Probing and Steering Uncertainty in Biomedical Language Models: Representational Structure and Behavioral Limits

Debmalya Pal
University of California, San Diego
Department of Computer Science and Engineering
d2pal@ucsd.edu

Abstract

Biomedical language models can generate overly confident clinical statements despite incomplete or ambiguous evidence. We study whether linguistic uncertainty (the hedged epistemic stance expressed in phrases such as "consistent with" or "cannot exclude") is encoded in model representations and can be controlled without retraining. Across six biomedical language models spanning two architectures (causal decoders and bidirectional encoders), we show that uncertainty is captured by robust low-dimensional linear structure in hidden states. We then apply activation steering to manipulate this representation directly, increasing hedged generation in decoder models and inducing targeted uncertainty-related shifts in encoder representations. Together, these results show that epistemic stance is not merely a surface linguistic phenomenon but an interpretable and controllable feature of biomedical language model representations, with implications for safer and more calibrated clinical text generation.

1 Introduction

Biomedical language models are increasingly proposed as components of clinical decision support systems, from radiology report generation to literature-assisted diagnosis. Yet a gap persists between benchmark performance and clinical deployability: language models trained on biomedical text tend to produce assertions in a confident declarative register, even when the clinical evidence is incomplete. Prior work has documented this tendency in the factual domain (Singhal et al., 2023; Umapathi et al., 2023) and in models' failure to express calibrated uncertainty in their outputs (Xiong et al., 2024); the corresponding problem for *linguistic hedging style* in clinical text is the focus of this work.

Throughout this work, we use *uncertainty* to refer to the *linguistic expression* of epistemic hedging in clinical text (the use of modal constructions and speculative cues that mark epistemic stance toward a proposition), rather than to a model's confidence calibration over factual claims (Geng et al., 2024). Expressing this uncertainty ("findings are consistent with," "cannot exclude") is a learned professional skill with real safety consequences (Panicek and Hricak, 2016): confident model outputs invite action without confirmation, while hedged outputs prompt consideration of alternatives.

Existing behavioral control methods operate either at the input level (prompting, brittle (Zhao et al., 2021; Lu et al., 2022)) or the parameter level (fine-tuning, expensive). A third approach from mechanistic interpretability, *activation steering* (Turner et al., 2023; Zou et al., 2023; Panickssery et al., 2024), intervenes directly on hidden states at inference time along directions encoding target concepts, requiring no retraining. Whether this prospect is realistic for clinical uncertainty in biomedical models is an open empirical question.

Two conditions must hold: uncertainty must be linearly encoded in hidden states, and steering along this direction must shift outputs without collapsing other generation properties. We investigate both systematically across six biomedical language models spanning two architectural families (causal decoders and bidirectional encoders), parameter scales ranging from 110M to 2.7B, and five pretraining corpora. This multi-model scope is deliberate: a finding that holds only for a single checkpoint could reflect checkpoint-specific memorization of speculation cues; establishing it across architectures, scales, and corpora is evidence that linear uncertainty encoding is a property of biomedical LM training rather than an artifact. Our five contributions are:

Code: github.com/DebmalyaPal/Probing-and-Steering-Uncertainty-in-Biomedical-Language-Models

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1066–1079

- We show uncertainty is linearly decodable across all six models (80–87% 5-fold CV accuracy), establishing it as a general property of biomedical LMs rather than a model-specific artifact.
- We document a 15-fold Cohen's d improvement for mean vs. last-token pooling in BioGPT (1.33 vs. 0.09), with implications for probing biomedical LMs broadly.
- We introduce a principled two-stage layer selection (2/3-depth anchor ± 2 window) validated empirically: gains range from $+0.25$pp to $+2.50$pp, with BioBERT showing the most striking improvement (L10 over L8 anchor).
- We show that dynamic hidden-norm calibration prevents generation quality collapse across the full tested steering range for decoder models, and characterize qualitatively distinct steering regimes for BioGPT and BioMedLM arising from differences in scale and hidden dimension.
- We introduce a five-metric encoder representation-steering suite (projection shift, probability shift, flip rate, centroid cosine shift, on-axis fraction) and demonstrate consistent, architecturally uniform shifts across all four encoder models despite distinct pretraining corpora.

2 Related Work

Uncertainty in biomedical text. The BioScope corpus (Vincze et al., 2008) provides sentence-level speculation annotations used as ground truth in our work. Prior classification work (Agarwal and Yu, 2010; Adel and Schütze, 2015) treated hedging as a surface-level feature; we ask whether it constitutes an *internal representational property* of biomedical LMs. Recent clinical LLM calibration work (Umapathi et al., 2023; Singhal et al., 2023) focuses on factual accuracy rather than the expression of epistemic stance, leaving our question largely unaddressed.

Probing representations. Probing (training classifiers on frozen hidden states) has established a layered hierarchy of linguistic information in LMs (Hewitt and Manning, 2019; Tenney et al., 2019; Belinkov, 2022). In biomedical models, probing has focused on domain knowledge (Sung et al., 2021; Vulić et al., 2020); pragmatic and stance-like properties are understudied. A key caveat is that decodability does not imply task relevance (Ravichander et al., 2021), a distinction directly relevant to our steering results.

Activation steering. The linear representation hypothesis (Park et al., 2024) underlies activation addition (Turner et al., 2023), representation engineering (Zou et al., 2023), and contrastive activation addition (Panickssery et al., 2024). Methodological concerns include surface-feature confounds (Elazar et al., 2021) and out-of-distribution degradation at large perturbation magnitudes (Zou et al., 2023). Application to biomedical models is, to our knowledge, unexplored. We address this gap across six models with two architectural families and full multi-metric evaluation.

Multi-model comparative probing. Most probing studies target a single model or checkpoint (Tenney et al., 2019; Hewitt and Manning, 2019); systematic comparison across architectures remains less common. Notable exceptions include Vulić et al. (2020), who compare BERT variants on lexical knowledge, and Belinkov (2022), who survey cross-model probing methodology. Our study extends this comparative lens to a pragmatic property (epistemic stance) across two architectural families with different training objectives, examining both where uncertainty representations form and whether they support controlled steering.

3 Methods

The methodology has two stages: (i) train linear probes on frozen hidden states to identify the layer and direction $\mathbf{w}_{\hat{\ell}}$ that best separates uncertain from certain sentences (§§3.3–3.4); (ii) inject a scaled multiple of that direction at inference time via a forward hook, with no retraining (§§3.5–3.6).

3.1 Dataset

We use the BioScope corpus (Vincze et al., 2008), which provides token-level scope annotations for speculation and negation. We collapse these to sentence-level binary labels: sentences containing any speculation cue are treated as *uncertain*, all others as *certain*. This simplification discards scope information but yields a clean binary contrast set aligned with our probing objective. Parsing all three sub-corpora (abstracts, full papers, clinical reports) yields 3,065 uncertain and 13,007 certain sentences. After restricting to sentences of 30–200 characters, we sample 200 of each class (seed 42) to form a balanced 400-sentence contrast set $\mathcal{D}$. Files are enumerated alphabetically before sampling for cross-platform reproducibility. Despite length filtering, a small residual difference remains

Table 1: Models evaluated (float32, no quantization).

Model	Arch.	L	Params	Pretrain
BioGPT	Dec.	24	345M	PubMed
BioMedLM	Dec.	32	2.7B	PubMed
BioBERT	Enc.	12	110M	PubMed+PMC
ClinicalBERT	Enc.	12	110M	MIMIC-III
BlueBERT	Enc.	12	110M	PubMed+MIMIC
SciBERT	Enc.	12	110M	Broad sci.

(27.1 vs. 24.6 tokens on average), which we later control via orthogonalization.

3.2 Models and representation extraction

We study six biomedical LMs across two families (Table 1): two causal decoders, BioGPT (Luo et al., 2022) and BioMedLM (Bolton et al., 2024), and four bidirectional encoders, BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), BlueBERT (Peng et al., 2019), and SciBERT (Beltagy et al., 2019). For sentence s tokenized into T tokens, we extract mean-pooled hidden states at layer ℓ:

$$\mathbf{h}_\ell(s) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{H}_\ell(s)[t,:] \tag{1}$$

We use mean pooling uniformly; Section 4.1 shows it yields Cohen's $d = 1.33$ vs. 0.09 for last-token pooling, motivating this choice.

3.3 Linear probing

For each layer ℓ, we train a logistic regression probe:

$$p(y{=}\text{unc.} \mid \mathbf{h}_\ell(s)) = \sigma\Big(\mathbf{w}_\ell^\top \mathbf{h}_\ell(s) + b_\ell\Big) \tag{2}$$

with ℓ_2-regularization $C = 0.1$. Evaluation uses stratified 5-fold CV on the 400-sentence set; we report mean accuracy, std, and 95% CI.

3.4 Principled layer selection

Stage 1: 2/3-depth anchor.

$$\ell^* = \text{round}\big(\tfrac{2N}{3}\big) \tag{3}$$

This heuristic is grounded in probing studies showing that middle-to-upper layers encode the most abstract linguistic properties (Tenney et al., 2019; Belinkov, 2022). For decoders it also balances representational maturity against propagation distance (enough remaining blocks for the injected signal to influence generation). The formula gives $\ell^* = 8$ for all 12-layer BERT variants and $\ell^* = 16$ for BioGPT; the latter coincides with the layer examined in Luo et al. (2022).

Table 2: Steering layer selection. $\ell^* = \text{round}(2N/3)$; $\hat{\ell}$ = window-best (Eq. 4). † = window differs from anchor.

Model	N	ℓ^*	Acc@ℓ^*	$\hat{\ell}$	Acc@$\hat{\ell}$	Std
BioGPT	24	16	85.50%	16	85.50%	2.44%
BioMedLM†	32	21	84.00%	22	84.25%	1.90%
BioBERT†	12	8	84.00%	10	86.50%	2.40%
ClinicalBERT†	12	8	82.00%	9	82.75%	3.79%
BlueBERT	12	8	80.75%	8	80.75%	5.35%
SciBERT†	12	8	84.75%	10	85.00%	3.42%

Stage 2: Local empirical refinement. Probe accuracy curves are locally flat across several adjacent layers. We search within ± 2 layers of the anchor, selecting by highest mean accuracy (ties broken by minimum std):

$$\hat{\ell} = \underset{\ell \in [\ell^*-2,\, \ell^*+2] \cap \{1,\ldots,N\}}{\text{argmax}} \big(\bar{a}_\ell,\ -s_\ell\big) \tag{4}$$

where $\bar{a}_\ell$ is the mean CV accuracy and s_ℓ the standard deviation at layer ℓ. The tuple $(\bar{a}_\ell, -s_\ell)$ encodes lexicographic tie-breaking: select the layer with highest mean accuracy; break ties by lowest standard deviation. Layer 0 (embedding output) is excluded as it is not a hookable transformer-block module. Table 2 and Figure 2 show that the window correctly identifies BioBERT L10 over the L8 anchor (+2.50pp); BioGPT and BlueBERT show zero gain (anchor = window-selected layer), and the remaining three models (BioMedLM, ClinicalBERT, SciBERT) gain between +0.25pp and +0.75pp.

3.5 Hidden-norm calibration and steering directions

Steering magnitude is parameterized fractionally:

$$\alpha_{\text{eff}} = \alpha_{\text{frac}} \cdot \|\mathbf{h}_{\hat{\ell}}\|_{\text{est}} \tag{5}$$

where $\|\mathbf{h}_{\hat{\ell}}\|_{\text{est}}$ is estimated dynamically from 8 generic biomedical sentences. This makes α_{frac} interpretable as a fraction of typical hidden-state magnitude across layers and models. For BioGPT L16, dynamic calibration yields ≈ 205, versus the static value 616 used in our uncalibrated baseline (derived from a single prompt without averaging), a $3\times$ discrepancy that produced perplexity collapse (> 400) at $\alpha_{\text{frac}} = 0.25$; calibrated norms prevent this entirely.

We construct two steering directions at $\hat{\ell}$, both derived from the probe in Eq. 2. For the steering direction the probe is retrained on all 400 sentences (rather than per-fold subsets) to obtain a maximally

stable direction; Finding 3 shows this recovers per-fold CV accuracy completely.

The **probe direction** is:

$$\mathbf{v}_{\text{probe}} = \mathbf{w}_{\hat{\ell}}^{\text{full}} \tag{6}$$

The **orthogonalized direction** removes two candidate confound directions from $\mathbf{v}_{\text{probe}}$: a *length direction* $\mathbf{v}_{\text{len}}$ (regression weights predicting token count) and a *hedge-count direction* $\mathbf{v}_{\text{hedge}}$ (regression weights predicting hedging-vocabulary occurrences). We define the orthogonalization operator as:

$$\text{orth}(\mathbf{u}, \mathbf{v}) = \mathbf{u} - \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{v}\|^2}\mathbf{v} \tag{7}$$

and apply it sequentially:

$$\mathbf{v}_{\text{ortho}} = \text{orth}\big(\text{orth}(\mathbf{v}_{\text{probe}},\ \mathbf{v}_{\text{len}}),\ \mathbf{v}_{\text{hedge}}\big) \tag{8}$$

The resulting direction is orthogonal to both confounds while retaining maximum alignment with the original probe direction. For BioGPT L16, $\cos(\mathbf{v}_{\text{probe}}, \mathbf{v}_{\text{ortho}}) = 0.95$, confirming that orthogonalization makes only a small angular adjustment. Note that this procedure controls for hedge-count as a scalar surface feature; it does not fully separate lexical from semantic content, since the probe direction itself is derived from annotations defined partly by the presence of hedging vocabulary.

3.6 Steering protocols

Decoder steering. We register a forward hook at $\hat{\ell}$ that adds $\alpha_{\text{eff}} \cdot \hat{\mathbf{v}}$ to all token positions during both prefill and each autoregressive decoding step. We sweep $\alpha_{\text{frac}} \in \{0, 0.025, \ldots, 0.25\}$ (9 values) for probe and ortho directions; 20 seeds $\times$ 5 radiology prompts $=$ 100 samples per configuration. BioMedLM uses a coarser step of 0.05 (6 values) due to its larger memory footprint; the monotonically increasing response gives no indication that finer resolution would alter the qualitative findings. Outputs are scored on hedge score, lexical diversity, token validity, and perplexity (see Appendix E).

Encoder steering. For encoder models (no generation), we inject $\alpha_{\text{eff}} \cdot \hat{\mathbf{v}}$ at $\hat{\ell}$ and compare modified vs. original mean-pooled hidden states across 80 held-out test sentences. We compute five metrics: projection shift Δ_{proj}, probability shift Δ_{prob}, decision-boundary flip rate, centroid cosine shift $\Delta_{\cos}$, and on-axis fraction (formal definitions in Appendix E).

Table 3: Probe CV accuracy at selected steering layer $\hat{\ell}$.

Model	$\hat{\ell}$	L0 acc	Acc@$\hat{\ell}$	Std
BioGPT	16	82.00%	85.50%	2.44%
BioMedLM	22	55.25%	84.25%	1.90%
BioBERT	10	76.00%	86.50%	2.40%
ClinicalBERT	9	77.75%	82.75%	3.79%
BlueBERT	8	77.00%	80.75%	5.35%
SciBERT	10	80.50%	85.00%	3.42%

Compute. All canonical experiments use Python 3.11.9, PyTorch 2.5.1+cu121, Transformers 4.49.0, NVIDIA GPU, CUDA 12.1, float32. bfloat16/float16 produce equivalent probe accuracies on the same CUDA device (within 0.25pp). Platform-level non-determinism (CUDA vs. MPS) can shift layer selection within the ± 2 window; CUDA results are canonical.

4 Results

4.1 Pooling method determines observable structure

At BioGPT layer 16, **last-token pooling** yields Cohen's d=0.088 (57% accuracy), while **mean pooling** yields d=1.331 (85.50% probe CV accuracy at this layer). Last-token representations are dominated by the final content token (e.g. *pneumonia*), which is informative about topic but not epistemic stance. Mean pooling surfaces distributed sentence-level properties including hedging. All subsequent analyses use mean-pooled hidden states.

4.2 Uncertainty is linearly decodable across all six models

Table 3 and Figure 1 report probe accuracy across models.

Finding 1: Robust cross-model decodability. All six models achieve 80–87% CV accuracy at the selected steering layer. The narrow spread across models with different architectures, sizes, and pretraining corpora indicates that linear encoding of epistemic stance is a general property of biomedical LMs rather than a checkpoint-specific artifact.

Finding 2: Layer-0 accuracy reveals family-level differences. BioGPT achieves 82% at layer 0; uncertainty is already largely encoded in the token embeddings through surface lexical statistics. BioMedLM achieves only 55% at layer 0 (near chance), meaning its uncertainty encoding is built entirely by transformer computation despite shar-

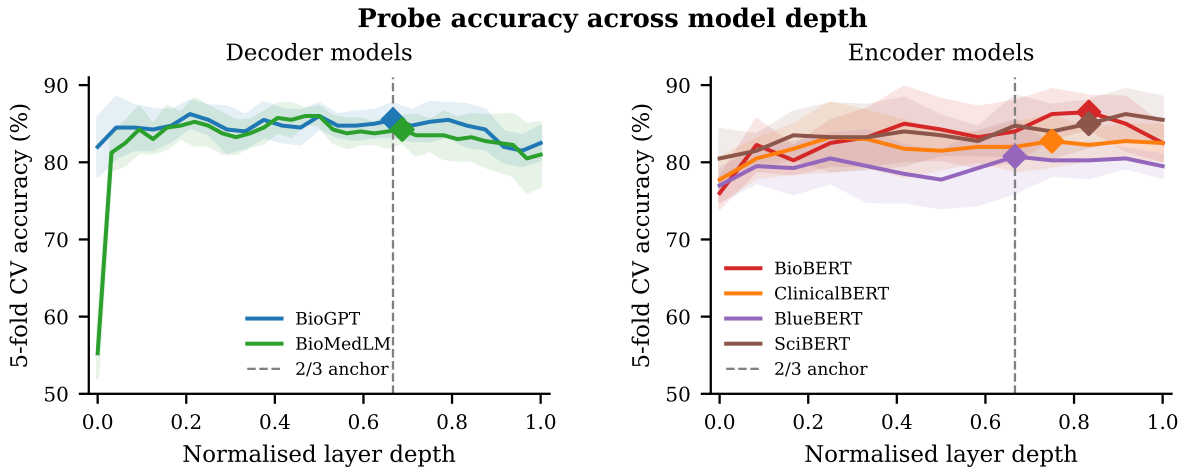

Figure 1: Probe accuracy across normalised layer depth for all models. Shaded bands are 95% CI. Diamonds mark the window-selected layer $\hat{\ell}$. Dashed line marks the 2/3-depth anchor.

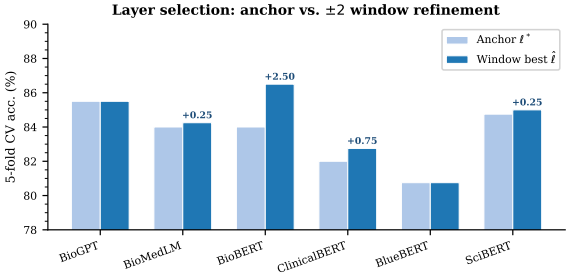

Figure 2: Anchor vs. window-best accuracy for all models. Annotations show percentage-point gain from the ±2 refinement; BioGPT and BlueBERT show zero gain (anchor = window-selected layer).

ing the same pretraining corpus as BioGPT. Encoder models occupy an intermediate position (76–81%), consistent with bidirectional attention enabling richer early contextualization even within the embedding layer's immediate context. This family-level contrast is clearly visible in Figure 1.

Finding 3: The uncertainty direction is stable and concentrated. The non-trivial question is whether the learned direction is *consistent*: does a single direction computed once on all 400 sentences, frozen across folds, recover the same accuracy as a per-fold adaptive probe free to re-orient across the full representation space? We test this by projecting onto $\mathbf{v}_{\text{probe}}$ and training only a scalar threshold per fold. For BioGPT L16, the frozen direction achieves 85.50%, matching the per-fold adaptive probe (2.44% std); encoder models recover $> 98\%$ of per-fold accuracy. This confirms that uncertainty is concentrated along a single *consistent* geometric axis. Orthogonalizing against sentence length and hedge-count reduces accuracy by < 1pp ($85.50\% \rightarrow 85.00\%$), confirming the direction encodes epistemic stance, not surface statistics. Full layer-by-layer tables appear in Appendix F.

4.3 Decoder steering: controlled range, non-monotonic peak

Figure 3 shows BioGPT decoder steering at L16 with calibrated norms ($\approx$ 205). Four findings emerge.

Finding 1: Clean steering at low-to-mid magnitudes. Hedge score rises from 0.35 (baseline) to 0.90 at $\alpha_{\text{frac}} = 0.125$ (probe direction, $2.6\times$ increase) while lexical diversity (0.94) and perplexity (20.3) remain near baseline (0.95 and 17.0). This is the signature of a clean semantic intervention.

Finding 2: Calibration prevents collapse. With static norm 616, perplexity at $\alpha_{\text{frac}} = 0.25$ exceeds 400. With calibrated norm $\approx$ 205 ($3\times$ smaller effective perturbation), perplexity peaks at 43.7 ($2.6\times$ baseline) and lexical diversity stays at or above 0.91 throughout the tested range.

Finding 3: Non-monotonic hedge response. The probe-direction hedge score peaks at $\alpha_{\text{frac}} = 0.125$ ($H = 0.90$), then oscillates between 0.80 and 0.88 above the peak ($0.84 \rightarrow 0.80 \rightarrow 0.88$ at $\alpha_{\text{frac}} \in \{0.15, 0.20, 0.25\}$), indicating the model transitions through different response regimes even within the collapse-free range. The ortho direction mirrors this with a slightly higher peak ($H = 0.97$ at $\alpha_{\text{frac}} = 0.125$), consistent with the probe direc-

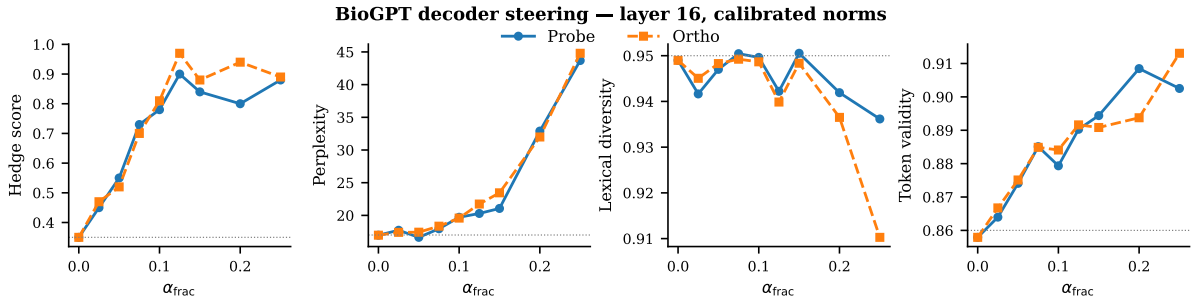

Figure 3: BioGPT decoder steering at layer 16 (calibrated norms, $\|\mathbf{h}_{16}\|_{\text{est}} \approx 205$). Dotted horizontal lines mark baseline values. Both directions show a non-monotonic hedge peak at $\alpha_{\text{frac}} = 0.125$ (Table 8) and no collapse through $\alpha_{\text{frac}} = 0.25$. The visual peak may appear at ≈ 0.10 due to marker interpolation; the numerical peak is at 0.125.

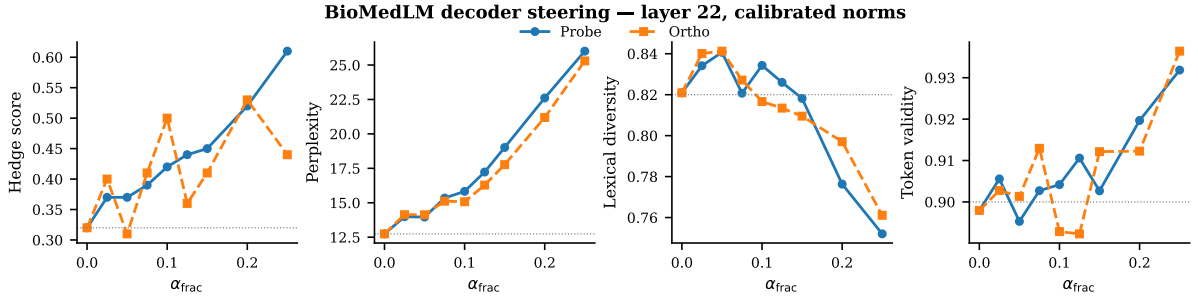

Figure 4: BioMedLM decoder steering at layer 22 (calibrated norms, $\|\mathbf{h}_{22}\|_{\text{est}} \approx 228$, $d = 2560$). Dotted horizontal lines mark baseline values. The hedge response is monotonically increasing with no collapse through $\alpha_{\text{frac}} = 0.25$, in contrast to BioGPT's non-monotonic profile (Figure 3).

tion carrying a small residual surface-feature component that the ortho direction removes. Note also that PPL dips slightly below baseline at $\alpha_{\text{frac}} = 0.05$ (16.6 vs. 17.0): at low magnitudes, the steering vector may nudge outputs closer to the PubMed training manifold (hedge vocabulary is dense there), producing a transient perplexity reduction; this effect disappears at higher magnitudes. Full tables are in Appendix G.

BioMedLM decoder steering at layer 22, calibrated norms. Figure 4 shows the BioMedLM steering profile. At BioMedLM layer 22 ($\|\mathbf{h}\|_{\text{est}} \approx 228$, $d = 2560$), three features distinguish the steering profile from BioGPT's.

Gradual monotonic response. Under the probe direction, hedge score rises from 0.32 (baseline) to 0.61 at $\alpha_{\text{frac}} = 0.25$, a near-linear $1.9\times$ increase with no mid-range dip. Perplexity climbs from 12.7 to 26.0 ($2.0\times$), and lexical diversity declines gradually ($0.82 \rightarrow 0.75$). In contrast to BioGPT, whose hedge score peaks non-monotonically at $\alpha_{\text{frac}} = 0.125$ and then dips, BioMedLM's response is monotonically increasing throughout with no generation quality collapse within the tested range.

Probe direction outperforms ortho. The orthogonalized direction achieves hedge 0.44 at $\alpha_{\text{frac}} = 0.25$ versus 0.61 for the probe direction, a reversal of the BioGPT pattern, where ortho peaked at $H = 0.97$ versus probe's $H = 0.90$. This suggests that for BioMedLM the surface-feature component of the probe direction contributes constructively to the steering effect rather than acting as a confound.

Greater distributional resilience. We attribute BioMedLM's wider collapse-free range to two factors. Its hidden dimension ($d = 2560$) is $2.5\times$ larger than BioGPT's ($d = 1024$): a fixed fractional perturbation covers a proportionally smaller fraction of the representation space per coordinate. Additionally, BioMedLM's near-chance layer-0 accuracy (55.25%) indicates that its uncertainty representation is built through transformer computation rather than surface lexical cues, yielding a steering direction that is more diffuse in activation space and less prone to out-of-distribution excursions at moderate magnitudes. Full numerical results appear in Appendix G.

4.4 Encoder representation steering

Figure 5 shows all five metrics across the four encoder models (probe direction). Key observations:

(i) Consistent positive shifts. All models show strongly positive Δ_{proj} and Δ_{prob}, confirming the steering vector reliably shifts representations toward the uncertain pole.

(ii) Flip-rate saturation. Flip rate rises rapidly to $\approx 50\%$ by $\alpha_{\text{frac}} = 0.10$ and then saturates, indicating a fixed ceiling of flippable sentences is reached once perturbation exceeds a threshold.

(iii) Centroid cosine sign reversal. $\Delta_{\cos}$ is slightly positive at small α_{frac} but turns negative above ≈ 0.10. The probe direction and the mean uncertain-centroid direction are correlated but not identical; large perturbations along the former overshoot the latter, moving representations *away* from the uncertain centroid even as probe-axis projections keep increasing.

(iv) On-axis fraction. For the probe direction, the on-axis fraction is 1.000 by construction. For the ortho direction, it falls to 0.97–0.98 (model-dependent), confirming that orthogonalization introduces only a small geometric deviation (see Appendix G).

(v) Model ordering. SciBERT shows the largest shifts ($\Delta_{\text{proj}} = 1.41$, $\Delta_{\text{prob}} = 0.40$ at $\alpha_{\text{frac}} = 0.05$); BlueBERT shows the smallest (0.77, 0.30), consistent with their relative probe accuracies.

4.5 Cross-model patterns and architectural effects

The probing and steering results together reveal several regularities that illuminate how architecture, scale, and pretraining interact with uncertainty encoding.

Layer-0 accuracy reflects pretraining objective and scale. The gap between the two decoder models at layer 0 is the sharpest contrast in our study: BioGPT achieves 82.00% while BioMedLM achieves only 55.25%, despite both being causal LMs trained on PubMed. BioGPT (345M parameters) apparently relies on surface lexical statistics (hedging tokens are correlated with their local embedding context), whereas BioMedLM (2.7B parameters) abstracts uncertainty into higher-order contextual representations that are not yet formed at layer 0. Encoder models occupy an intermediate position (76–81%), consistent with bidirectional attention enabling richer early contextualization across the full sentence. The pattern is consistent with scale (or the associated increase in hidden dimension and training compute) driving the shift from surface-dependent to computation-dependent uncertainty encoding, though the two decoders differ simultaneously in multiple factors and a controlled ablation would be needed to isolate scale as the causal variable.

Peak accuracy converges despite divergent starting points. Despite the large layer-0 variation, all six models converge to 80–87% at their selected steering layers. This convergence implies a ceiling (set either by the BioScope contrast set composition or by the intrinsic linearly accessible information in 400 sentences) that all models reach regardless of how early their uncertainty representation forms. The encoder ordering at peak accuracy (SciBERT 85.0%, BioBERT 86.5%, ClinicalBERT 82.75%, BlueBERT 80.75%) correlates with pretraining corpus breadth and biomedical specificity, suggesting that broader domain exposure slightly sharpens the uncertainty axis even when absolute accuracy differences are modest.

Encoder steering response is architecturally uniform. All four BERT-family encoders show nearly identical α_{frac} response curves for Δ_{proj} and Δ_{prob} despite being pretrained on corpora ranging from clinical notes (MIMIC-III) to broad scientific text. The decoder contrast is visible in Figure 9. The flip-rate saturation ceiling (50–54%) is consistent across all four, suggesting it reflects BioScope test-set composition rather than model-specific geometry. The model ordering on shift magnitude (SciBERT $>$ ClinBERT $>$ BioBERT $>$ BlueBERT) correlates with peak probe accuracy, confirming that a sharper uncertainty axis produces stronger representational steering effects. This uniformity has a practical implication: uncertainty steering may be transferable across BERT-family encoder variants without per-model calibration of the steering direction, provided probe accuracy is used to select the appropriate layer.

Layer selection interacts with architecture. The ± 2 window refinement matters most for encoders (BioBERT $+2.50$pp, SciBERT $+0.25$pp, ClinicalBERT $+0.75$pp) and is trivially zero for BioGPT. In BERT-family models, probe accuracy can vary non-monotonically near the 2/3 anchor as the `[CLS]` representation transitions between syntactic and semantic processing stages. For causal decoders, the layer-by-layer accuracy profile is flat

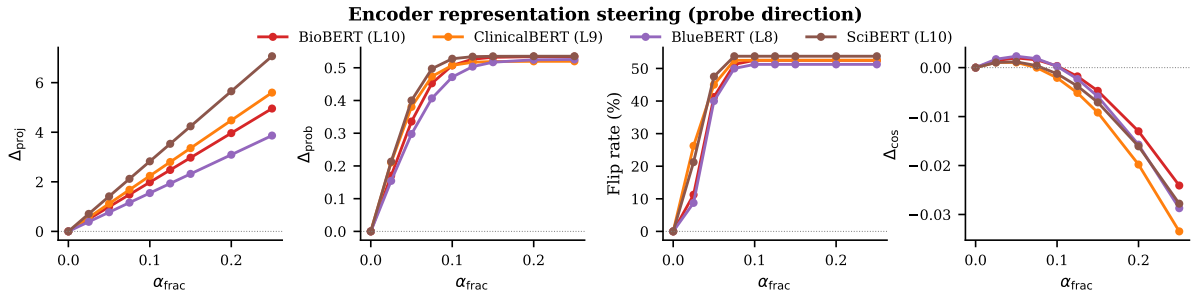

Figure 5: Encoder representation steering (probe direction) across all four models. Top-left: Δ_{proj} (all models positive, linear in α). Top-right: Δ_{prob} (saturates above $\alpha \approx 0.10$). Bottom-left: flip rate (saturates at $\approx 50\%$). Bottom-right: Δ_{cos} (sign reversal above $\alpha \approx 0.10$, dotted line marks zero).

across adjacent layers (Table 5): any layer in the ± 2 window performs equivalently, making precise selection less consequential. This asymmetry supports applying the full window search to all architectures while noting that it adds most value for bidirectional encoders.

5 Discussion

Uncertainty is cleanly decodable across all six models. The narrow range of peak accuracies (80–87%) across models with different architectures, sizes, and pretraining corpora suggests that linear encoding of epistemic stance is a general property of biomedical LMs trained on scientific or clinical text. The layer-0 contrast (BioGPT at 82% vs. BioMedLM at 55%) is particularly informative: for BioGPT, much of the signal is already present in the token embeddings through surface lexical cues, whereas BioMedLM appears to build much of the uncertainty-associated signal through transformer computation despite sharing the same pretraining corpus.

The uncertainty direction is stable but not purely semantic. Orthogonalizing against sentence length and hedge-count reduces probe accuracy by less than 1pp, which shows the representation is not reducible to these two surface statistics. However, this should not be read as evidence that the direction is purely semantic: the probe is trained on BioScope annotations that are themselves defined partly by the presence of hedging vocabulary. Orthogonalizing against a scalar hedge-count feature cannot fully decouple lexical from conceptual content. What we can say is that the direction captures more than a simple hedge-word lookup, and that this residual structure supports activation steering, but the precise mix of lexical and higher-level epistemic content remains an open question.

Layer selection is not arbitrary. The 2/3-depth anchor is a principled starting point, and the ± 2 window consistently identifies equal-or-better layers without over-searching. The BioBERT result ($+2.50$pp, L10 over L8) is the clearest demonstration: the accuracy plateau is not uniformly flat, and the window correctly finds the local optimum. The restriction to ± 2 is deliberate: wider windows risk selecting layers at qualitatively different processing stages.

Dynamic calibration is essential; decodability $\neq$ steerability. For decoders, the critical lesson is that hidden-norm calibration determines whether steering is controlled or catastrophic: a $3\times$ static over-estimate produced collapse at $\alpha_{\text{frac}} = 0.25$; dynamic calibration eliminates it across the entire tested range. Yet even controlled steering is non-monotonic: the peak hedge effect occurs at intermediate magnitudes and weakens above it, suggesting the model transitions between response regimes. For encoders, the Δ_{cos} sign reversal demonstrates that even geometrically precise steering along the probe direction does not guarantee movement toward the uncertain cluster: linear decodability and directional steerability are not equivalent properties.

PPL measurement under the unsteered model. Perplexity is scored under the *unsteered* model p_θ (Eq. 10), not the steered one, because the steered model's own likelihood is inflated along the steering direction. This provides a model-independent fluency proxy, but some PPL increase may reflect lexical style shift (more hedging vocabulary) rather than genuine quality degradation; token validity and lexical diversity serve as complementary

checks.

Multi-metric evaluation is necessary. In the uncalibrated regime (PPL > 400), hedge score alone would suggest strong success. The same logic applies to encoder steering: a large Δ_{proj} paired with negative $\Delta_{\cos}$ is ambiguous without the full metric suite. For steering interventions in safety-sensitive clinical domains, we argue that the target-behavior metric must be paired with at least one fluency/diversity metric (decoders) or one probability/boundary-crossing metric (encoders).

Decoder scale shapes the steering regime. The BioGPT–BioMedLM comparison constitutes a natural experiment: two causal LMs trained on the same corpus but differing in scale by $8\times$ (345M vs. 2.7B parameters) and hidden dimension by $2.5\times$ (1024 vs. 2560). Their steering profiles differ qualitatively. BioGPT exhibits a non-monotonic hedge peak at $\alpha_{\text{frac}} = 0.125$ with a sharp perplexity inflection above it, suggesting its representation space tightly couples the uncertainty direction to generation quality. BioMedLM exhibits a monotonic, collapse-free profile across the same range, and its ortho direction underperforms its probe direction, a reversal of the BioGPT pattern. These observations suggest that model scale modulates not merely the magnitude of the steering response but its qualitative character: larger models appear to afford wider collapse-free operating ranges, with the trade-off of lower peak hedge gains within that range. Whether this is a general scaling property or specific to the PubMed pretraining distribution remains open.

Encoder pretraining distribution has limited influence on steering geometry. The near-identical steering response curves across four BERT-family encoders pretrained on corpora ranging from clinical notes (MIMIC-III) to broad scientific text constitute our strongest evidence that linear uncertainty encoding is corpus-independent when the training objective is masked language modeling. If hedging representation were primarily a function of clinical text exposure, ClinicalBERT would be expected to show qualitatively different behavior from SciBERT; the curves are instead nearly superimposable. Small quantitative differences (SciBERT showing the largest shifts, BlueBERT the smallest) are better explained by differences in probe accuracy—a proxy for the sharpness of the uncertainty axis—rather than corpus content.

6 Limitations

Encoder generation. We measure representation shifts in encoder models only; behavioral impact on downstream NLP tasks (named entity recognition, relation extraction, information extraction) remains open. The $\approx 50\%$ flip rate confirms meaningful prediction changes at the probe level, but task-level evaluation is required to assess real-world impact.

Contrast set. The 400-sentence BioScope contrast set mixes genres (abstracts, full papers, clinical reports) and speculation types (epistemic hedging, conditionals, hypotheticals). A narrower annotation scheme focused specifically on clinical epistemic uncertainty might yield a sharper representation and higher steering fidelity.

Orthogonalization. Only sentence length and hedge-word count are controlled; specific modal constructions, passive voice, negation scope, and syntactic bigrams remain uncontrolled confounds. The claim that uncertainty encoding survives orthogonalization should be read as holding for these two confounds specifically.

Hedge metric. Our 27-term lexicon detects the presence of hedging markers but not their clinical appropriateness. A model generating repetitive hedging vocabulary outside coherent clinical content would receive a high score. Expert radiologist evaluation of output quality is an essential next step for any clinical safety claim.

Single prompt domain. Decoder experiments use radiology prompts exclusively. Steering behavior may differ for other clinical genres such as pathology reports, discharge summaries, or progress notes, where the distribution of uncertainty cues differs substantially.

Window size. The ± 2 window is not cross-validated across held-out model families. Results are empirically insensitive to ± 1 or ± 3 for all six models in this study, but the optimal window width may differ for models with sharper or flatter probe accuracy curves.

References

Heike Adel and Hinrich Schütze. 2015. Comparing approaches to the coreference resolution of clinical speculation and negation. In *Proceedings of the BioNLP 2015 Workshop*.

Shashank Agarwal and Hongfang Yu. 2010. Detecting hedging in biomedical text using a semantic-based approach. In *Proceedings of the BioNLP 2010 Workshop*.

Emily Alsentzer, John R. Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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)*, pages 3615–3620.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. BioMedLM: A 2.7b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

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

Jiayang Geng, Fei Cai, Yida Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8086–8098.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

David M. Panicek and Hedvig Hricak. 2016. How the language of radiology reports can foster (or foil) clinical communication. *Journal of the American College of Radiology*, 13(12):1394–1395. Editorial.

Arjun Panickssery, Nick Gabrieli, Julia Bowman, Meg Tong, Evan Hubinger, and Alexander Mallen. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3363–3377.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Alex Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical hallucination test bed for large language models. *arXiv preprint arXiv:2307.15343*.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7222–7240.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in large language models. In *Proceedings of the 12th International Conference on Learning Representations*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Experimental Details

Hyperparameters. Generation: temperature 0.8, top-p 0.9, max new tokens 60, min new tokens 20, repetition penalty 1.2. Probe: $C = 0.1$, max_iter 2000, random_state 42. 5-fold stratified CV, seed 42. Sample sizes: 200+200 BioScope sentences for probe training; 80 balanced sentences (held out) for encoder steering; 20 seeds $\times$ 5 prompts $=$ 100 generations per decoder configuration.

Reproducibility. Fixed random seeds throughout. BioScope XML files enumerated alphabetically before sampling; contrast set is platform-independent. Code and results available at the repository linked in the first-page footnote.

B Implementation Details

Compute environment. Python 3.11.9, PyTorch 2.5.1+cu121, Transformers 4.49.0, Linux 6.1.0, NVIDIA GPU, CUDA 12.1. Float32 is the canonical dtype.

Dtype robustness. bfloat16 and float16 produce probe accuracies within 0.25pp of float32 on the same CUDA device. BioMedLM bfloat16 shows minor steering drift at $\alpha_{\text{frac}} \geq 0.20$; qualitative findings are unchanged.

Cross-platform note. CUDA and Apple MPS use different floating-point kernels and RNG sequences, producing numerically distinct hidden states even with identical seeds. Within a fixed CUDA platform, results are reproducible.

C Numerical Precision Robustness

Table 4 reports probe accuracy at the selected steering layer $\hat{\ell}$ under three floating-point precisions on the same CUDA device. All six models return identical accuracies across float32, float16, and bfloat16, confirming that numerical precision does not affect the qualitative or quantitative probing findings. Minor steering drift is observed for BioMedLM bfloat16 at $\alpha_{\text{frac}} \geq 0.20$ (hedge score within ± 0.02 of float32); all qualitative findings are unchanged. Float32 is the canonical dtype for all reported results.

D Hedge Marker Lexicon

may, might, could, would, should; suggest, suggests, suggestive; possible, possibly, possibility; probable, probably, probability; likely, unlikely;

Table 4: Probe accuracy at $\hat{\ell}$ across floating-point precisions (same CUDA device). All differences are zero across all models.

Model	$\hat{\ell}$	float32	float16	bfloat16
BioGPT	16	85.50%	85.50%	85.50%
BioMedLM	22	84.25%	84.25%	84.25%
BioBERT	10	86.50%	86.50%	86.50%
ClinicalBERT	9	82.75%	82.75%	82.75%
BlueBERT	8	80.75%	80.75%	80.75%
SciBERT	10	85.00%	85.00%	85.00%

appear, appears, appearing; seem, seems; suspect, suspected, suspicion; consistent with, cannot exclude, rule out.

E Formal Evaluation Metrics

Lexical diversity.

$$D(g) = \frac{|\text{unique words}(g)|}{|\text{words}(g)|} \tag{9}$$

Perplexity.

$$\text{PPL}(g) = \exp\left(-\frac{1}{|g|}\sum_{t=1}^{|g|}\log p_\theta(g_t \mid g_{<t})\right) \tag{10}$$

where p_θ is BioGPT evaluated without steering hooks.

Token validity.

$$V(g) = \frac{|\{t \in g : t \in \mathcal{W}\}|}{|\{t \in g : t \text{ alphabetic}\}|} \tag{11}$$

where $\mathcal{W}$ is the set of valid English words (checked against the NLTK words corpus).

Length.

$$L(g) = |g| \tag{12}$$

Encoder steering metrics.

$$\Delta_{\text{proj}} = (\mathbf{h}^{\text{mod}} - \mathbf{h}^{\text{orig}}) \cdot \hat{\mathbf{v}}_{\text{probe}} \tag{13}$$

$$\Delta_{\text{prob}} = p(y{=}1 \mid \mathbf{h}^{\text{mod}}) - p(y{=}1 \mid \mathbf{h}^{\text{orig}}) \tag{14}$$

$$\text{flip}\% = \frac{|\{s : \hat{y}^{\text{mod}} \neq \hat{y}^{\text{orig}}\}|}{|S|} \tag{15}$$

$$\Delta_{\cos} = \cos(\mathbf{h}^{\text{mod}}, \bar{\mathbf{h}}_u) - \cos(\mathbf{h}^{\text{orig}}, \bar{\mathbf{h}}_u) \tag{16}$$

$$f_{\text{axis}} = \frac{|\Delta\mathbf{h} \cdot \hat{\mathbf{v}}_{\text{probe}}|}{\|\Delta\mathbf{h}\|} \tag{17}$$

where $\bar{\mathbf{h}}_u$ is the mean hidden state of uncertain training sentences, $\hat{y}$ denotes the probe's predicted class, and $\Delta\mathbf{h} = \mathbf{h}^{\text{mod}} - \mathbf{h}^{\text{orig}}$.

F Full Probing Results

Tables 5, 6, and 7 list layer-by-layer 5-fold CV probe accuracies for all six models. *=2/3 anchor; ^=window-selected. Figures 6–7 show bar-chart views of the decoder sweeps.

Table 5: BioGPT probe accuracy, all 25 layers. $\ell^* = 16$, $\hat{\ell} = 16$.

Layer	CV Acc	Std	95% CI
0	82.00%	4.38%	[78.16%, 85.84%]
1	84.50%	4.47%	[80.58%, 88.42%]
2	84.50%	3.26%	[81.64%, 87.36%]
3	84.25%	3.01%	[81.61%, 86.89%]
4	84.75%	2.24%	[82.79%, 86.71%]
5	86.25%	1.98%	[84.52%, 87.98%]
6	85.50%	2.09%	[83.67%, 87.33%]
7	84.25%	3.38%	[81.29%, 87.21%]
8	84.00%	2.40%	[81.89%, 86.11%]
9	85.50%	2.59%	[83.23%, 87.77%]
10	84.75%	2.40%	[82.64%, 86.86%]
11	84.50%	2.44%	[82.36%, 86.64%]
12	86.00%	1.85%	[84.37%, 87.63%]
13	84.75%	1.63%	[83.32%, 86.18%]
14	84.75%	2.05%	[82.95%, 86.55%]
15	85.00%	1.98%	[83.27%, 86.73%]
16*^	**85.50%**	**2.44%**	**[83.36%, 87.64%]**
17	84.75%	2.24%	[82.79%, 86.71%]
18	85.25%	2.85%	[82.75%, 87.75%]
19	85.50%	2.44%	[83.36%, 87.64%]
20	84.75%	3.11%	[82.02%, 87.48%]
21	84.25%	2.88%	[81.73%, 86.77%]
22	82.00%	2.27%	[80.01%, 83.99%]
23	81.50%	2.24%	[79.54%, 83.46%]
24	82.50%	2.50%	[80.31%, 84.69%]

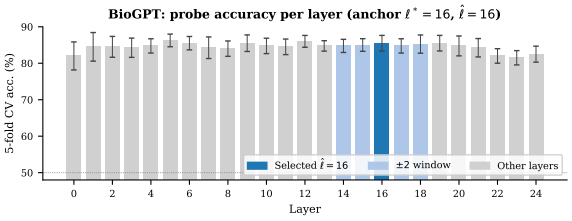

Figure 6: BioGPT probe accuracy per layer. Blue = selected L16; light blue = ±2 window; grey = other layers.

G Detailed Steering Results

Figure 10 shows the steering geometry: ortho direction on-axis fraction (0.97–0.98, constant across α) and probe direction $\Delta_{\cos}$ (sign reversal above $\alpha \approx 0.10$).

Table 6: BioMedLM probe accuracy, all 33 layers. $\ell^* = 21$, $\hat{\ell} = 22$.

Layer	CV Acc	Std	95% CI
0	55.25%	3.69%	[52.02%, 58.48%]
1	81.25%	2.65%	[78.93%, 83.57%]
2	82.50%	3.19%	[79.71%, 85.29%]
3	84.25%	3.38%	[81.29%, 87.21%]
4	83.00%	2.09%	[81.17%, 84.83%]
5	84.50%	3.14%	[81.75%, 87.25%]
6	84.75%	1.85%	[83.12%, 86.38%]
7	85.25%	3.47%	[82.21%, 88.29%]
8	84.75%	2.85%	[82.25%, 87.25%]
9	83.75%	2.93%	[81.18%, 86.32%]
10	83.25%	2.44%	[81.11%, 85.39%]
11	83.75%	1.98%	[82.02%, 85.48%]
12	84.50%	2.27%	[82.51%, 86.49%]
13	85.75%	2.09%	[83.92%, 87.58%]
14	85.50%	2.44%	[83.36%, 87.64%]
15	86.00%	1.05%	[85.08%, 86.92%]
16	86.00%	1.37%	[84.80%, 87.20%]
17	84.25%	1.90%	[82.59%, 85.91%]
18	83.75%	2.17%	[81.85%, 85.65%]
19	84.00%	1.85%	[82.37%, 85.63%]
20	83.75%	1.53%	[82.41%, 85.09%]
21*	84.00%	2.40%	[81.89%, 86.11%]
22$^\wedge$	**84.25%**	**1.90%**	**[82.59%, 85.91%]**
23	83.50%	2.98%	[80.88%, 86.12%]
24	83.50%	2.24%	[81.54%, 85.46%]
25	83.50%	3.47%	[80.46%, 86.54%]
26	83.00%	3.14%	[80.25%, 85.75%]
27	83.25%	3.38%	[80.29%, 86.21%]
28	82.75%	3.79%	[79.43%, 86.07%]
29	82.50%	4.15%	[78.87%, 86.13%]
30	82.25%	4.54%	[78.27%, 86.23%]
31	80.50%	5.05%	[76.08%, 84.92%]
32	81.00%	4.71%	[76.87%, 85.13%]

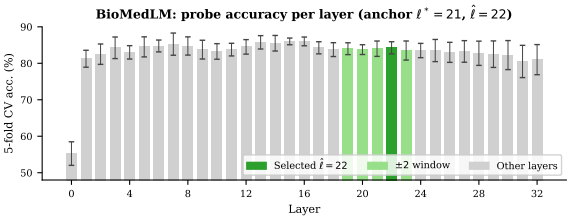

Figure 7: BioMedLM probe accuracy per layer. Green = selected L22; light green = ±2 window.

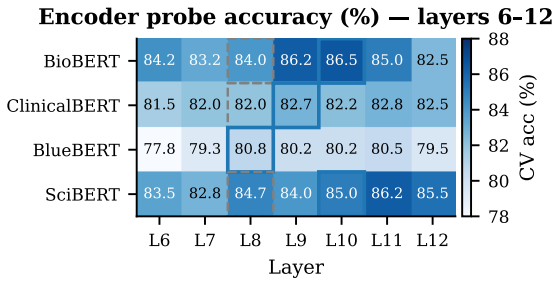

Figure 8: Encoder probe accuracy heatmap for layers 6–12. Blue boxes mark selected $\hat{\ell}$; dashed boxes mark the 2/3 anchor L8.

Table 7: Encoder probe accuracy (%) across all 13 layers. Anchor $\ell^* = 8$; *=anchor; $^\wedge$=window-selected; $^{*\wedge}$=anchor and window-selected coincide.

L	BioBERT	ClinBERT	BlueBERT	SciBERT
0	76.0±2.2	77.8±3.2	77.0±2.4	80.5±4.3
1	82.2±3.8	80.5±2.9	79.5±2.4	81.5±2.4
2	80.2±2.1	81.8±3.6	79.2±3.8	83.5±3.5
3	82.5±3.3	83.2±5.1	80.5±3.6	83.2±5.0
4	83.2±4.2	83.0±4.1	79.5±5.3	83.2±4.7
5	85.0±5.4	81.8±4.1	78.5±4.2	84.0±5.0
6	84.2±4.5	81.5±2.7	77.8±4.2	83.5±2.9
7	83.2±4.5	82.0±2.1	79.2±5.4	82.8±2.1
8*	84.0±4.5	82.0*±3.6	**80.8**$^{*\wedge}$±5.4	84.8*±4.5
9	86.2±3.6	**82.8**$^\wedge$±3.8	80.2±2.2	84.0±3.9
10$^\wedge$	**86.5**$^\wedge$±2.4	82.2±2.2	80.2±2.6	**85.0**$^\wedge$±3.4
11	85.0±4.0	82.8±2.4	80.5±1.4	86.2±3.6
12	82.5±2.7	82.5±2.7	79.5±1.7	85.5±3.4

Acc in %; ± denotes std over 5-fold CV.

Table 8: BioGPT decoder steering, layer 16, calibrated norms ($\|\mathbf{h}_{16}\| \approx 205$). Bold = peak hedge.

α	Hedge	Lex.div.	Tok.val.	PPL
Probe direction				
0.000	0.35	0.95	0.86	17.0
0.025	0.45	0.94	0.86	17.7
0.050	0.55	0.95	0.87	16.6
0.075	0.73	0.95	0.89	18.0
0.100	0.78	0.95	0.88	19.7
0.125	**0.90**	0.94	0.89	20.3
0.150	0.84	0.95	0.89	21.1
0.200	0.80	0.94	0.91	32.9
0.250	0.88	0.94	0.90	43.7
Ortho direction				
0.000	0.35	0.95	0.86	17.0
0.025	0.47	0.95	0.87	17.4
0.050	0.52	0.95	0.88	17.4
0.075	0.70	0.95	0.88	18.3
0.100	0.81	0.95	0.88	19.6
0.125	**0.97**	0.94	0.89	21.7
0.150	0.88	0.95	0.89	23.5
0.200	0.94	0.94	0.89	32.0
0.250	0.89	0.91	0.91	44.8

Table 9: BioMedLM decoder steering, layer 22, calibrated norms ($\|\mathbf{h}_{22}\| \approx 228$). Bold = peak hedge.

α	Hedge	Lex.div.	Tok.val.	PPL
Probe direction				
0.000	0.32	0.82	0.90	12.7
0.050	0.37	0.84	0.90	14.0
0.100	0.42	0.83	0.90	15.8
0.150	0.45	0.82	0.90	19.0
0.200	0.52	0.78	0.92	22.6
0.250	**0.61**	0.75	0.93	26.0
Ortho direction				
0.000	0.32	0.82	0.90	12.7
0.050	0.31	0.84	0.90	14.1
0.100	0.50	0.82	0.89	15.1
0.150	0.41	0.81	0.91	17.8
0.200	0.53	0.80	0.91	21.2
0.250	**0.44**	0.76	0.94	25.3

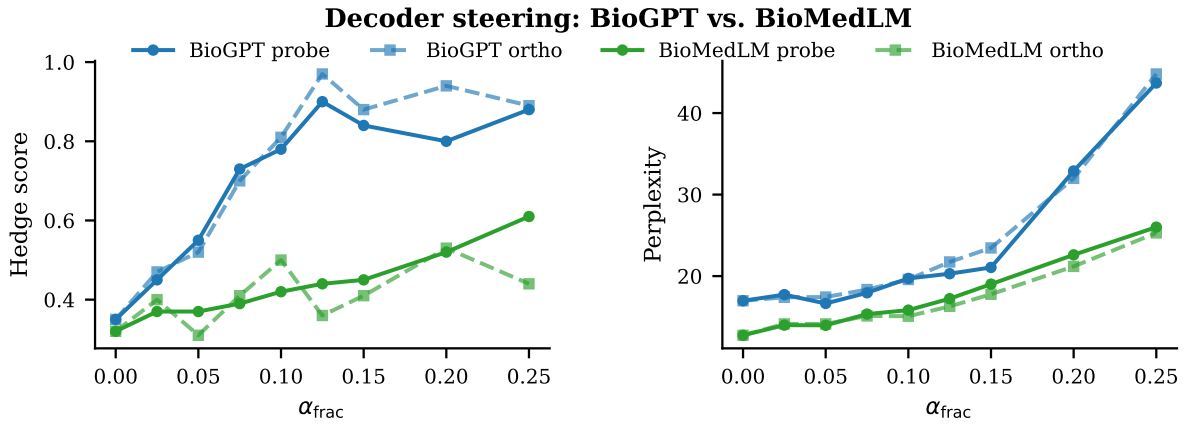

Figure 9: Side-by-side decoder steering comparison: BioGPT (L16, $\|\mathbf{h}\| \approx 205$, $d = 1024$) vs. BioMedLM (L22, $\|\mathbf{h}\| \approx 228$, $d = 2560$), probe and ortho directions. Left: hedge score. Right: perplexity. BioGPT exhibits a non-monotonic peak and steeper perplexity rise; BioMedLM shows a monotonic, collapse-free profile.

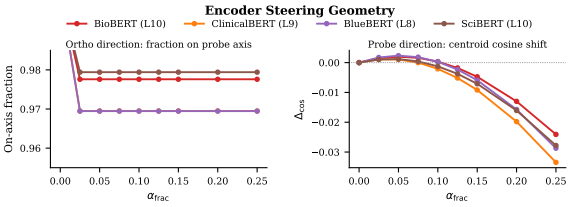

Figure 10: Left: ortho direction on-axis fraction (constant, 0.97–0.98 depending on model). Right: probe direction $\Delta_{\cos}$ showing sign reversal above $\alpha \approx 0.10$.

Table 10: Encoder steering (probe direction) at selected layers, averaged over 80 test sentences. On-axis = 1.000 for all entries (probe direction by construction).

Model	α	Δ_{proj}	Δ_{prob}	flip%	$\Delta_{\cos}$
BioBERT L10	0.00	+0.00	+0.000	0.0%	+0.000
	0.05	+0.99	+0.336	41.2%	+0.002
	0.10	+1.98	+0.507	52.5%	+0.000
	0.15	+2.97	+0.532	52.5%	−0.005
	0.25	+4.96	+0.534	52.5%	−0.024
ClinBERT L9	0.00	+0.00	+0.000	0.0%	+0.000
	0.05	+1.12	+0.381	45.0%	+0.001
	0.10	+2.24	+0.507	52.5%	−0.002
	0.15	+3.36	+0.519	52.5%	−0.009
	0.25	+5.60	+0.520	52.5%	−0.034
BlueBERT L8	0.00	+0.00	+0.000	0.0%	+0.000
	0.05	+0.77	+0.298	40.0%	+0.002
	0.10	+1.55	+0.472	51.2%	+0.000
	0.15	+2.32	+0.517	51.2%	−0.006
	0.25	+3.87	+0.525	51.2%	−0.029
SciBERT L10	0.00	+0.00	+0.000	0.0%	+0.000
	0.05	+1.41	+0.400	47.5%	+0.001
	0.10	+2.83	+0.528	53.8%	−0.001
	0.15	+4.24	+0.535	53.8%	−0.007
	0.25	+7.07	+0.535	53.8%	−0.028