Comparing Clinical and General LLMs on Knowledge Boundaries and Robustness

Xingmeng Zhao
University of Texas at San Antonio
xingmeng.zhao@utsa.edu

Ke Yang
University of Texas at San Antonio
ke.yang@utsa.edu

Anthony Rios
University of Texas at San Antonio
anthony.rios@utsa.edu

Abstract

Large language models (LLMs) often know the correct answer internally even when their expressed output is wrong, which raises questions about how this knowledge is represented and whether domain adaptation changes it. We study how continued pretraining on domain corpora affects what a model knows and how reliably it can use this knowledge, with a focus on biomedical data. Comparing a general-purpose LLM with a clinical LLM obtained through continued pretraining on clinical text, we find that both retain similar levels of probe-accessible factual knowledge, yet the stability of self-monitoring signals is substantially reduced after domain pretraining. For example, the variance of error-detection performance nearly doubles in the biomedical model. An analysis of embedding geometry suggests that this reduced stability is associated with representations becoming more isotropic, with anisotropy decreasing from about 0.47 to 0.37. These results indicate that continued domain pretraining tends to reorganize rather than expand what the model knows, and can unintentionally weaken the consistency of error-detection signals, with implications for building reliable domain-adapted LLMs[1].

1 Introduction

Large language models (LLMs) have rapidly advanced the state of the art across a wide range of tasks. However, their tendency to produce hallucinations, meaning plausible yet factually incorrect statements, raises serious concerns about reliability, particularly in high-stakes domains such as medicine. Recent evidence indicates that exposing an LLM to new factual information during supervised fine-tuning can increase hallucination rates and slow learning: examples containing new knowledge are learned more slowly than those aligned with existing knowledge, and once acquired, hallucination rates grow roughly linearly with the proportion of novel facts in the fine-tuning data [1]. In addition, head-to-head evaluations on medical question answering show that domain-adapted LLMs rarely outperform general-purpose models; medical models win only about 12% of comparisons and are significantly worse in more than one third [2]. Taken together, these findings suggest that most factual knowledge is acquired during pre-training and that naive fine-tuning on biomedical corpora may, in fact, degrade factuality [1].

To mitigate hallucinations, a complementary line of work examines a model's hidden activations to understand what it knows [3, 4]. Early results show that a simple classifier trained on hidden states

[1] Code will be released at https://github.com/Zephyr1022/knowledge-boundaries

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

can distinguish true from false statements with 71–83% accuracy, outperforming probability-based heuristics [5]. More recent work on three-digit addition finds that lightweight probes can decode both the model's predicted answer and the correct answer from hidden states, even when the output is wrong; these probes can also predict whether the output will be correct with over 90% accuracy and can guide selective re-prompting [6]. Visualizations of hidden representations on true/false datasets reveal clear linear structure, and simple difference-in-means probes generalize across datasets while causally influencing the model's answer [7]. However, prior work mainly studies general-domain models on arithmetic or binary truth settings, and it remains unclear whether similar internal signals exist in clinical LLMs or for long-tail biomedical knowledge.

Two factors make the biomedical setting particularly challenging. First, biomedical knowledge is long-tailed: many facts appear in only a handful of documents. A model's ability to answer fact-based biomedical questions correlates strongly with how many pre-training documents mention the subject and object [8]. Even after applying knowledge-editing methods, performance on long-tail biomedical facts remains substantially lower than on high-frequency ones, partly because biomedical triples often follow one-to-many relations [9]. Second, hallucination detection is under-explored in medicine. The MedHallu benchmark shows that state-of-the-art models struggle to identify hallucinated answers in PubMedQA; even GPT-4o reaches only 0.625 F1 on the hardest category [10]. Moreover, hallucinations that are semantically close to the ground truth are the most difficult to detect [10], and general-purpose models outperform fine-tuned medical models on this task. These observations suggest that domain-specific fine-tuning may reorganize internal representations in ways that reduce the robustness of self-monitoring.

We present a lightweight and unified framework for biomedical knowledge probing and error detection. Given a factual triple template $T(s, r)$ with subject s and relation r, we first perform *external probing* by prompting the LLM to generate the corresponding object o and recording its top-k predictions. We then perform *internal probing* on residual-stream hidden states at the final subject token to decode (a) the object the model *believes* is correct and (b) a proxy for the ground-truth object. Building on prior work showing that simple probes can decode answers and detect errors from hidden states [5, 6, 11], we train logistic and MLP probes to determine whether the model's answer is correct. We evaluate this framework on both base and clinical LLMs (Mistral-7B, Llama-3, BioMistral) [12, 13, 14, 15], with a focus on long-tail biomedical triples where subject–object co-occurrences are infrequent [8, 9]. Finally, our probing approach contributes to mechanistic interpretability: analyzing linear directions in hidden states provides insight into how LLMs represent biomedical relations [7, 5, 16].

Motivated by these gaps, we ask: *Do clinical (domain-adapted) LLMs differ from general LLMs in how they internally represent and monitor factual knowledge?* We adapt the probing framework for arithmetic reasoning [6] to biomedical knowledge triples $\langle s, r, o\rangle$ and make three contributions:

1. **Cross-domain probing.** We develop simple circular, logistic, and MLP probes that decode both the model's predicted object and a proxy for the ground-truth object from hidden states at each layer. We find that, while both general and clinical models *know* latent biomedical facts internally, clinical models show much higher variance in error-detection signals across layers. This matters because a model cannot reliably detect or correct its own mistakes if the self-monitoring signal shifts from one layer to the next.
2. **Error detection and geometry.** We train lightweight classifiers to detect mismatches between the model's predicted and ground-truth objects, and analyze the geometry of hidden activations. General models exhibit more anisotropic and structured representations that support stable error-detection signals, whereas clinical models are more isotropic, making self-monitoring noisier and less reliable.
3. **Implications for safety.** We show that continued domain pretraining can reorganize internal representations in ways that weaken a model's ability to self-monitor. Lightweight probes provide a low-overhead tool for real-time error detection and highlight potential risks when deploying domain-adapted LLMs in clinical settings.

2 Related Work

Long-Tail Biomedical Knowledge. Biomedical knowledge follows a long-tailed distribution: many subject–object pairs occur in only a handful of training documents. An LLM's ability to answer fact-based questions depends strongly on how often the subject and object appeared during pretraining [8]. Kandpal et al. [8] show that even very large models struggle with rare biomedical facts, requiring orders of magnitude more parameters to match performance on well-supported facts. While knowledge editing can inject missing facts, Yi et al. [9] find that edited models still perform substantially worse on long-tail biomedical triples than on high-frequency ones, in part because biomedical relations are often one-to-many, which limits edit generalization. These challenges motivate us to stratify our probing results by fact popularity to examine whether internal knowledge signals and error-detection cues differ between common and rare biomedical facts.

Mechanistic Interpretability. Mechanistic interpretability (MI) seeks to uncover the circuits, features, and directions inside neural networks that implement high-level behaviors. Recent work shows that truth-related signals are encoded in hidden activations and can be isolated with simple linear probes [7]; linear classifiers trained on hidden states can detect when a model is lying [5], and relational concept directions can be identified and used to causally steer model outputs [17]. We build on these insights by applying MI-inspired probing to domain-adapted biomedical models and analyzing how continued pretraining reshapes the geometry of internal representations.

Error Detection in LLMs. Reliable hallucination detection is critical for trustworthy LLM deployment. Azaria & Mitchell [5] showed that classifiers trained on hidden activations can distinguish true from false statements and provide more reliable confidence estimates than softmax probabilities. While many subsequent works explore probability- or consistency-based hallucination detectors, performance remains limited in the medical domain: on MedHallu, even GPT-4o achieves only moderate F1 scores, and hallucinations that are semantically close to the truth are the most difficult to detect [10]. We build on this line of work by comparing logistic, circular, and MLP probes for binary error detection and analyzing how their performance varies across layers and between general-purpose and clinical LLMs.

3 Method

We study biomedical triples $\langle s, r, o \rangle$ by prompting the model with a simple template $T(s, r)$ and extracting the residual–stream vector at layer l for the final subject token, $\mathbf{x}_l \in \mathbb{R}^d$. On this same representation $\mathbf{x}_l$, we train two decoders: an *internal* decoder that predicts the ground-truth object o (what the model "knows"), and an *external* decoder that predicts the model's own output $f_\theta(s, r)$ (what the model will "say"), following the probing framework for arithmetic reasoning of Sun, Stolfo & Sachan [6]. Because prior work shows that probe expressiveness influences what aspects of knowledge can be recovered [6], we use three complementary probe types: circular (captures low-dimensional geometric structure), logistic (linear and interpretable), and MLP (non-linear) to test whether our findings hold across probe capacities. We then convert the outputs of these decoders into a lightweight correctness score.

Probing Internal and External Knowledge. Both decoders share the same architecture and differ only in their training targets. We consider K candidate objects and use three lightweight probes applied to $\mathbf{x}_l$: (i) *circular*, which projects onto $(\mathbf{w}_1, \mathbf{w}_2)$, computes an angle $\theta = atan2(\mathbf{w}_2^\top \mathbf{x}_l,\, \mathbf{w}_1^\top \mathbf{x}_l)$, and predicts $\hat{k} = \lfloor (\theta / 2\pi)\, K \rfloor$; (ii) *logistic*, which computes logits $\mathbf{z} = \mathbf{W}\mathbf{x}_l + \mathbf{b}$ and predicts $\hat{k} = \arg\max_i z_i$; and (iii) *MLP*, which computes $\mathbf{h} = \mathrm{ReLU}(\mathbf{W}_1 \mathbf{x}_l + \mathbf{b}_1)$, then $\mathbf{z} = \mathbf{W}_2 \mathbf{h} + \mathbf{b}_2$, and predicts $\hat{k} = \arg\max_i z_i$. We report *internal accuracy* for decoders trained to recover o and *external accuracy* for decoders trained to recover $f_\theta(s, r)$.

Probing for Error Detection. To detect when a model produces an incorrect biomedical fact, we apply three lightweight probes to the hidden activations: (i) a *Logistic* probe that provides a simple and interpretable linear classifier for detecting errors, (ii) an *MLP* probe that captures non-linear patterns in how correct and incorrect answers are represented, and (iii) a *Joint Circular Error Detector* that compares the model's internal belief and expressed answer using angular representations. The circular detector maps hidden activations to an angle $\theta = \mathrm{atan2}(w_1^\top x,\, w_2^\top x)$ and flags a triple as incorrect when the angular distance between the internal and external predictions exceeds a learned threshold. This yields an intuitive geometric signal for identifying errors, while keeping the probes lightweight and easy to interpret.

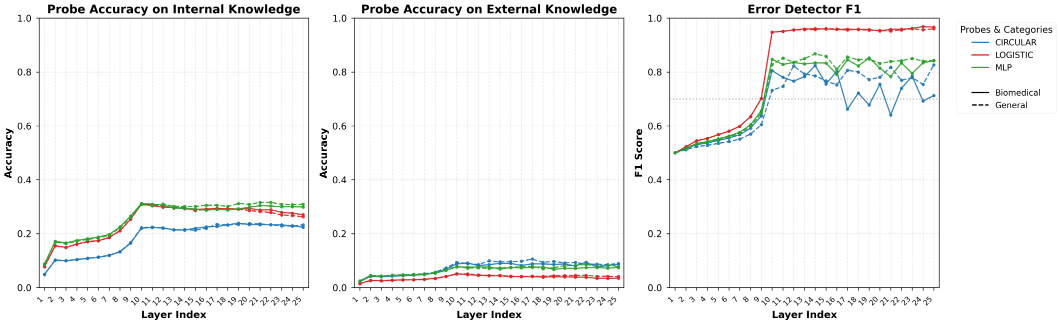

Figure 1: Layer-wise probe performance on general models (solid) and biomedical models (dashed). Left, internal $\mathrm{Acc@10}$. Center, external $\mathrm{Acc@10}$. Right, error-detector F1. Biomedical models exhibit notably higher variability in error detection at later layers despite similar average recall.

4 Experiment

We evaluate knowledge and error-detection signals in six large language models, consisting of three general-purpose models and their corresponding biomedical adaptations, across three datasets.

Language Models. We compare three open-source base models, Gemma-2-9B [18], Llama-3-8B [13], and Mistral-7B-Instruct-v0.1 [12], with their biomedical counterparts: Meditron3-Gemma2-9B [19], MMed-Llama-3-8B [20], and BioMistral-7B [14]. This pairing isolates the effect of domain-specific medical adaptation, which has been reported to not reliably improve factual recall or reduce hallucinations [2, 1].

Datasets and Prompts. We use three relation-based biomedical datasets. The first is MedLAMA, which contains UMLS triples from Meng et al. [21] for probing factual knowledge; we mark triples with fewer than ten PubMed co-occurrences as long-tail, following Kandpal et al. [8] and Yi et al. [9]. The second is a drug–symptom dataset from Berkowitz et al. [22] with 165 positive controls and 234 negative controls for assessing pharmacovigilance. The third consists of drug–drug interaction triples extracted from SemMedDB [23]. For each relation (e.g., "may treat"), we design a simple natural-language template such as "`[X] might treat [Y]`" to prompt the models; the full set of templates is provided in Appendix 1.

Evaluation Metrics. We assess how well a model can recall the correct object using Accuracy@K (Acc@1, Acc@5, Acc@10), which measures whether the correct answer appears within the model's top-K predictions. For example, Acc@10 counts a prediction as correct if the gold object o is included anywhere among the top-10 generated candidates for a triple $\langle s, r, o\rangle$. For error detection, we label a case as *correct* if the model's top-1 prediction matches o and *incorrect* otherwise, train lightweight probes on hidden states to predict this label, and report F1 scores. Following Meng et al. [21], we use "not appearing in the top-10" as the operational definition of an incorrect output.

To understand *where* knowledge becomes accessible inside the model, we apply probes across model layers: for layers $L \in \{1, \ldots, 25\}$, we extract residual-stream activations at the final subject token, train probes independently at each layer, and identify the depth with the best validation performance. Finally, we measure *retrieval failures*, cases where the model internally encodes the correct object but does not express it in the top-10 outputs, by training a probe to detect such failures and reporting F1 scores.

Experiment Results. Figure 1 presents layer-wise probe performance for general and biomedical models. Internal and external $\mathrm{Acc@10}$ remain similarly low across both model types, indicating that continued domain pretraining does not add or remove probe-accessible knowledge. The key difference is *stability*: biomedical models show markedly higher layer-to-layer fluctuation in error-detection F1, especially in later layers (Table 2; e.g., circular probes SD $= 0.053$ vs. 0.027 for the corresponding general model). This matters because reliable self-monitoring requires a consistent signal over the forward pass; if the signal fluctuates with depth, downstream confidence-estimation or correction modules cannot rely on it.

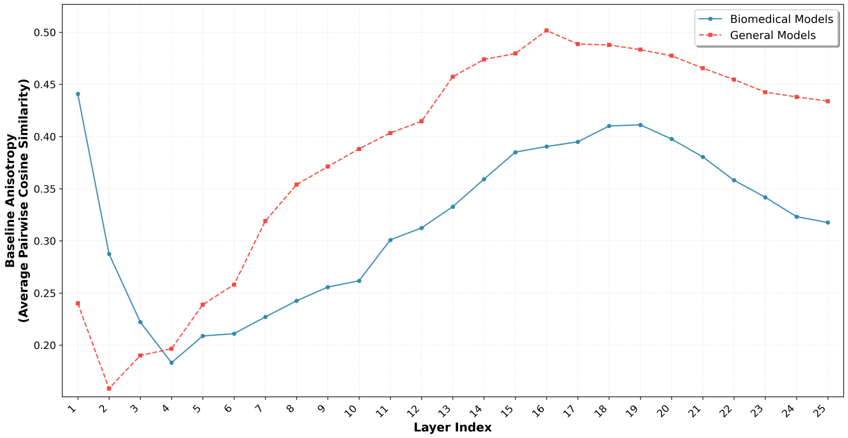

Figure 2: Baseline anisotropy across layers. General LLMs (red) show higher anisotropy. Clinical LLMs (blue) are more isotropic. Lower anisotropy coincides with more volatile error-detection.

To examine what may underlie this instability, we analyze representation geometry. Following Ethayarajh [24], we estimate *anisotropy* by measuring the average cosine similarity between random directions and layer-wise representations. As shown in Figure 2, general models are more anisotropic (peak ≈ 0.61) than their biomedical counterparts (peak ≈ 0.35). Prior work has argued that greater anisotropy reflects representations that are concentrated along a small number of dominant directions, which can make certain features easier to separate with simple linear readouts. In contrast, more isotropic representations are more uniformly spread, which can dilute these directions and make small differences harder to linearly decode.

Here, we observe that layers with lower anisotropy coincide with higher variability in probe performance (Figure 1, right). While this does not, by itself, establish causation, it suggests a plausible link: more isotropic representations may provide weaker or less stable linear signals for error detection, contributing to the observed volatility across layers. Taken together, these results support the view that continued domain pretraining tends to *reorganize*, rather than expand, probe-accessible knowledge, and that this reorganization can make internal error-detection signals less consistent. This aligns with recent findings that some domain-adaptation regimes can inadvertently increase hallucination rates.

5 Limitations and Conclusion

Our results show that general-purpose and biomedical LLMs contain similar amounts of probe-accessible knowledge, but their internal error-detection signals differ sharply in stability: across circular, logistic, and MLP probes, the biomedical model has nearly twice the across-layer standard deviation in error-detector F1 compared to the general model (Table 2). An isotropy analysis indicates a plausible mechanism: domain-specific fine-tuning reduces anisotropy, yielding more isotropic representations that are less amenable to simple linear separation of correct vs. incorrect states. Practically, this implies that fine-tuning reorganizes internal geometry without adding new knowledge and can make self-monitoring less reliable, consistent with reports that instruction tuning increases hallucination rates [1, 2]. As a simple diagnostic, we propose measuring the variability of internal error-detection signals to assess robustness. More broadly, our unified probing and error-detection framework links mechanistic interpretability to safety: by extracting both predicted and ground-truth objects from hidden states, lightweight probes act as oracles that anticipate hallucinations and, when plugged into re-prompting or abstention pipelines, improve reliability with minimal overhead. Although we focus on biomedical facts, the approach generalizes to other domains with long-tail knowledge and high-stakes decisions. Future work should study cross-model prediction [25], explore unsupervised clustering for richer error signals, and test whether stronger probes or representation editing can mitigate the variability introduced by domain-specific fine-tuning.

References

[1] Z. Gekhman, G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart & J. Herzig. “Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?” In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*(. 2024), pp. 7765–7784.

[2] D. Jeong, S. Garg, Z. C. Lipton & M. Oberst. “Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?” In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*(. 2024), pp. 12143–12170.

[3] Z. Gekhman, E. B. David, H. Orgad, E. Ofek, Y. Belinkov, I. Szpektor, J. Herzig & R. Reichart. “Inside-out: Hidden factual knowledge in llms”. In: *arXiv preprint arXiv:2503.15299* ((2025)).

[4] H. Orgad, M. Toker, Z. Gekhman, R. Reichart, I. Szpektor, H. Kotek & Y. Belinkov. “LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations”. In: *The Thirteenth International Conference on Learning Representations*.

[5] A. Azaria & T. Mitchell. “The Internal State of an LLM Knows When It’s Lying”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*(. 2023), pp. 967–976.

[6] Y. Sun, A. Stolfo & M. Sachan. “Probing for Arithmetic Errors in Language Models”. In: *arXiv preprint arXiv:2507.12379* ((2025)).

[7] S. Marks & M. Tegmark. “The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets”. In: *First Conference on Language Modeling*.

[8] N. Kandpal, H. Deng, A. Roberts, E. Wallace & C. Raffel. “Large language models struggle to learn long-tail knowledge”. In: *Proceedings of the 40th International Conference on Machine Learning*(. 2023), pp. 15696–15707.

[9] X. Yi, J. Lever, K. Bryson & Z. Meng. “Can We Edit LLMs for Long-Tail Biomedical Knowledge?” In: *arXiv preprint arXiv:2504.10421* ((2025)).

[10] S. Pandit, J. Xu, J. Hong, Z. Wang, T. Chen, K. Xu & Y. Ding. “Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models”. In: *arXiv preprint arXiv:2502.14302* ((2025)).

[11] A. Shrivastava & A. Holtzman. “Linearly Decoding Refused Knowledge in Aligned Language Models”. In: *arXiv preprint arXiv:2507.00239* ((2025)).

[12] A. Q. Jiang et al. *Mistral 7B*(. 2023). arXiv: `2310.06825 [cs.CL]`. URL: `https://arxiv.org/abs/2310.06825`.

[13] Meta AI. *Introducing Meta Llama 3: The most capable openly available LLM to date*. `https://ai.meta.com/blog/meta-llama-3/`. Accessed: 2025-08-21(. 2024).

[14] Y. Labrak, A. Bazoge, E. Morin, P.-a. Gourraud, M. Rouvier & R. Dufour. “BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains”. In: *62th Annual Meeting of the Association for Computational Linguistics (ACL’24)*(. 2024).

[15] M. S. Ankit Pal. *OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences*. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`(. 2024).

[16] M. Beigi, Y. Shen, R. Yang, Z. Lin, Q. Wang, A. Mohan, J. He, M. Jin, C.-T. Lu & L. Huang. “InternalInspector I2: Robust Confidence Estimation in LLMs through Internal States”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*(. 2024), pp. 12847–12865.

[17] D. Chanin, A. Hunter & O.-M. Camburu. “Identifying Linear Relational Concepts in Large Language Models”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*(. 2024), pp. 1524–1535.

[18] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. “Gemma 2: Improving open language models at a practical size”. In: *arXiv preprint arXiv:2408.00118* ((2024)).

[19] OpenMeditron Initiative. *Meditron-3 (Gemma2-9B): A Clinical Medicine–Specialized LLM*. `https://huggingface.co/OpenMeditron/Meditron3-Gemma2-9B`. Model Card; fine-tuned from google/gemma-2-9b(. 2025).

[20] P. Qiu, C. Wu, X. Zhang, W. Lin, H. Wang, Y. Zhang, Y. Wang & W. Xie. *Towards Building Multilingual Language Model for Medicine*(. 2024). arXiv: `2402.13963 [cs.CL]`.

Relation	**Manual Prompt**
may treat	[X] might treat [Y].
may prevent	[X] might prevent [Y].
adverse drug effect	The adverse effect of [X] is [Y].
interacts with	The [X] interacts with [Y].

Table 1: Templates $T(s, r)$ used to elicit $\mathbf{x}_l$. Here [X] is the subject s and [Y] is the object o.

Probe	Metric	Bio s.d.	Gen s.d.	Ratio
Circular	Acc@10	0.00730	0.00796	0.917
Circular	External_Acc@10	0.00441	0.00553	0.797
Circular	F1 (err. det.)	0.05287	0.02725	**1.940**
Logistic	Acc@10	0.00959	0.01311	0.732
Logistic	External_Acc@10	0.00460	0.00295	1.561
Logistic	F1 (err. det.)	0.00504	0.00353	**1.427**
MLP	Acc@10	0.00727	0.00477	1.526
MLP	External_Acc@10	0.00268	0.00473	0.567
MLP	F1 (err. det.)	0.01994	0.01292	**1.543**

Table 2: Across-layer standard deviations (s.d.) of probe metrics. Bio = biomedical model; Gen = general model. Larger s.d. indicates less stability across layers.

[21] Z. Meng, F. Liu, E. Shareghi, Y. Su, C. Collins & N. Collier. “Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*(. 2022), pp. 4798–4810.

[22] J. Berkowitz, D. Weissenbacher, A. Srinivasan, N. A. Friedrich, J. M. A. Cortina, S. Kivelson, G. G. Hernandez & N. P. Tatonetti. “Probing Large Language Model Hidden States for Adverse Drug Reaction Knowledge”. In: *International Conference on Artificial Intelligence in Medicine*. Springer(. 2025), pp. 55–64.

[23] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat & T. C. Rindflesch. “SemMedDB: a PubMed-scale repository of biomedical semantic predications”. In: *Bioinformatics* 28.23 ((2012)), pp. 3158–3160.

[24] K. Ethayarajh. “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”. 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)*(. 2019), pp. 55–65.

[25] J. Gallifant, S. Chen, K. Sasse, H. Aerts, T. Hartvigsen & D. S. Bitterman. “Sparse autoencoder features for classifications and transferability”. In: *arXiv preprint arXiv:2502.11367* ((2025)).

A Experiments Compute Resources

All experiments were conducted on 4 NVIDIA H100 GPUs. Probe training and evaluation completed within approximately one day of wall-clock time.

B Prompt Templates and Layerwise Stability

This appendix documents (i) the minimal natural-language templates we use to render (s, r, o) tuples into prompts for extracting layer representations $\mathbf{x}\ell$, and (ii) the layerwise stability of three probe families. In Table 1, `[X]` is the subject s and `[Y]` is the object o; instantiating $T(s, r)$ yields the input from which we read $\mathbf{x}\ell$. Table 2 reports the standard deviation across layers for each metric; larger values indicate less stable signals over depth. Notably, error-detection F1 varies substantially more in the biomedical model.

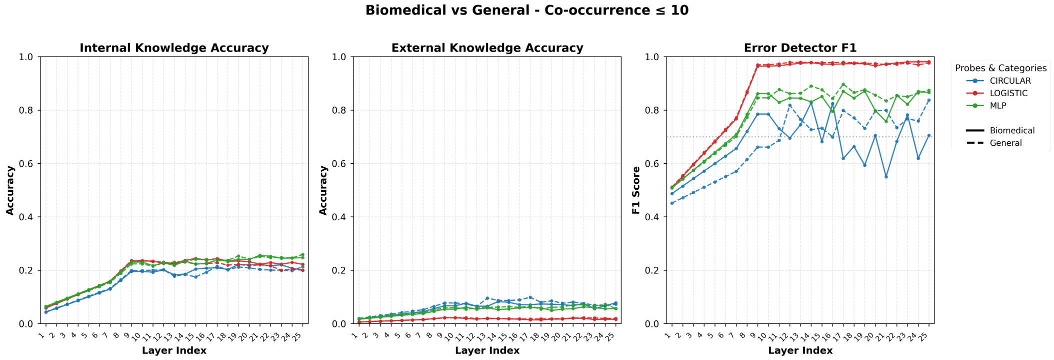

Figure 3: Layer-wise probe performance for low-frequency pairs (**co-occurrence** ≤ 10). Error detection is strong in deeper layers, with notably higher variability for the biomedical model (solid) relative to the general model (dashed).

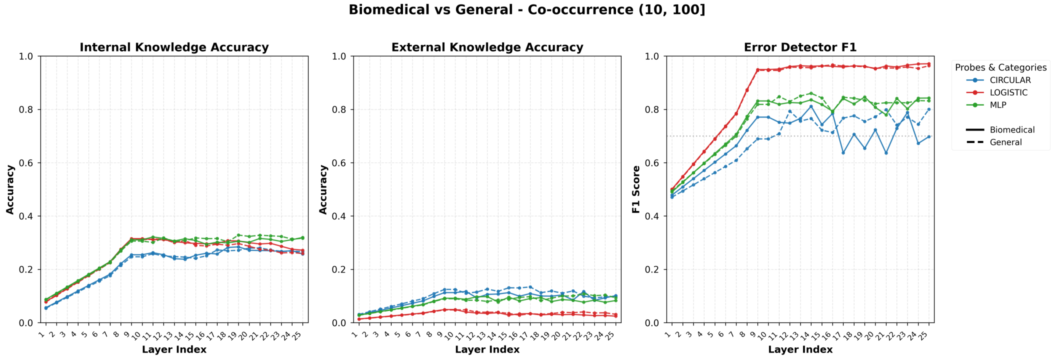

Figure 4: Layer-wise probe performance for **(10, 100]** co-occurrence. The stability gap persists: biomedical error-detection signals vary more across layers despite similar average recall.

C Additional Layer-wise Results by Co-occurrence

We stratify (s, o) pairs by corpus co-occurrence into four buckets (low to high frequency): ≤ 10, $(10, 100]$, $(100, 1000]$, and > 1000. For each bucket we plot layer-wise probe accuracy on internal and external knowledge (left/center) and error-detector F1 (right) for circular, logistic, and MLP probes. Across buckets, knowledge readout remains low while error detection strengthens with depth; the biomedical model shows consistently higher across-layer variability, especially in later layers.

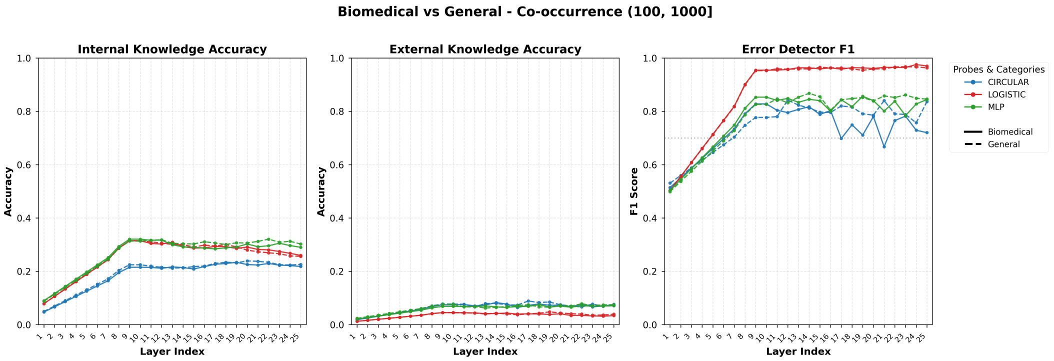

Figure 5: Layer-wise probe performance for **(100, 1000]** co-occurrence. General model (dashed) remains more stable across depth; biomedical model (solid) shows noisier F1 trajectories.

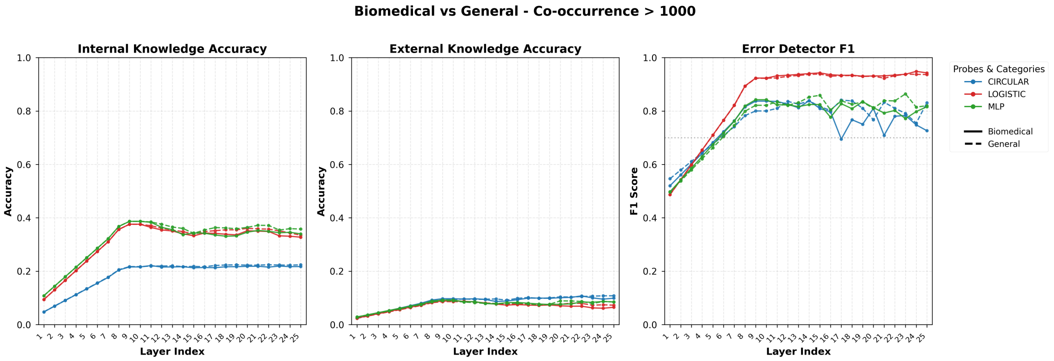

Figure 6: Layer-wise probe performance for high-frequency pairs (**co-occurrence** > 1000). Trends hold with slightly smoother curves; biomedical error-detection variability remains elevated at later layers.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that domain-specific finetuning reorganizes internal representations without adding or removing probe-accessible knowledge, and that this reduces stability of error detection. These claims match the experiments and conclusions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section "Limitations and Conclusion" explicitly discusses that findings are limited to biomedical triples and that variability may arise from finetuning regimes. It also notes the scope (long-tail biomedical knowledge) and proposes future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include formal theoretical results or proofs; it is an empirical study.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides details on models, datasets, prompt templates, probing methods, and evaluation metrics (Sections 2–3, Appendix). These are sufficient to reproduce the main findings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets are public (MedLAMA, PubMed-derived sets, SemMedDB). In addition, the code will be released on GitHub with instructions and setup details to enable faithful reproduction of the main experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 and Appendix specify datasets, relation templates, model variants, probing methods, and evaluation metrics, enabling understanding of the setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results report across-layer variability using standard deviations (Table 2) and compare probe families. Error bars/variance are included where relevant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were conducted on 4 NVIDIA H100 GPUs. Probe training and evaluation required approximately one day of wall-clock time.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. It uses only publicly available biomedical datasets and models without involving sensitive human subjects data.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The introduction and conclusion discuss risks of deploying domain-specific LLMs in medical settings, noting reduced robustness in self-monitoring. Positive impact includes lightweight tools for hallucination detection.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No new high-risk models or datasets are released; the work uses existing open-source models and benchmarks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used assets (Llama, Mistral, BioMistral, MedLAMA, SemMedDB, PubMedQA) are publicly licensed and properly cited in the bibliography.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce new datasets, models, or code artifacts.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The study does not involve crowdsourcing or human subject experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subject research was conducted; all datasets are from existing biomedical corpora.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The core methodology involves probing the hidden states of general and biomedical LLMs (Mistral, Llama, Gemma, BioMistral, Meditron, MMed-Llama), which is clearly described throughout the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.