations, or whether their apparent competence masks fragile reasoning that falters under perturbation.

Rather than treating models as black boxes evaluated solely by their predictions, mechanistic methods aim to identify the computational features and circuits through which models process information.[8] Sparse autoencoders (SAEs) have emerged as an effective technique for decomposing neural network activations into interpretable components.[9] They learn dictionaries of features, corresponding to human-interpretable concepts. When applied to general-domain LLMs, SAEs have revealed features encoding concepts ranging from protein sequences and legal language to abstract notions of deception and sycophancy.[10,11]

Open resources, including Gemma Scope, provide pretrained autoencoders for the Gemma model family, encompassing over 400 autoencoders with more than 30 million learned features.[12] Yet despite this methodological maturity and the clear relevance to clinical artificial intelligence governance, no published work has systematically applied SAEs for the mechanistic interpretability of medical LLMs. This gap is consequential: healthcare applications face stringent requirements for explainability, with regulatory guidance indicating that clinical decision support lacking interpretable mechanisms may require device-level oversight.

In this study, we sought to characterize a medical-LLM critical reasoning gap in evaluating complex cases and elucidate mechanisms of poor reasoning structure in realistic scenarios. Through prompt reasoning tests on realistic oncology cases, we identify weak points in prompt structure and trait attribution for accurate assessment. We also evaluate context and reasoning structures within two open-source, SOTA medical LLMs (MedGemma-27B-Text-IT and OpenBioLLM-Llama3-70B)[3,4,13] by training and systematically analyzing large-scale SAEs on model activations of clinical notes sampled from MIMIC-IV-Note, a corpus of over 330,000 deidentified discharge summaries representing diverse critical care encounters.[14,15] With these

methods, we find that both LLMs use diverse features to encode medical knowledge, that context-encoding features are causally necessary for clinical prediction, and that clinical safety interventions require architecture-awareness to improve clinical understanding.

Materials and Methods

SYNTHETIC ONCOLOGY CASE EVALUATION

To assess clinical reasoning stability under controlled conditions, we developed three synthetic oncology cases in collaboration with two board-certified physicians: a urethral carcinoma case (UC), a non-small cell lung cancer follow-up case (NSCLC-FU), and a non-small cell lung cancer initiate treatment case (NSCLC-IT). Cases were constructed in SOAP (Subjective, Objective, Assessment, and Plan) format, with the Subjective and Objective (S&O) sections serving as LLM input, with Assessment and Plan (A&P) sections serving as expected LLM outputs. For each case, we systematically removed individual medical terms (Table S1) from the input S&O prompt, generating 111 perturbations for Urethral Carcinoma (UC), 142 for non-small cell lung cancer initiate treatment (NSCLC-IT), and 102 for non-small cell lung cancer follow-up (NSCLC-FU). Then, we measured (i) cosine similarity when no terms are removed between LLM-generated A&P embeddings versus physician-written A&P reference embeddings, (ii) tumor staging accuracy, and (iii) treatment recommendation categories (management with or without perioperative management, systemic therapy, radiation with or without chemotherapy, further workup needed, or ambiguous) per NCCN treatment guidelines. We tested three prompting strategies: full A&P generation, direct staging/treatment queries, and chain-of-thought from TNM staging to clinical staging to treatment recommendation. For chain-of-thought, we also measured concordance between model-generated TNM staging and the model-generated clinical staging versus AJCC 9th edition-(or 8th edition for UC)-clinical staging. Detailed experiments were conducted on GPT-5 Thinking, MedGemma-27B-Text-IT, and OpenBioLLM-Llama3-70B to assess consistency and accuracy across model families.

SPARSE AUTOENCODER STUDY DESIGN AND DATA SOURCES

We conducted a computational study applying top-k SAE analysis to medical LLMs using deidentified clinical notes from the MIMIC-IV-Note (Medical Information Mart for Intensive Care, version 2.2) database. For details on medical LLM selection, please see supplementary methods. MIMIC-IV-Note is a publicly accessible critical care database containing free-text clinical notes from Beth Israel Deaconess Medical Center from 2008 to 2019, comprising 331,794 deidentified discharge summaries about chief complaint, history of present illness, hospital course, physical exam findings, and discharge diagnoses across 145,915 unique patients.[14] We also utilized the MIMIC-IV-Ext-BHC extension database, which provides 270,033 clinical notes due to standardized Brief Hospital Course sections derived from discharge summaries suitable for language model analysis.[15] From this corpus, we randomly sampled 50,000 discharge notes for SAE training, ensuring representation across the full temporal range and clinical diversity of the database. This study was conducted under the MIMIC-IV data use agreement and did not require additional institutional review board approval, as all data were previously deidentified.

SPARSE AUTOENCODER ARCHITECTURE AND TRAINING

SAEs learn dictionaries of features by reconstructing dense LLM activation vectors (see supplementary methods) using sparse combinations of learned latent directions. We employed the Top-K SAE, which enforces sparsity by retaining only the K-highest magnitude latent activations during the forward pass rather than relying on L1 regularization, following the training scheme of OpenAI.[9] We use a 12x expansion factor when training our SAE (65,536 latent for MedGemma, 98,304 latent for OpenBioLLM), with a fixed sparsity k of 128. These expansion factors and sparsity levels were selected based on prior work demonstrating that 10-15x

expansion with approximately 100 active features yields interpretable decompositions with acceptable reconstruction fidelity[9] and our own small-scale experiments (**Figs. S1-4**; Supplementary Methods). The SAE was trained using the Adafactor optimizer with a learning rate $2*10^{-4}$, a micro-batch of 32,768 tokens, gradient clipping at 1.0, and an auxiliary loss coefficient of 0.03125 with auxiliary sparsity of 512 for dead latent recovery (**Table S1**). Training proceeded for 200,000 steps with validation every 2,000 steps. The reconstruction loss stabilized after having seen approximately 1 billion tokens. We report results from the calibrated checkpoints at this convergence point.

FEATURE DECOMPOSITION AND POLYSEMANTICITY ANALYSIS

We measured the unique latent patterns for each token to determine how the medical LLMs represent polysemous medical terms versus monosemantic medical terms (see supplementary methods for annotation strategy). We compute the pairwise Jaccard similarity between feature sets activated for tokens in different sense categories (cardiac arrest versus respiratory arrest). High overlap indicates features that activate regardless of sense (i.e., lexical representation), and low overlap indicates sense-discriminating features (i.e., semantic representation). We fit univariate L2-regularized logistic regression probes for each latent using Adam optimization to distinguish medical entities from background text. For each latent, we computed the area under the receiver operating characteristic curve (AUROC) for binary classification of each entity type. Latent for each sense of each polysemous term were also fitted using a generalized linear mixed model to measure the entropy of each polysemous term.

ABLATION AND CAUSAL ANALYSIS

To assess whether identified features causally influence LLM predictions, we performed activation ablation experiments measuring the change in cross-entropy loss (ΔCE) when specific latent directions were zeroed:

$$\Delta CE = \frac{1}{|M|}\sum_{t\in M}\log\frac{P_{base}(y_t \mid y_{<t})}{P_{SAE}(y_t \mid y_{<t})}\ , \tag{1}$$

where M represents the valid tokens. Positive ΔCE indicates prediction degradation from the intervention. For each medical entity category, we identified sense-linked latent from probe analysis and measured ΔCE when ablating those features during forward passes on held-out clinical notes. We modeled the ablation effects using mixed-effects regression with random intercepts for document and fixed effects for medical context category, quantifying which clinical domains exhibited feature-mediated prediction changes.

RETRIEVAL EXPERIMENTS FOR SENSE DISAMBIGUATION

To validate the functional relevance of lexical features versus semantic features, we designed a two-tier structure where we re-rank high magnitude latent (lexical identity) and lower magnitude latent (contextual semantics). Given a query token such as “discharge”, we retrieve the nearest-neighbor tokens of the same word and evaluate whether the retrieved tokens shared the query’s semantic sense. Terms were only included for analysis if they had ≥30 total token instances and ≥2 senses with ≥5 instances each. For each query token, we computed cosine similarity against all-other same word tokens using three feature subsets: full (all 128 latent), lexical only (top-10 latent), and semantic-only (latent 11-128). The two-stage re-ranking retrieves the top-k candidates by full similarity first, then re-ranks these *k* candidates by semantic-only similarity. We swept candidate pool sizes $k \in \{5,10,15,20,\ldots,100\}$ to identify optimal retrieval depth. Performance was measured by precision at rank 1 (P@1) and mean average precision (MAP). For each *k*, we computed the improvement over baseline ($\Delta P@1 = P@1_{re-rank} -$

$P@1_{full\ similarity}$) across all query tokens. Statistical significance was assessed using paired t-tests comparing re-ranked versus baseline P@1 for each query, with 95% confidence intervals computed as 1.96*SE.

Results

BENCHMARK PERFORMANCE DOES NOT PREDICT CLINICAL REASONING STABILITY

Benchmarking tasks in LLMs have largely involved question-and-answer on relatively routine medical situations, so we evaluated whether medical AI systems maintain consistent clinical reasoning under realistic conditions by presenting three synthetic oncology cases to leading medical LLMs: GPT-5 Thinking, MedGemma-27B-Text-IT, and OpenBioLLM-Llama3-70B. Each case was constructed in a standard SOAP format by two board-certified medical oncology physicians, to mimic a real scenario seen in an outpatient oncology clinic visit. We systematically removed individual medical terms (see methods) from each case, yielding 355 total perturbation trials per model (**Fig. 1A**). We prompted several models to produce an A&P from the S&O, which displayed marked variability compared to physician-produced A&P, when no medical terms were perturbed (**Fig. 1B**). Using two open source medical LLMs and GPT-5 Thinking, each model was prompted iteratively for each perturbed version of the notes to produce an A&P, under two prompt formats (produce a full A&P or direct prompting for clinical stage and treatment), which displayed marked variability between models in tumor staging, depending on prompt format (**Fig. 1C**). Staging accuracy varied substantially with prompt structure in model- and case-specific manners. For UC, OpenBioLLM accuracy shifted from 45.9% to 99.1% based solely on whether staging was embedded in a full A&P or queried directly; MedGemma showed similar instability (74.8% to 98.2%). GPT-5 Thinking was largely incorrect on NSCLC-IT (93-100% incorrect). Most concerning, on NSCLC-FU, where clinical information was intentionally insufficient, MedGemma provided definitive staging 100% of the time when directly prompted, whereas GPT-5 appropriately indicated uncertainty. This inconsistent reliability across models and cases undermines clinical trust regardless of any single high-accuracy result.

MedGemma and OpenBioLLM exhibited sensitivity to clinically irrelevant information (**Fig. 1D**). Removing “creatinine 0.80” from the UC prompt, a term with no relevance to tumor staging, resulted in MedGemma changing its staging from correct (Stage III) to incorrect Stage II. Conversely, when clinically relevant terms such as “invades the anterior vagina” were removed, OpenBioLLM maintained its original staging, with similar staging variability for MedGemma direct prompting when staging is unknown (NSCLC-FU). GPT-5 demonstrates resistance to irrelevant text perturbations and correctly changes its answers in a clinically aligned manner for relevant terms (“anterior vagina”, UC) but also remains consistent even when incorrect (NSCLC-IT).

Treatment recommendations showed comparable variability (**Fig. 1E**). For NSCLC-IT, OpenBioLLM recommended definitive treatment options in 99.3% of full A&P responses but provided an ambiguous treatment recommendation in 100% of direct queries. When staging is unknown, we find the models tend to provide a spread of treatment recommendations (NSCLC-FU). These findings indicate that benchmark performance, where all tested models achieve state-of-the-art accuracy on standardized medical examinations, fails to predict reasoning stability in clinical contexts.

CHAIN-OF-THOUGHT PROMPTING DOES NOT RESOLVE REASONING DEFICITS

Because the models provided more treatment recommendations when staging was unknown, we next tested whether explicitly prompting models to reason stepwise through TNM staging before providing treatment recommendations would improve consistency (**Fig. 2A**). Under

chain-of-thought (CoT) prompting, models first extracted TNM components, then derived clinical stage, then recommended treatment, mimicking clinical reasoning workflows. CoT prompting did not reliably improve staging accuracy. When models' TNM-derived clinical stages were compared against AJCC 9th edition staging guidelines (or 8th edition for UC) across all 355 perturbations, systematic errors persisted. GPT-5 achieved 97.5% concordance, while MedGemma achieved only 26.8% and OpenBioLLM achieved 31.0% (**Fig. 2B**). For example, MedGemma reported "T4N3M1b (Stage IVB)", which should be Stage IVA according to the AJCC 9th edition. This pattern suggests memorized associations rather than principled applications of staging algorithms.

Treatment concordance between CoT and direct prompting strategies varied substantially by case complexity (**Fig. 2C**). For the UC case, all models recommended appropriate definitive management regardless of prompting strategy. However, for NSCLC-FU (staging to be determined), MedGemma recommended definitive management in 98% of CoT responses, while GPT-5 appropriately recommended workup in 100% of responses. OpenBioLLM recommended radiation or systemic therapy in over 90% of responses without addressing the staging ambiguity. GPT-5 provided correct treatment independent of correct staging (**Fig. 2D-E**), while MedGemma and OpenBioLLM both display a dependence on staging to arrive at the correct treatment ($p<0.001$, Fisher's exact test).

SPARSE AUTOENCODER ANALYSIS REVEALS INTERPRETABLE CLINICAL CONCEPT ORGANIZATION

To understand why models fail at clinical reasoning despite benchmarking success, we trained top-k sparse autoencoders (SAE) to decompose model activations into interpretable components (**Fig. 3A**). Dimensionality reduction on MedGemma's SAE revealed that learned features cluster into coherent clinical concept groups (**Fig. 3B**). HDBSCAN clustering identified five distinct groupings corresponding to medication lists for cardiopulmonary patients, complex medicine inpatient cases, neurovascular procedures, alcohol-related adverse events, and general medical documentation. Local lexical coherence analysis demonstrated that tokens with high feature overlap cluster spatially (**Fig. 3C**), indicating non-random organization of clinical concepts in feature space. The highest effect sizes were seen among the clusters associated with complex medicine inpatient cases (fever in immunocompromised, respiratory decompensation), suggesting an association between clinically complex cases and reliance on token frequency.

Qualitative examination confirmed clinical interpretability (**Table 1**). Feature 418 was activated maximally for dental terminology. Feature 56 bundled echocardiogram terminology ("size", "thickness", "motion", "diameter"), co-activated due to their textual proximity in the clinical notes, illustrating how next-token prediction objectives shape feature organization. Feature 52, associated with lexical meanings of household furniture ("pilar", "chairs", "mattress"), instead activated for gastrointestinal content ("stool softener", "stool extended constipation"), demonstrating unexpected clinical semantics.

ARCHITECTURALLY DISTINCT MODELS ENCODE MEDICAL POLYSEMY DIFFERENTLY

Medical terminology frequently carries multiple meanings depending on context. We examined 12 polysemous terms and 4 monosemantic terms (**Tables S2-3**) to determine whether models reliably distinguish contextual meanings. MedGemma exhibited a distinctive two-tier organization (**Fig. 4A**). The top 10 highest-magnitude features showed 77.8% ± 8.7% overlap for polysemous terms (n=182 samples) and 78.8% ± 4.1% for monosemantic terms (n=2,392 samples). However, when examining all 128 active features, overlap dropped to 16.3% ± 0.3% for polysemous terms (n=183 samples) and 17.0% ± 4.9% for monosemantic terms (n=20,000

comparisons). This 4.6-fold reduction indicates that low-magnitude features encode contextual meaning rather than word identity. OpenBioLLM showed no such separation. Top-10 polysemantic feature overlap was 13.6% ± 3.1% (n=320 samples) and 17.9% ± 2.1% for monosemantic terms (n=2,661 samples), similar to the full 128-latent overlap of 10.3% ± 0.7% (polysemous, n=317 samples) and 11.6% ± 6.2% (monosemantic, n=20,000 samples). This finding suggests information is distributed uniformly across all features regardless of magnitude. The difference between architectures in top-10 polysemous overlap was significant ($p<0.001$, two-sample Z-test). Notably, OpenBioLLM showed significantly lower k=10 overlap for polysemous versus monosemantic terms ($p<0.001$, empirical distribution test), while MedGemma showed no such difference ($p=0.75$).

CONTEXT-ENCODING FEATURES ARE CAUSALLY NECESSARY FOR CLINICAL PREDICTION AND ARE REPRESENTED BY DISTRIBUTED BUNDLES OF FEATURES

We tested whether context-specific features are functionally necessary for the LLM through targeted ablation experiments (**Fig. 4B**). For each of the 12 polysemous terms, we identified features associated with specific clinical meanings via co-occurrence statistics and "disabled" them, measuring the change in cross-entropy loss (ΔCE; **Equation 1**) for next-token prediction. Disabling context-encoding features in MedGemma produced severe prediction degradation (mean ΔCE =16.43 ± 1.59 nats, n=445 ablations), representing near-complete loss of predictive capability. Effects were consistent across clinical domains: cardiovascular (16.64±1.25 nats, n=31 ablations), renal (17.02±1.13 nats, n=29 ablations), and administrative (16.02±1.47 nats, n=19 ablations). OpenBioLLM showed qualitatively similar but smaller effects (mean ΔCE =6.69 ± 0.66 nats, n=400 ablations). The top senses show the specific lemma and its contribution to the ΔCE **(Fig 4C)**. Mixed effects regression with random intercepts for clinical notes confirmed these effects were domain-general (**Fig. 4D**). For MedGemma and OpenBioLLM, no clinical context showed significant deviation from the overall mean ($p>0.05$ for all domains, gamma mixed effects), thus disabling context-specific features produced statistically equivalent prediction degradation within each model.

We then tested whether individual features encode broad medical significance by training univariate logistic regression probes per latent to distinguish clinical entities from background text (65,536 probes for MedGemma; 98,304 probes for OpenBioLLM; **Fig. 4E**). Mean AUROC was 0.479±0.037 for MedGemma and 0.472±0.044 for OpenBioLLM, both indistinguishable from chance (0.50), with no probe achieving an AUROC above 0.70. No single feature reliably identifies medical content, meaning medical knowledge is distributed across thousands of specialized features. Mixed-effects analysis of normalized feature entropy revealed divergent architecture-specific patterns. MedGemma showed uniformly high entropy across all clinical contexts with no significant context effects ($p>0.05$, beta mixed effects). In contrast, OpenBioLLM showed significant context-dependent variation. Dermatological contexts exhibited elevated entropy relative to baseline (coefficient 2.46, 95% CI 1.23-3.68, $p=0.002$, beta mixed effects), followed by clinical contexts (coefficient 2.06, 95% CI 0.83-3.28, $p = 0.007$) and respiratory contexts (coefficient 1.54, 95% CI 0.54-2.55, $p=0.011$). This suggests that while MedGemma maintains consistent representational complexity across specialties, OpenBioLLM develops domain-specific representational patterns with greater feature dispersion in certain clinical contexts.

ARCHITECTURE-AWARE RETRIEVAL IMPROVES DISAMBIGUATION FOR MEDGEMMA BUT HARMS FOR OPENBIOLLM

The architectural differences we identified suggest a practical intervention. If MedGemma concentrates word identity in high-magnitude features and contextual meaning in low-magnitude features, retrieval systems could improve disambiguation by filtering identity-encoding features

when re-ranking. We tested this hypothesis using a same-word sense retrieval task with terms meeting minimum thresholds of 30 total instances and at least 2 senses with 5 or more instances each (**Fig. 4F**). We employ a two-stage retrieval where we retrieve top-k candidates using all 128 features, then re-rank using only features 11-128. For MedGemma (n=147 queries across 8 polysemous terms), two-stage retrieval produced a significant improvement. At optimal k=30, precision at rank one improved by 10.2% (95% CI: 1.6-18.8; p=0.022, paired t-test). This same intervention harmed OpenBioLLM (n=396 queries across 10 polysemous terms). At k=40, precision decreased by 2.0% (95% CI: -3.7 to -0.3; p=0.021). Degradation was significant at multiple pool sizes (k = 15-50; $p<0.05$). This difference between architectures demonstrates that optimization strategies derived from one architecture may actively harm another.

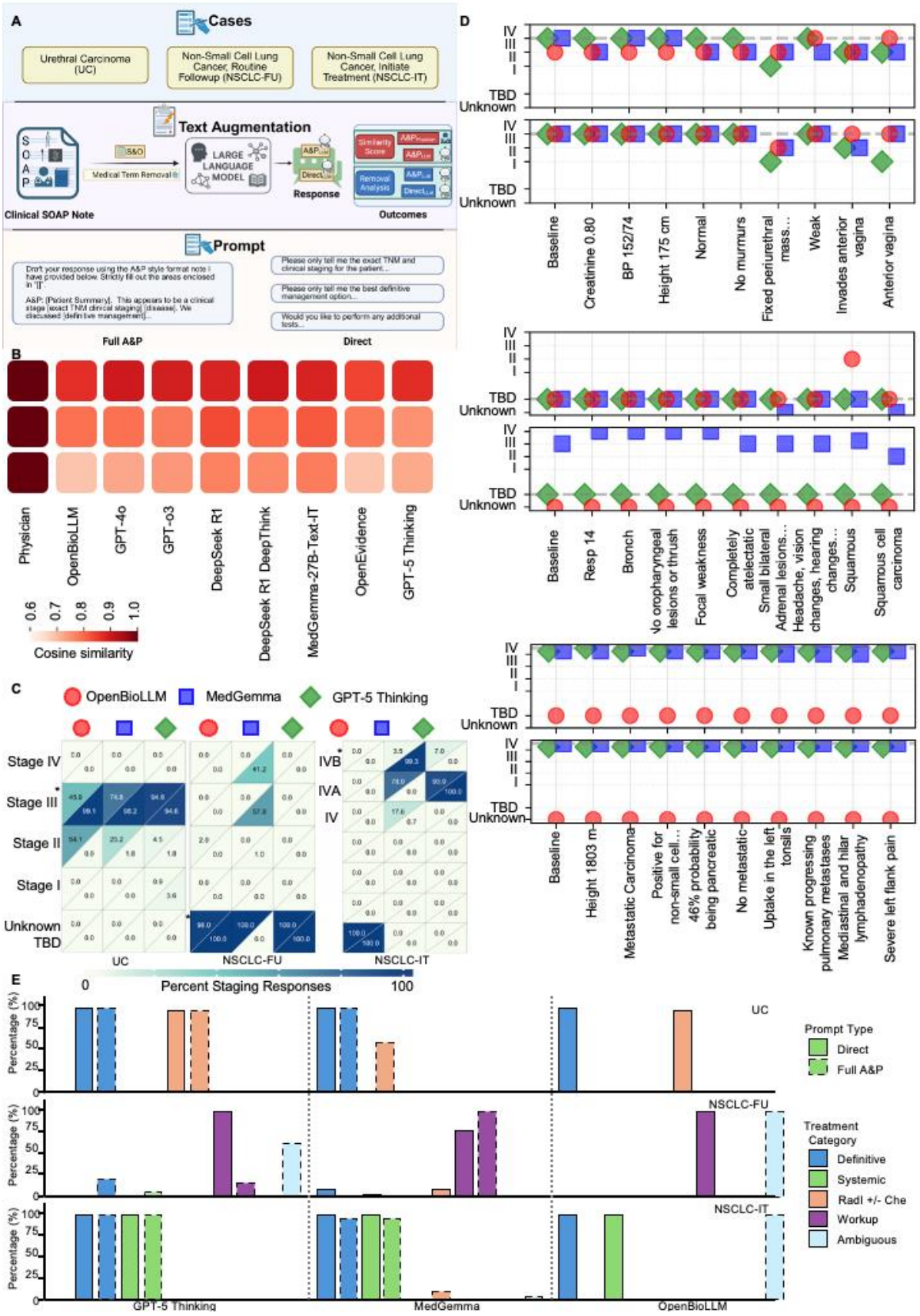
A
Cases
Urethral Carcinoma (UC)
Non-Small Cell Lung Cancer, Routine Followup (NSCLC-FU)
Non-Small Cell Lung Cancer, Initiate Treatment (NSCLC-IT)
Text Augmentation
S&O
Medical Term Removal
LARGE LANGUAGE MODEL
Response
Clinical SOAP Note
Outcomes
Prompt
Full A&P
Direct
B
Physician
OpenBioLLM
GPT-4o
GPT-o3
DeepSeek R1
DeepSeek R1 DeepThink
MedGemma-27B-Text-IT
OpenEvidence
GPT-5 Thinking
Cosine similarity
C
OpenBioLLM
MedGemma
GPT-5 Thinking
Stage IV
Stage III
Stage II
Stage I
Unknown
TBD
IVB
IVA
IV
UC
NSCLC-FU
NSCLC-IT
Percent Staging Responses
D
Baseline
Creatinine 0.80
BP 152/74
Height 175 cm
Normal
No murmurs
Fixed periurethral mass...
Weak
Invades anterior vagina
Anterior vagina
Resp 14
Bronch
Focal weakness
Squamous
Squamous cell carcinoma
Height 1803 m
Metastatic Carcinoma
No metastatic
Uptake in the left tonsils
Known progressing pulmonary metastases
Mediastinal and hilar lymphadenopathy
Severe left flank pain
E
Percentage (%)
UC
NSCLC-FU
NSCLC-IT
Prompt Type
Direct
Full A&P
Treatment Category
Definitive
Systemic
Radl +/- Che
Workup
Ambiguous
GPT-5 Thinking
MedGemma
OpenBioLLM

Figure 1. Text augmentation leads to unstable LLM clinical reasoning and decision-making. (a) Three realistic oncology cases—Urethral Carcinoma (UC), a clinical follow-up for NSCLC (NSCLC-FU), and an initial treatment visit for Non-Small Cell Lung Cancer (NSCLC-IT)—were tested as SOAP notes, with Subjective & Objective (S&O) serving as input, with medical terms being removed and different prompt strategies applied. (b) Cosine similarity scores comparing text embeddings from LLM Assessment & Plan (A&P) outputs compared to the physician A&P embedding when no terms are removed, with a higher score indicating greater similarity to the physician note. (c) LLM output tumor staging separated by model (columns: GPT-5 Thinking, OpenBioLLM, and MedGemma) and full-A&P vs. Direct Prompt (upper left vs. lower right), the heatmap represents the percentage of input terms that yield that stage when deleted from the prompt, with the * indicating the correct tumor staging. (d) Selected terms from the input prompt and the model's clinical stage output when the term or phrase is removed from the prompt, separated by full A&P prompt or direct prompt strategies. Green diamond – GPT5 Thinking, blue square – MedGemma, orange circle – OpenBioLLM, dotted line – ground truth. (e) Percentage of output predictions of treatment category (definitive management, systemic therapy, radiation +/- chemotherapy, workup, and ambiguous treatment options) for all terms deleted from the prompt, split by model and full A&P vs. direct prompt.

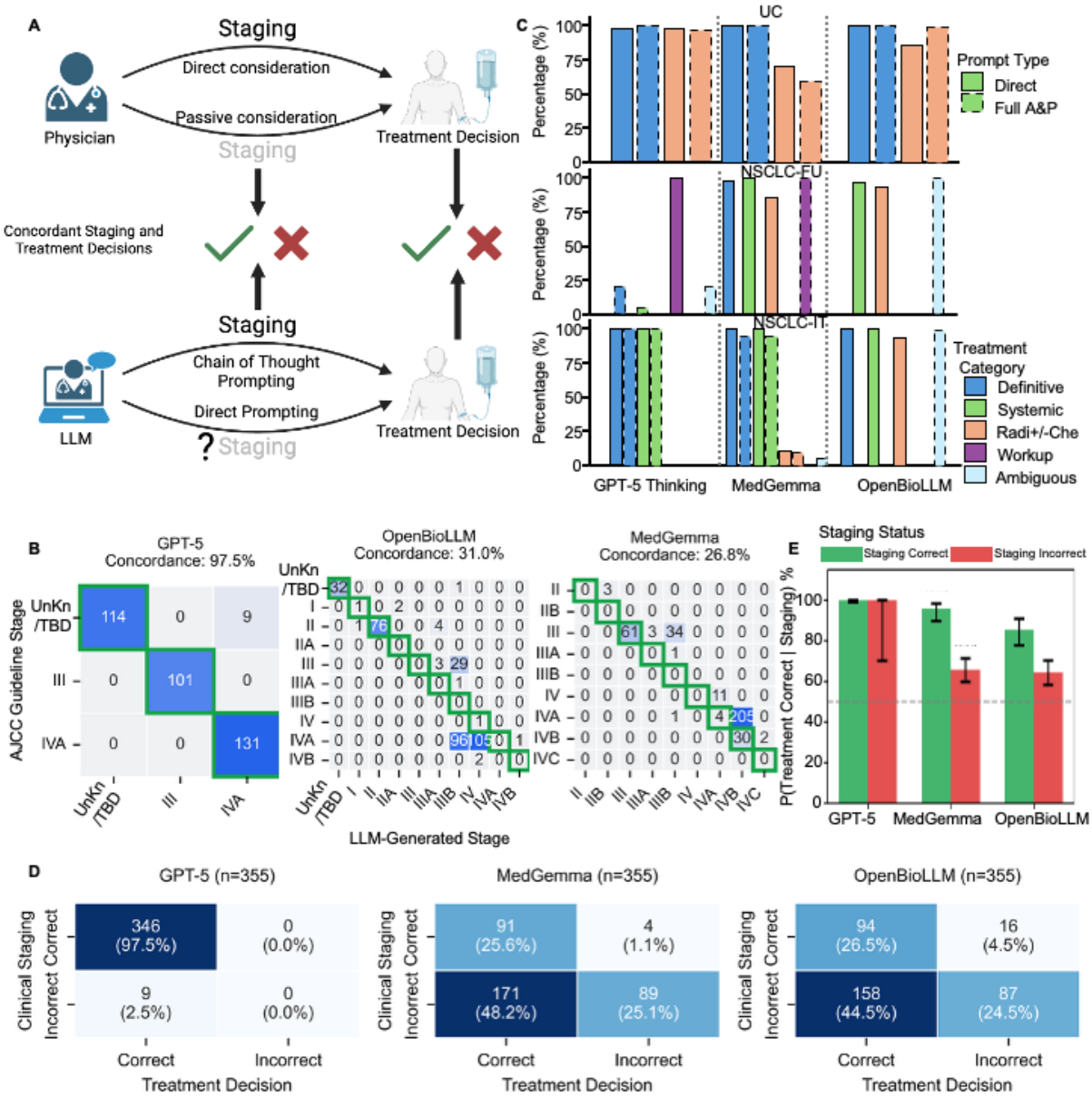

Figure 2. Chain-of-thought prompting reveals LLM memory mapping using tumor staging. (a) Schematic of the chain-of-though (CoT) evaluation: models first extract TNM staging, then derive clinical stage, then recommend treatment, compared against direct treatment queries to measure concordance to a physician's clinical thinking. (b) Confusion matrices comparing LLM-derived clinical staging (columns) against AJCC 8/9th edition staging guidelines (rows) for each model and case. Overall AJCC staging concordance was 97.5% for GPT-5, 26.8% for MedGemma, and 31.0% for OpenBioLLM. (c) Treatment category proportions by model and prompting strategy (CoT versus Direct) across three oncology cases (NSCLC-FU, NSCLC-IT, UC). Treatment categories include definitive management, systemic therapy, radiation +/- chemotherapy, and ambiguous treatment options, defined by NCCN treatment guidelines. Notable findings include OpenBioLLM shifting from 99.3% definitive to 100% ambiguous for NSCLC-IT between CoT and Direct prompting, and MedGemma recommending definitive treatment in 98.0% of NSCLC-FU CoT responses despite staging uncertainty. (d) Confusion matrix of LLM clinical staging versus LLM treatment decision. MedGemma correctly staged 97.5%

of responses with correct downstream treatment. MedGemma and OpenBioLLM correctly staged only 25.6% and 26.5% of responses, respectively, with correct downstreatm treatment. (e) Correct treatment dependency on correct stage. GPT-5 gave correct treatments 100% of the time, making it independent of correct staging (p=1.0, Fisher's exact test). MedGemma and OpenBioLLM are dependent on correct staging (p<0.001, Fisher's exact test).

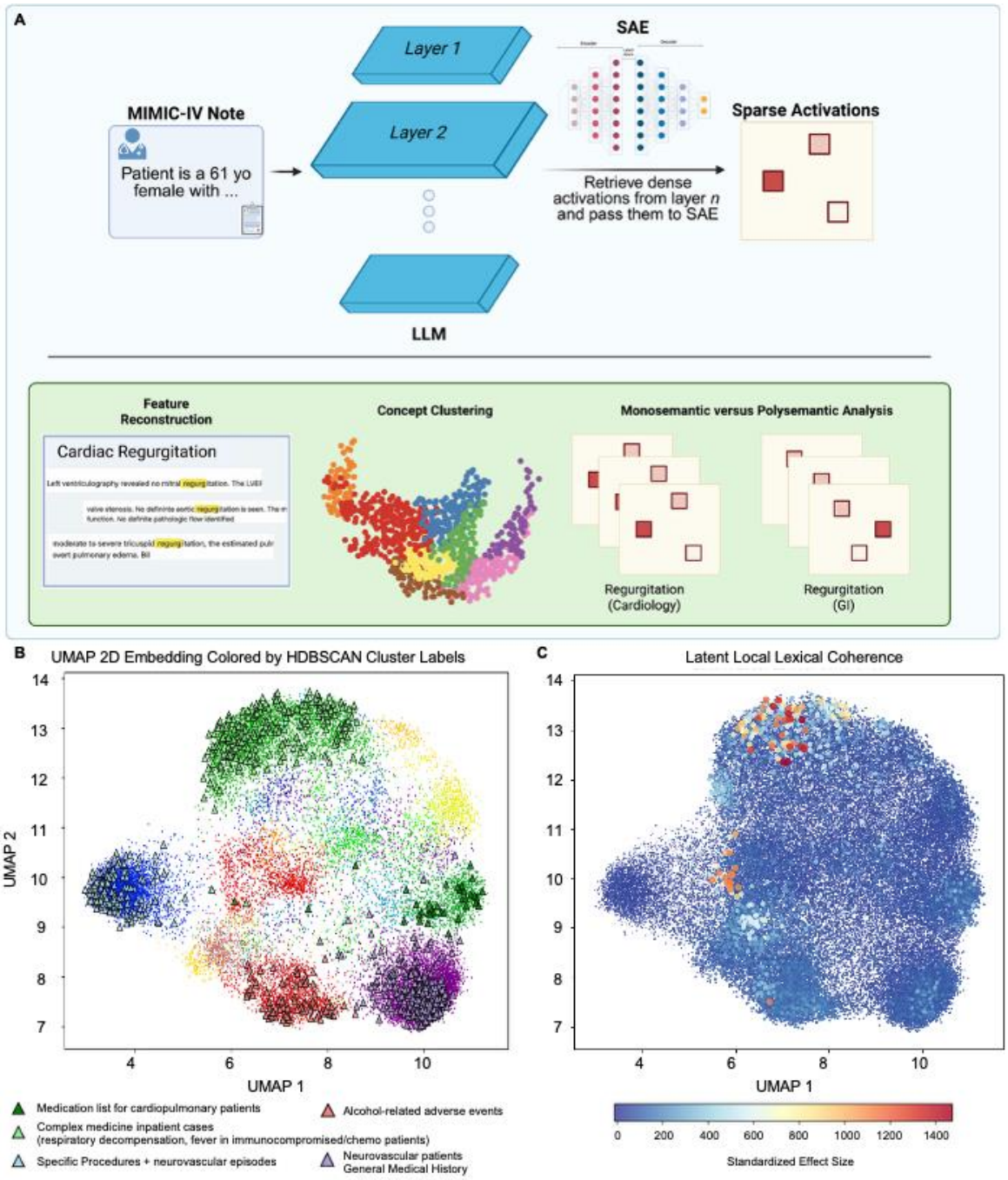

Figure 3. Sparse autoencoder reveals medical concepts in MedGemma-27B-Text-IT. (a) Graphical representation of the sparse autoencoder training. MIMIC-IV notes are sent to the medical LLM, retrieving dense activations from an internal layer. The sparse autoencoder (SAE) learns a sparse representation of the activations, from which we then perform downstream

analyses. (b) A UMAP embedding of SAE latents clustered on their dictionary values using HDBScan, with the cluster color based on HDBScan cluster labels. Diamond-shaped points are randomly selected latents from 5 of the largest latent clusters, which share conceptual value. Each point is represented by a concept listed below in the legend, which summarizes the top terms for all latents in a cluster. (c) A UMAP embedding of SAE latents, with scaled values for the standardized effect size of local lexical coherence among top terms of each latent, represented by both point color and size. Large values represent more local overlap of top latent terms, above expected word frequency, and small values represent less overlap among top latent terms within each latent's local neighborhood, than the expected term frequency overlap. Aggregation of higher value points shows local concept dependence on the presence of specific terms above the expected random occurrence frequency.

Table 1. Representative sparse autoencoder latents demonstrate interpretable clinical concept encoding.

Latent	Top Tokens	Top Examples	Interpretation
418	periodontal; tooth; dental; molar; fitting; denture; amalgam; wisdom; caries	"Pt presents with heavily restored dentition with multiple missing teeth. The upper molar as noted on the image as left side shows evidence of periapical pathology" "DENTITION: Multiple maxillary teeth are absent. There is a periapical lucency in the location tooth #6 and #7 (3: 84)" "Chronic Fatigue Syndrome Endocarditis, 1980s following dental procedure"	Dentistry-specific findings with lexical meaning related to dental materials and pathology
538	Steady; steadily; speedy; slow; deliberate; leisurely; stubbornly	"cerebellar exam intact, steady gait" "Exam was notable for a steady gait" "sensation grossly intact throughout, steady gait"	Neurological exam findings on gait with lexical meanings related to gait descriptors
56	Size; thickness; motion; height; circumference; diameter	"the left ventricular cavity size is normal. Regional left ventricular wall motion is normal." "The left ventricular cavity size is normal. There is moderate global left ventricular hypokinesis" "Right ventricular chamber size and free wall motion are normal."	Echocardiogram descriptions of heart chambers with lexical meaning of size/dimension except motion is a distinct lexical meaning; these concepts are bundled on textual proximity since "size" and "motion" are collocated neighbors.
53	Pillar; chairs; mattress	"stool softener, NOT a laxative…" "stool extended constipation…" "stool softeners, and should eat foods…"	GI disruption examples with lexical meaning of household furniture items

Table 1. Representative sparse autoencoder latents demonstrate interpretable clinical concept encoding. Selected latents from MedGemma illustrating the relationship between

lexical neighborhoods (top tokens by activation magnitude) and clinical semantics (top-activating text examples). Latent 418 encodes dentistry-specific terminology. Latent 538 captures neurological gait descriptors used in physical examination documentation. Latent 56 bundles echocardiogram terminology containing semantically distinct terms (“size”, “motion”) co-activated due to textual proximity in clinical notes, illustrating how MedGemma’s next-token prediction objectives shape feature organization. Latent 53 demonstrates unexpected lexical-semantic mapping, in which tokens associated with furniture (“pillar”, “chairs”, “mattress”) instead activate for gastrointestinal clinical content (“stool softener”, “stool extended constipation”).

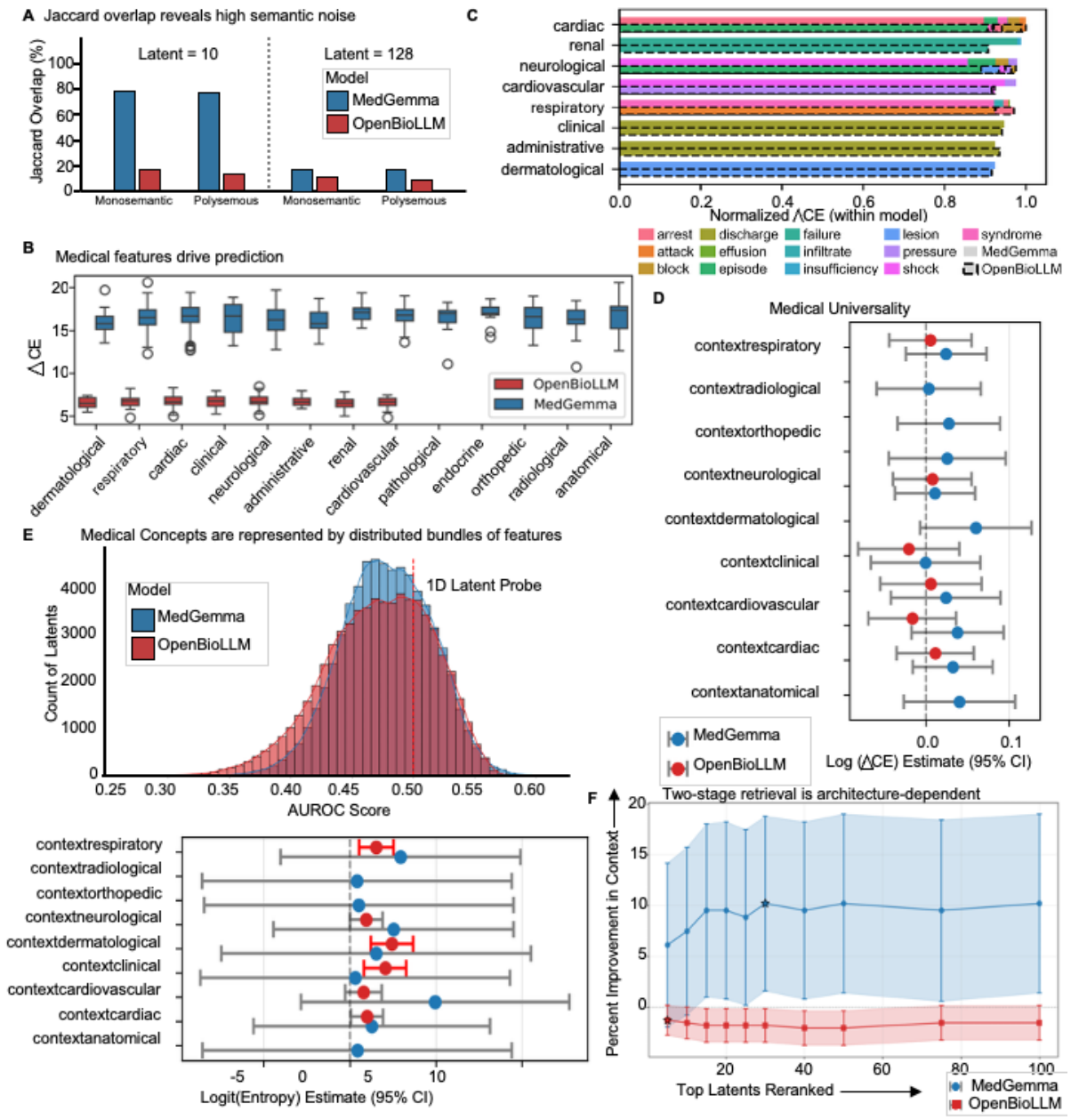

Figure 4. Sparse autoencoder analysis reveals architecture-specific encoding of medical polysemy. (A) Jaccard overlap between feature sets activated for polysemous versus monosemantic medical terms at k=10 (top features only) and k=128 (all active features).

MedGemma exhibits 77.8% overlap at k=10 regardless of word sense, dropping to 16.9% at k=128, indicating hierarchical separation of lexical identity (high-magnitude features) from medical contextual meaning (low-magnitude features). OpenBioLLM shows uniformly low overlap (10-18%) at both k values, indicating distributed clinical representations. The between-architecture difference at k=10 (64.2%) was significant ($p < 0.001$). (B) Change in cross-entropy loss (ΔCE) when ablating context-encoding features, grouped by clinical sense. MedGemma ablations produced mean ΔCE of 16.43 ± 1.59 nats (n=445 ablations) versus 6.69 ± 0.66 nats for OpenBioLLM (n=400 ablations). (C) The lemmas with highest ablation impact by sense. (D) Mixed-effects model coefficients for $\log(\Delta CE)$ by clinical context. Neither model showed significant context-dependent ablation effects ($p > 0.05$, gamma mixed effects), indicating domain-general mechanisms. (E) Distribution of single-latent probe AUROC scores for distinguishing clinical entities from background text and mixed-effects model coefficients for normalized feature entropy by clinical context.. Mean AUROC was 0.479 for MedGemma and 0.472 for OpenBioLLM, both indistinguishable from chance, with no probe exceeding 0.70. MedGemma showed no significant context effects (all $p > 0.05$, beta mixed effects); OpenBioLLM showed elevated entropy in dermatological ($p = 0.002$, beta mixed effects), clinical ($p = 0.007$, beta mixed effects), and respiratory ($p = 0.011$, beta mixed effects) contexts. (F) Two-stage retrieval schematic and results. Standard retrieval ranks by full feature similarity; two-stage retrieval first retrieves candidates, then re-ranks using only semantic features (positions 11–128). MedGemma showed gains of 10.2% in precision at rank 1 ($p = 0.022$, n=147 queries, paired t-test); OpenBioLLM showed loss of 2.0% ($p = 0.021$, n=395 queries, paired t-test). Stars indicate optimal k values.

Discussion

Our findings reveal a fundamental disconnect between benchmark performance and clinical reasoning capability in medical large language models (LLMs). LLMs achieving 85-90% accuracy on medical licensing examinations exhibited dramatic instability when processing realistic oncology cases.[3,13] OpenBioLLM tumor staging accuracy shifted from 45.9% to 99.1% based solely on prompt format, MedGemma generated a definitive staging 100% of the time despite intentionally insufficient clinical information, and GPT-5 was only 7% accurate on understanding clinical tumor stage IVB (**Fig. 1C**). Concordance with AJCC 8/9th edition guidelines reached only 26.8%-31.0% for open-source models versus 97.5% for GPT-5 (**Fig. 2B**), and MedGemma and OpenBioLLM are dependent on correct staging for correct treatment response (**Fig. 2D-E**). These findings align with growing evidence questioning standard evaluation paradigms. Script concordance testing reveals that models achieving near-expert scores on multiple-choice examinations consistently underperform clinicians on tasks requiring probabilistic reasoning under uncertainty, with systematic overconfidence manifesting as bias toward extreme responses.[16] These limitations have prompted calls to retire medical licensing examinations as primary benchmarks, given their fundamental inadequacy as metrics for clinical utility.[5] Evaluations using real patient data from MIMIC-IV demonstrate a 16-25% gap between LLM and physician diagnostic accuracy when models must gather clinical information rather than select from provided options.[5] The MedS-Bench evaluation spanning 11 clinical tasks found that even GPT-4 struggles with complex reasoning despite strong performance on medical question-answering, underscoring the gap between benchmark success and clinical demands.[17] Our perturbation analysis extends these observations by demonstrating that instability emerges not only across task types but within single clinical scenarios when superficial input characteristics change.

Our sparse autoencoder analysis provides a representational explanation for these phenomena. MedGemma organizes polysemous medical terms hierarchically, with the top-10 features per token showing 77.8% overlap regardless of clinical context, encoding word identity, while low-

magnitude features showed only 16.9% overlap, encoding contextual meaning (**Fig. 4A**). OpenBioLLM distributed information uniformly, with 10–18% overlap at all feature magnitudes. This architectural difference (hierarchical versus distributed) means that when a clinical decision support system processes a query about "arrest," MedGemma's high-magnitude features activate similarly for respiratory arrest and cardiac arrest, potentially retrieving cardiac protocols when respiratory guidance is needed. Ablation experiments confirmed these context-encoding features are causally necessary, degrading next token prediction by 16.4 nats in MedGemma and 6.7 nats in OpenBioLLM (**Fig. 4B**). These effects were domain-general, in which degradation in radiological contexts was statistically similar to cardiac domains (**Fig. 4D**), suggesting both architectures rely on distributed feature coalitions rather than localized “medical knowledge” detectors (**Fig. 4E**). Medical knowledge therefore is shown to be compartmentalized across thousands of specialized features, with dental terminology in one feature, neurological gait in another (Table 1), reinforcing theoretical models where neural networks pack more features than neurons.[18] This organization therefore challenges safety approaches that rely on detecting or steering broad content categories. The practical consequences of this representational heterogeneity are significant. Retrieval-augmented generation has emerged as a leading approach for grounding LLM outputs in verified clinical knowledge, with studies demonstrating that RAG can improve accuracy to 96.4% with zero hallucinations in surgical fitness assessment, significantly outperforming human clinicians.[19] However, RAG effectiveness varies substantially across LLM architectures, and systematic reviews identify a lack of standardized evaluation frameworks for RAG-based clinical applications.[20,21] Our two-stage retrieval intervention, designed to filter identity-encoding features during re-ranking, improved MedGemma's precision by 10.2% (p=0.022, paired t-test) but degraded OpenBioLLM's performance by 2.0% (p=0.021; **Fig. 4F**). This 12.2% difference demonstrates that optimization strategies derived from one architecture may actively harm another, consistent with evidence that prompt engineering effectiveness varies dramatically across LLM families (κ= -0.002 to0.984)[22] and that safety fine-tuning produces architecture-dependent outcomes.[23] The mechanistic basis for why these architectures developed different encoding strategies remains unclear. Possible factors include training data composition, optimization dynamics, or model capacity.[24–27] However, regardless of origin, our findings demonstrate that these representational differences have direct, measurable consequences on clinical AI safety. Current regulatory frameworks and institutional validation protocols typically assume model-agnostic safety properties[28], but our results suggest this assumption is incorrect.

Due to the significant storage and computational demands of this work, we analyzed two model architectures. Additional architectures (Phi, Qwen, GPT-OSS) may exhibit distinct patterns. Our SAE analysis used 64-token context windows, following prior work,[9] which may potentially underrepresent document-level semantic relationships. MIMIC-IV represents single-center intensive care documentation from Beth Israel Deaconess Medical Center, limiting generalization to outpatient or international settings. The three synthetic oncology cases and retrieval tasks (n=147, n=395), while demonstrating mechanistic principles, represent simplified proxies for production clinical decision support systems.

In conclusion, our results demonstrate that benchmark equivalence does not imply functional equivalence, interpretability findings do not transfer across architectures, and safety interventions must be validated architecture-by-architecture. Current regulatory frameworks and institutional validation protocols typically assume model-agnostic safety properties. Our findings indicate this assumption requires reconsideration. We recommend that healthcare institutions evaluating medical AI systems should test clinical reasoning stability under realistic perturbations, validate retrieval and safety interventions on the specific architecture being

deployed, and incorporate mechanistic interpretability methods to characterize how candidate systems encode clinical knowledge.

Disclosures

The authors declare no conflicts of interest.

This work was funded by the National Institutes of Health (NIH; T32GM139784, U54AG075931) and by the Pelotonia Institute of Immuno-Oncology.

During the preparation of the manuscript, the authors used Claude Opus 4.5 for editing assistance to improve clarity, grammar, and readability. All AI-generated suggestions were reviewed and revised by the authors, who take full responsibility for the accuracy and integrity of the final content. No new intellectual content was generated by the AI tool.

The MIMIC datasets used for our work can be found at physionet.org/content/mimic-iv-note/2.2/ and physionet.org/content/labelled-notes-hospital-course/1.2.0/.

References

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

2. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31(3):943-950. doi:10.1038/s41591-024-03423-7

3. Sellergren A, Kazemzadeh S, Jaroensri T, et al. MedGemma Technical Report. *arXiv*. Preprint posted online July 12, 2025. doi:10.48550/arXiv.2507.05201

4. Advancing Open-source Large Language Models in the Medical & Healthcare Domain. March 30, 2025. Accessed December 22, 2025. https://huggingface.co/blog/aaditya/openbiollm

5. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613-2622. doi:10.1038/s41591-024-03097-1

6. McCoy LG, Swamy R, Sagar N, et al. Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study. *NEJM AI*. 2025;2(10):AIdbp2500120. doi:10.1056/AIdbp2500120

7. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun Med*. 2025;5(1):330. doi:10.1038/s43856-025-01021-3

8. A Mathematical Framework for Transformer Circuits. Accessed December 23, 2025. https://transformer-circuits.pub/2021/framework/index.html

9. Gao L, Tour TD la, Tillman H, et al. Scaling and evaluating sparse autoencoders. *arXiv*. Preprint posted online June 6, 2024. doi:10.48550/arXiv.2406.04093

10. Simon E, Zou J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nat Methods*. 2025;22(10):2107-2117. doi:10.1038/s41592-025-02836-7

11. Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S. Zoom In: An Introduction to Circuits. *Distill*. 2020;5(3):10.23915/distill.00024.001. doi:10.23915/distill.00024.001

12. Lieberum T, Rajamanoharan S, Conmy A, et al. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. In: Belinkov Y, Kim N, Jumelet J, Mohebbi H, Mueller A, Chen H, eds. *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics; 2024:278-300. doi:10.18653/v1/2024.blackboxnlp-1.19

13. Pal A, Minervini P, Motzfeldt AG, Gema AP, Alex B. openlifescienceai/open_medical_llm_leaderboard. 2024. https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard

14. Johnson A, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV-Note: Deidentified free-text clinical notes. doi:10.13026/1N74-NE17

15. Aali A, Van Veen D, Arefeen Y, et al. MIMIC-IV-Ext-BHC: Labeled Clinical Notes Dataset for Hospital Course Summarization. doi:10.13026/5GTE-BV70

16. Raji ID, Daneshjou R, Alsentzer E. It's Time to Bench the Medical Exam Benchmark. *NEJM AI*. 2025;2(2):AIe2401235. doi:10.1056/AIe2401235

17. Wu C, Qiu P, Liu J, et al. Towards evaluating and building versatile large language models for medicine. *npj Digit Med*. 2025;8(1):58. doi:10.1038/s41746-024-01390-4

18. Elhage N, Hume T, Olsson C, et al. Toy Models of Superposition. *arXiv*. Preprint posted online September 21, 2022. doi:10.48550/arXiv.2209.10652

19. Ke YH, Jin L, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digit Med*. 2025;8(1):187. doi:10.1038/s41746-025-01519-z

20. Yang R, Ning Y, Keppo E, et al. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst*. 2025;2(1):2. doi:10.1038/s44401-024-00004-1

21. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*. 2025;4(6):e0000877. doi:10.1371/journal.pdig.0000877

22. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digit Med*. 2024;7(1):41. doi:10.1038/s41746-024-01029-4

23. Raza S, Bamgbose O, Ghuge S, Tavakoli F, Reji DJ. Developing Safe and Responsible Large Language Models -- A Comprehensive Framework. *arXiv*. Preprint posted online April 1, 2024. doi:10.48550/arXiv.2404.01399

24. Ainslie J, Lee-Thorp J, de Jong M, Zemlyanskiy Y, Lebron F, Sanghai S. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In: Bouamor H, Pino J, Bali K, eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2023:4895-4901. doi:10.18653/v1/2023.emnlp-main.298

25. Zheng Z, Wang Y, Huang Y, et al. Attention heads of large language models. *Patterns*. 2025;6(2). doi:10.1016/j.patter.2025.101176

26. Xiong R, Yang Y, He D, et al. On layer normalization in the transformer architecture. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol 119. ICML'20. JMLR.org; 2020:10524-10533.

27. Team G, Kamath A, Ferret J, et al. Gemma 3 Technical Report. *arXiv*. Preprint posted online March 25, 2025. doi:10.48550/arXiv.2503.19786

28. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*. 2024;6(6):e428-e432. doi:10.1016/S2589-7500(24)00061-X