

Reverse Probing: Supervised Token-level Uncertainty Quantification for Large Language Models in Clinical Text

Bushi Xiao

University of Florida
Gainesville, FL
xiaobushi@ufl.edu

Sarvesh Soni

U.S. National Library of Medicine
Bethesda, MD
sarvesh.soni@nih.gov

Daisy Zhe Wang

University of Florida
Gainesville, FL
daisyw@cise.ufl.edu

Abstract

As large language models are increasingly deployed for clinical text, ensuring they can reliably signal their own uncertainty becomes critical. Most existing uncertainty quantification (UQ) methods are designed for open-domain generation and cannot localize uncertainty at the token or span level in long clinical text. We propose Reverse Probing, the first uncertainty quantification framework specialized for clinical summarization, which estimates token-level uncertainty directly from pre-existing clinical summaries. Rather than sampling new outputs, Reverse Probing treats the text as a probe into the model’s internal state, extracting uncertainty features from four categories of signals. We evaluate Reverse Probing on two expert-annotated clinical datasets and outperform eight adapted baselines, achieving up to $4\times$ higher AUPRC while running over $15\times$ faster than sampling-based methods. Feature analysis reveals that the log-logit difference induced by the source record and the neighborhood context are the most consistent predictors across all models. This study offers interpretable insights into how models internally respond to unsupported clinical content.

1 Introduction

Uncertainty quantification (UQ) for Large Language Models (LLMs) has been widely studied in general domains, yet its application to clinical text remains limited. Clinical text presents distinct challenges from general tasks: it is dense with terminology, errors often occur at a single word or number within an otherwise correct sentence, and small mistakes can carry serious consequences. Most existing UQ methods operate at the case level and are poorly suited for fine-grained, token-level uncertainty in long clinical text.

UQ is also fundamentally distinct from hallucination detection. Rather than using external methods to detect whether a model’s output is factually cor-

rect, UQ asks how confident the model is in its own predictions. This is a form of self-assessment that reflects model reliability independent of ground truth. Yona et al. (2026) formalizes this distinction, redefining hallucination as a confident error and arguing that models should express honest uncertainty instead of optimizing solely for accuracy.

Consequently, token-level UQ offers a more actionable pathway for model alignment and refinement. The prevailing assumption is that UQ requires generation: sample the model’s outputs multiple times and treat divergence as a proxy for uncertainty. This method works for short question-answering tasks, but is unsuitable for clinical settings, where generating new clinical notes is hard to evaluate reliably, repeated sampling over long documents is computationally expensive, and expert-annotated clinical text is scarce.

Traditionally, token-level probing is designed for encoder-only transformer models like BERT (Devlin et al., 2019), which evaluate confidence by predicting a masked token within a bidirectional context. Decoder-only LLMs operate auto-regressively and do not natively support such token-masking method. We bridge this architectural gap with a teacher forcing strategy to imitate the mask prediction on bi-direction models. Though the method is not equivalent to a native MLM prediction, the model is mimicking a pseudo mask prediction task.

Reverse Probing reflects a deliberate inversion of the standard UQ paradigm. Conventional methods ask the model to generate text and then evaluate the output. We reverse this direction: instead of examining what the model produces, we feed existing clinical text back into the LLM to extract a comprehensive snapshot of its internal activations. A well-supported token inherently aligns with and anchors to the relevant clinical records within the representation space. Conversely, when a token is unsupported, this factual grounding is absent. We harness this absence of alignment as a latent uncer-

tainty signal, effectively turning the text itself into a probe of the model rather than a product of it.

Our contributions are as follows:

- We introduce Reverse Probing¹, the first token-level UQ framework for clinical summaries. Our framework eliminates multi-sampling to reach an exceptional computational efficiency.
- We establish the first benchmark for UQ on clinical summaries. We conduct a systematic evaluation comparing Reverse Probing against eight baselines across six LLMs on two expert-annotated datasets.
- We analyse where the uncertainty signal comes from. The log-logit difference induced by the source record (`delta_energy`) and local neighborhood context are the most consistent predictors across models and feature configurations.

2 Related Work

2.1 Uncertainty Quantification

Traditionally, uncertainty quantification relies on sampling-based approaches. These methods require the model to generate multiple output candidates. The uncertainty is then quantified from the variance or distribution of the multiple generated text. SelfCheckGPT (Manakul et al., 2023), Semantic Uncertainty (Kuhn et al., 2023), and Semantic Entropy (Farquhar et al., 2024) follow this paradigm. Hou et al. (2024) further decomposes uncertainty into aleatoric and epistemic components. However, these methods require task-level new text generation without localizing uncertainty to specific tokens or spans.

Several recent methods derive uncertainty from model internals. Liu et al. (2024); Vazhentsev et al. (2025); Li et al. (2025) show that hidden activations and attention heads encode uncertainty information. EUQ (Huang et al., 2026) on evidential uncertainty also performs on internal signals, which decomposes uncertainty into conflict and ignorance to flag misbehaving vision-language model responses at the level of a full generation. Among all methods above, only Semantic Entropy (SE) Probes

(Kossen et al., 2024) adopts a supervised formulation, which trains a lightweight probe on hidden states to predict the semantic entropy.

Despite these advances, all existing methods are designed and evaluated on general-domain tasks, and score a generation as a whole. Unsupported content in clinical summary is sparse and crucial, with specialized terminologies, so uncertainty has to be localized rather than aggregated into a paragraph-level single score. More importantly, the summaries are pre-existing text rather than being generated by the model. No existing method fits this setting, which requires working post hoc on a fixed summary, localizing at the span level, and operating without access to the generation process.

2.2 Clinical Explainability and Faithfulness

Explainability methods have been widely applied to clinical decision support, but rarely to clinical uncertainty itself. SHAP (Lundberg and Lee, 2017) remains the standard framework for feature-level explanation, while Hur et al. (2025) shows that natural language explanations outperform raw SHAP scores for clinical decision-making. Qureshi et al. (2025) find that no single XAI method dominates across all clinical criteria. Several works argue that combining UQ with explainability improves reliability in high-stakes settings (Salvi et al., 2025; Dubey et al., 2025; Lu et al., 2024). However, none of the work applies this combination specifically in clinical text.

Faithfulness evaluation for clinical summaries also presents a related challenge. Standard metrics correlate poorly with expert judgment (Adams et al., 2023), and DocLens (Xie et al., 2024) improves this with claim-level evaluation. Croxford et al. (2025) validate LLM-as-a-Judge as a scalable alternative to human annotation. Wang et al. (2025) uses semantic consistency across sampled outputs as a semantic signal for radiology reports, but their approach still depends on sampling and operates at the sentence level. Token-level faithfulness evaluation over clinical text remains an open problem.

3 Methodology

In clinical summaries, the majority of the sentences are factually supported, while unsupported content appears only in sparse, scattered spans. This makes fine-grained UQ fundamentally different from case-level evaluations, where standard methods are too

¹<https://github.com/kitayamachingtak/ReverseProbing>

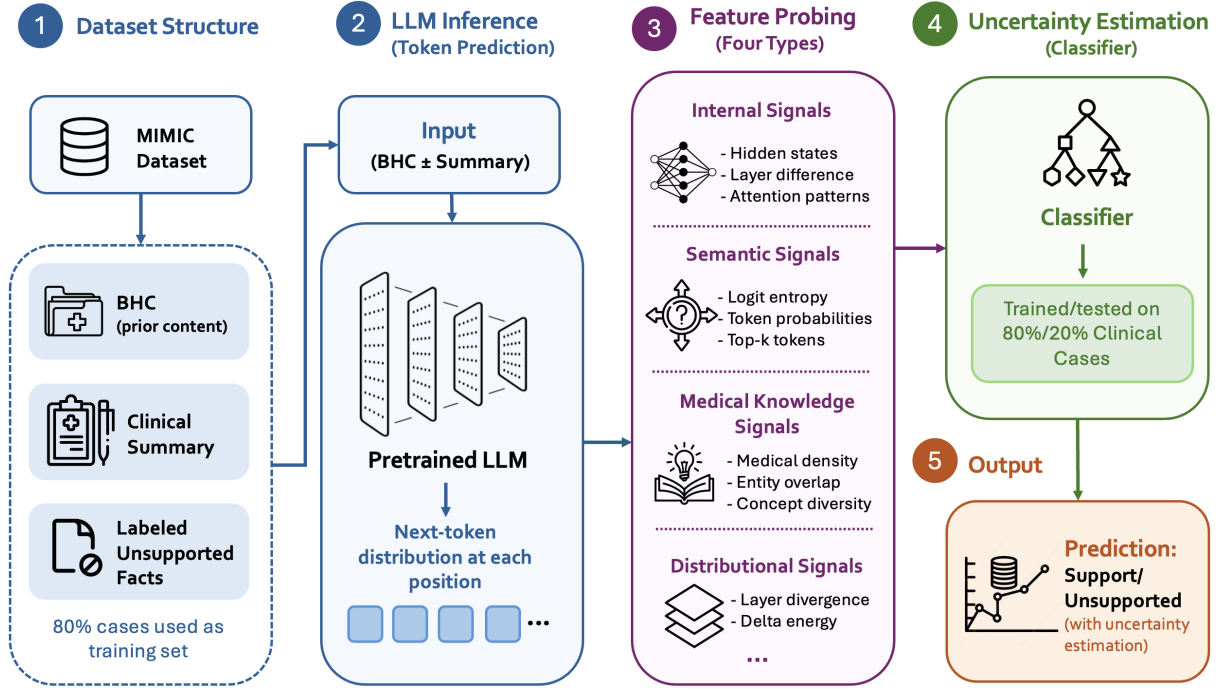


Figure 1: Overview of the Reverse Probing framework. Each summary is passed through a frozen LLM twice, once with the BHC text and once without, yielding distributional, semantic, internal, and medical-knowledge features at every token position. A supervised classifier then predicts token-level uncertainty.

coarse to capture localized errors.

Our approach is inspired by masked probing for bidirectional encoder models, where a token is hidden and the model is asked to recover it from its surroundings. That estimate is exactly what we want. However, decoder-only LLMs are auto-regressive and offer no masking mechanism. Teacher forcing gives us the same quantity without one: at every position the model predicts the token rather than reading it, so each token is effectively masked from its own prediction. This pseudo-masking turns a generative model into a token-level probe without modifying its weights.

The key insight behind Reverse Probing is to treat the source record as an attention anchor. In clinical domain, each discharge summary is written from the Brief Hospital Course (BHC), a long clinical note documenting the patient’s admission and treatment. When a summary token is well supported, the model’s attention concentrates on the BHC and its prediction sharpens once the record is present. When a token is unsupported, that concentration is absent or misdirected, and removing the BHC changes little. We therefore read the absence of grounding as a signal in its own right, by comparing activations with and without the BHC. This is precisely where the "reverse" in Reverse Probing comes from. While traditional methods

generate new text to check for factual accuracy, our approach does the opposite: it uses the existing text as a probe to locate the key features and uncover exactly where the model’s internal layers are struggling to align with clinical evidence.

Concretely, we feed the clinical summary with and without the BHC into a frozen LLM as shown in Figure 1, extract features from its internal activations and output distributions, and train a classifier to predict which part of the internal representation will be activated when encountering an unsupported span. We organize features into four categories described in Section 3.2.

3.1 Dataset

This work uses Hallucinations-MIMIC-DI and Hallucinations-Generated-DI introduced by Heggelmann et al. (2025). Both datasets are derived from MIMIC-IV-Note (Johnson et al., 2023), a large-scale corpus of de-identified clinical notes from the Beth Israel Deaconess Medical Center. Each clinical case consists of a BHC paired with a discharge summary. The BHC serves as the source document recording the patient’s clinical course, while the summary condenses this information for the patient. Hallucinations-MIMIC-DI contains 100 doctor-written discharge summaries, and Hallucinations-Generated-DI contains 100

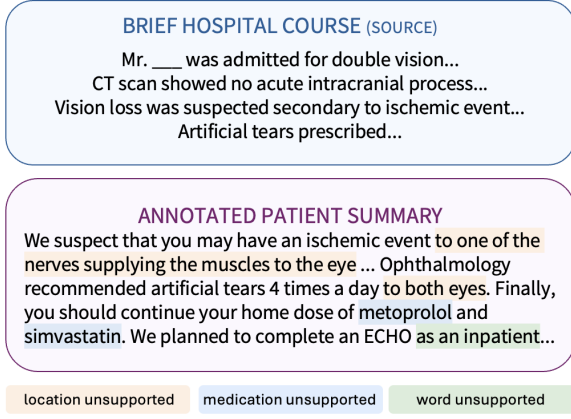


Figure 2: An example from Hallucinations-MIMIC-DI. The BHC serves as the source for the patient discharge summary. Highlighted spans in the summary mark unsupported facts annotated by experts, colored by label types. Dataset Hallucinations-Generated-DI follows the same structure.

LLM-generated summaries (Hegselmann et al., 2024), while the BHC text from both datasets originates from MIMIC-IV-Note. Both datasets provide span-level annotations of unsupported facts by two trained medical experts, where each annotation records the start and end offsets of the unsupported span with error types, as illustrated in Figure 2. These two datasets are highly imbalanced. In Hallucinations-MIMIC-DI, unsupported spans account for 7.33% of total tokens with an average of 2.86 spans per summary, whereas in Hallucinations-Generated-DI the ratio drops to 2.05% with an average of 1.14 spans per summary.

3.2 Feature Extraction

We feed the existing summary back through the model under teacher forcing, as shown below.

```
Brief Hospital Course: {BHC}
i = 1: Summary:           → p1
i = 2: Summary: The       → p2
i = 3: Summary: The patient → p3
      ⋮
```

At position i the model gives a distribution over the vocabulary, $\mathbf{p}_i = P(\cdot \mid \text{BHC}, t_{<i})$, conditioned on the clinical record and on the summary tokens that came before. We read that distribution at every position, along with the hidden states and attention rows there, and turn each position into one feature vector. Causal masking means position i depends only on its own prefix, so a single pass covers the

whole summary at once. Reverse Probing thus works at the token level while costing two forward passes per summary, no matter how long the summary is. Repeating the pass without the BHC gives a second distribution $\mathbf{p}_i^-$ at each position, and the contrast between the two forms one of four feature categories: distributional, semantic, internal, and medical knowledge.

We propose four types of sources that potentially affect the uncertainty: internal signals, medical knowledge, semantic, and distributional signals. To study how feature granularity affects performance, we define four configurations of increasing detail levels: 93, 120, 204, and up to 454 or 886 features. All models are evaluated on the 93, 120, and 204 configurations. We collect 454 as the maximum features for 7B and 8B models, while we collect 886 features for 70B models. This is because 70B models have 80 layers and therefore they need more layer-wise and attention features to sample. Complete details of all features are provided in Appendix D.

Internal Signals. From the with-BHC pass, at each sampled layer l , we extract the ℓ_2 norm, mean, and the standard deviation of the hidden state $\mathbf{h}_i^{(l)}$, together with the L2 distance and cosine similarity between consecutive sampled layers.

Attention is characterized along three axes. Let $\alpha_{i,j}^{(l,h)}$ be the weight from position i to j at layer l , head h , and let $\mathcal{L}_{\text{BHC}}$ index the BHC prefix. *Dispersion* is the entropy of the attention row:

$$H_i^{(l,h)} = - \sum_j \alpha_{i,j}^{(l,h)} \log \alpha_{i,j}^{(l,h)}, \quad (1)$$

distinguishing tokens that draw on a few specific positions from those whose attention is diffuse. *Grounding* is the total mass directed at the clinical record:

$$A_i^{(l,h)} = \sum_{j \in \mathcal{L}_{\text{BHC}}} \alpha_{i,j}^{(l,h)}, \quad (2)$$

which measures whether a token is supported by the record or only by the surrounding summary. *Stability* captures how much this grounding is revised as representations are refined. A well-supported token should not need to reallocate attention across depth. We measure it as the KL divergence between the same head’s attention rows at adjacent sampled layers:

$$\text{Drift}_{l \rightarrow l'}^{(h)} = D_{\text{KL}}(\tilde{\alpha}_i^{(l,h)} \parallel \tilde{\alpha}_i^{(l',h)}), \quad (3)$$

where $\tilde{\alpha}$ denotes the row after ε -smoothing and renormalization.

To capture cumulative rather than per-layer grounding, we apply attention rollout (Abnar and Zuidema, 2020): $\mathbf{R}_\ell = \prod_{k=0}^{\ell} (0.5 \mathbf{A}_k + 0.5 \mathbf{I})$ with row-wise renormalization at each step, and record the accumulated mass on $\mathcal{L}_{\text{BHC}}$, its entropy, and its maximum weight at a shallow, a middle, and a deep checkpoint. This traces how far information from the record actually reaches by each depth.

Semantic Signals. Given logits $\mathbf{z}$ and $\mathbf{p} = \text{softmax}(\mathbf{z})$, we compute Shannon entropy $H(\mathbf{p}) = -\sum_v p_v \log p_v$, the top-two margin $p_{(1)} - p_{(2)}$, the Gini coefficient of $\mathbf{p}$, and a free-energy score $\mathcal{E} = -\log \sum_v \exp(z_v)$, which reflects the overall evidence the model assigns at this position without softmax normalization. Together these describe how concentrated or diffuse the prediction is. In traditional methods, uncertainty is considered to be high when the prediction is diffused.

Medical Knowledge Signals. These encode domain context around each token. Named entity recognition is run offline with scispaCy (Neumann et al., 2019) against UMLS/MedDRA (Bodenreider, 2004), yielding entity types and confidence scores. Medical density measures the fraction of neighboring tokens recognized as medical entities across window sizes. Corpus-level frequency and IDF are computed over the full corpus. We also compute a conditional PMI, which standardizes PMI_i within the entity type e assigned to token i :

$$\text{cPMI}_i = \frac{\text{PMI}_i - \mu_e}{\sigma_e}, \quad (4)$$

where μ_e and σ_e are the mean and standard deviation of PMI over all tokens of type e . These corpus-level statistics are computed without reference to the annotations, so they do not leak label information across the train/test split.

Distributional Signals. The central quantity is the change the clinical record induces at each position. We measure it as the point-wise mutual information between the target token and the BHC,

$$\text{PMI}_i = \log p_{t_i} - \log p_{t_i}^-, \quad (5)$$

which is positive when the record raises the model’s belief in the token, and negative when the token is produced despite, rather than because of, the record. We also use a normalized variant $\text{nPMI}_i = \text{PMI}_i / (-\log p_{t_i})$, taking $-\log p_{t_i}$ as the normalizer.

We compare the two distributions as a whole through Δp , ΔH , $\Delta \mathcal{E}$, and the KL and Jensen–Shannon divergences between $\mathbf{p}$ and $\mathbf{p}^-$. Bringing in $\mathbf{p}^{\text{prior}}$ separates the entropy reduction due to the record from that of the surrounding text:

$$\underbrace{H(\mathbf{p}^{\text{prior}}) - H(\mathbf{p}^-)}_{\text{context}} + \underbrace{H(\mathbf{p}^-) - H(\mathbf{p})}_{\Delta H_{\text{BHC}}} \quad (6)$$

$$= H(\mathbf{p}^{\text{prior}}) - H(\mathbf{p}).$$

A token with $\Delta H_{\text{BHC}} \approx 0$ is resolved by context alone: fluent, but not grounded. Finally, neighborhood features aggregate token probabilities over sliding windows around i , since tokens inside the same unsupported span tend to share uncertainty patterns.

3.3 Classifier

We hypothesize that uncertainty is discretely distributed in the internal representations and that each model’s internal activation differs accordingly. Therefore, fixed parameters, thresholds, and weights cannot accurately capture uncertainty, nor can they fairly generalize across different models. To address this, we adopt a supervised machine learning approach to train tree classifiers XGBoost (Chen and Guestrin, 2016) and CatBoost (Prokhorenkova et al., 2018), and linear classifier Logistic Regression (Pedregosa et al., 2011) on extracted features to predict token-level uncertainty according to the change of contexts.

Data is split at the case level, treating each BHC-summary pair as one case, with 80% for training and 20% for testing. Our method predicts a continuous uncertainty score in $[0,1]$ for each token, and a threshold is then applied to determine whether a target token is flagged as uncertain. We then perform 5-fold cross-validation on test sets.

3.4 Experiments

We evaluate on six LLMs: Mistral-7B-v0.1 (Jiang et al., 2023), Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct (Dubey et al., 2024), BioMistral-7B (Labrak et al., 2024), Llama3-OpenBioLLM-8B and Llama3-OpenBioLLM-70B (Pal and Sankarababu, 2024). The first three are general-domain models; the latter three are pretrained or fine-tuned on biomedical data.

No existing UQ method targets clinical summaries directly, so we adapted eight general-domain baselines to our setting: Token Entropy (Ma et al., 2025), Semantic Energy (Ma et al.,

Method	MIMIC-DI				Generated-DI			
	BioMistral-7B		OpenBioLLM-8B		BioMistral-7B		OpenBioLLM-8B	
	AUCROC	AUPRC	AUCROC	AUPRC	AUCROC	AUPRC	AUCROC	AUPRC
Token Entropy	0.4337	0.0658	0.4832	0.0767	0.4421	0.0366	0.5363	0.0440
Semantic Energy	0.4714	0.0810	0.5375	0.0914	0.4697	0.0436	0.5534	0.0444
Semantic Entropy	0.6448	0.0868	0.5804	0.0854	0.5361	0.1050	0.5475	0.0735
Sliding Window Entropy	0.5127	0.0899	0.5877	0.1267	0.5884	0.0678	0.6940	0.1041
Decomposing Uncertainty	0.5938	0.0825	0.5657	0.0775	0.5925	0.1322	0.5959	0.1410
SE Probes	0.7204	0.1313	0.6015	0.0587	0.4061	0.1075	0.5909	0.1553
Uncertainty Aware Heads	0.5508	0.0810	0.5488	0.0852	0.6150	0.0575	0.6403	0.0682
Attention Chain	0.5195	0.0874	0.4913	0.1012	0.4549	0.0388	0.5098	0.0430
Reverse Probing (XGBoost)	0.8911	0.5478	0.8833	0.4976	0.8819	0.2316	0.9364	0.2892
Reverse Probing (CatBoost)	<u>0.8691</u>	<u>0.5178</u>	0.8848	0.5029	<u>0.8435</u>	0.2805	0.9144	0.2895
Reverse Probing (LRegression)	0.8508	0.3636	0.8545	0.4399	0.8122	0.1592	0.8807	0.2341

Table 1: Main results on MIMIC-DI and Generated-DI. Reverse Probing uses the max feature configuration (454 features). Baselines are adapted from general-domain methods as described in Section 3.4. Best results per column are **bold**; second best are underlined.

2025), Uncertainty Aware Heads (Vazhentsev et al., 2025), Attention Chain (Li et al., 2025), SE (Semantic Entropy) Probes (Kossen et al., 2024), Sliding Window Entropy (Sriramanan et al., 2024), Decomposing Uncertainty (Hou et al., 2024), and Semantic Entropy (Farquhar et al., 2024; Kuhn et al., 2023). Since our datasets consist of pre-existing clinical summaries rather than model-generated text, each baseline requires adaptation. Each adaptation preserves the core mechanism of the original method while replacing generation-dependent components with inference-only equivalents, ensuring a fair comparison as baselines.

For Semantic Entropy, we replace full-response sampling with sentence-level sampling: for each sentence in a summary, we prompt the LLM to explain it ten times, then compute semantic entropy over the resulting outputs. Sentences containing uncertain or hallucinated content tend to produce semantically divergent paraphrases, preserving the spirit of the original method.

For Sliding Window Entropy, we use the best hyperparameter configuration from grid search: window size 9, top- k 20.

For SE Probes method, we use the sentence-level semantic entropy scores obtained from the Semantic Entropy adaptation above as supervision signal to train a linear probe over hidden states. We extract hidden states from every 4 layers in a single forward pass, covering shallow, middle, and deep representations. At test time, the probe predicts uncertainty from a single forward pass without any sampling, replacing the need for multiple generations. We use the same 80/20 train-test split as our

other experiments.

For baselines originally designed around masked token prediction, we replace masking with a cloze-style prompt: each token is withheld and the LLM predicts it from context, after which we compute Shannon entropy over the predicted distribution.

4 Results and Analyses

4.1 Main Results

We evaluate all methods at the token level. We report the AUPRC and AUCROC for unsupported tokens across all documents. Micro F1 is reported in Appendix E. Rather than judging whether an entire summary is reliable, we assess each token individually against expert annotations. While these metrics partially overlap with hallucination detection, our objective is fundamentally distinct and more intricate: we measure how models’ internal representations activate when encountering unsupported facts.

Although our AUCROC values are high, this are partially affected by the severe class imbalance in both datasets: 7.3% unsupported tokens in MIMIC-DI, and only 2.05% unsupported tokens in Generated-DI. AUCROC inflates because the true negatives dominate the score. AUPRC is more informative here, so we utilize it as the main metric. Our results show consistent improvement on this metric across all models and configurations. Among all models and configurations, our two tree-structured classifiers XGBoost and CatBoost perform comparably. Logistic regression is consistently weaker. However, it still beats every baseline

Model	Best baseline	Baseline	RP
<i>MIMIC-DI</i>			
Mistral-7B	Attention Chain	0.0918	0.6022
BioMistral-7B	SE Probes	0.1313	0.5478
Llama3.1-8B	SW Entropy	0.1171	0.5140
OpenBioLLM-8B	SW Entropy	0.1267	0.5029
Llama3.1-70B	SW Entropy	0.1090	0.3934
OpenBioLLM-70B	SW Entropy	0.1244	0.3597
<i>Generated-DI</i>			
Mistral-7B	SE Probes	0.2067	0.3070
BioMistral-7B	Decomp. Unc.	0.1322	0.2805
Llama3.1-8B	SE Probes	0.1386	0.2761
OpenBioLLM-8B	SE Probes	0.1553	0.2895
Llama3.1-70B	SE Probes	0.1830	0.1939
OpenBioLLM-70B	SE Probes	0.1709	0.2073

Table 2: **Best baseline** is the best result among all eight baselines for that model and dataset (SW Entropy refers to Sliding Window Entropy), with its AUPRC in **Base-line**; **RP** is Reverse Probing AUPRC (best classifier, max feature configuration). Best result per dataset in **bold**.

on all six models on MIMIC-DI, which shows that the features themselves has already carried the signal. Comparing to linear models, tree classifiers perform better, which suggest the uncertainty has complex feature interactions.

Table 1 reports results of BioMistral-7B and OpenBioLlama-8B on AUCROC and AUPRC for all methods on both datasets. The two models are pretrained by medical data. Table 2 reports Reverse Probing AUPRC across model scale and pretraining condition. Reverse Probing outperforms all baselines by a large margin on both metrics. On MIMIC-DI, where 7.3% of tokens are unsupported, it reaches 0.6022 AUPRC with Mistral-7B against 0.0918 for the strongest baseline on that model, and the margin ranges from $2.9\times$ to $6.6\times$ across the six models. XGBoost achieves an AUPRC of 0.5478 with BioMistral-7B, compared to 0.1313 for the strongest baseline (SE Probes), a roughly $4\times$ improvement. The gap is smaller but consistent on Generated-DI, where positive cases are only 2.1%, and all methods score lower on all metrics.

Most of the baselines perform poorly. On MIMIC-DI, the strongest of the eight reaches 0.13 AUPRC against a random baseline of 0.07, and several fall below 0.5 AUCROC, meaning their scores rank unsupported tokens no better than chance and sometimes in the wrong direction. The gap between baselines and Reverse Probing reflects both the difficulty of clinical text and the advantage of token-level features extracted directly from internal model signals.

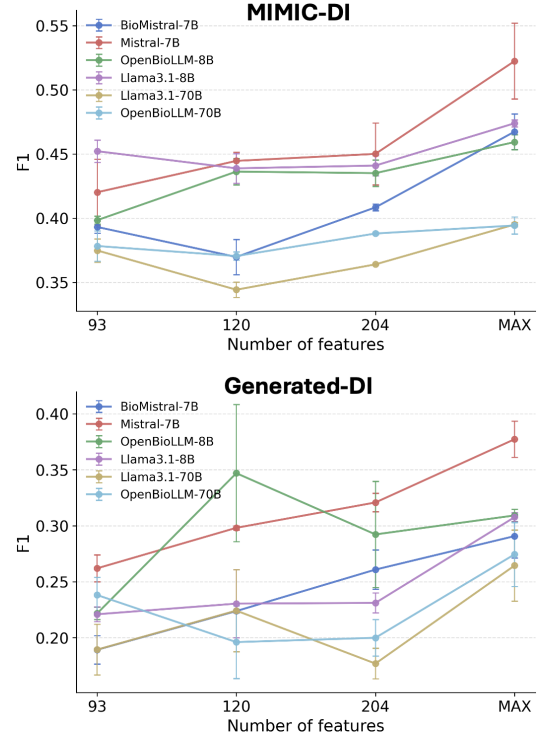


Figure 3: F1 vs. number of features on MIMIC-DI and Generated-DI. Each point is the mean over XGBoost and CatBoost. MAX denotes 454 features for 7/8B models and 886 for 70B models.

From efficiency point of view, Reverse Probing requires only two forward passes per summary: with and without BHC. This method executes 100 summaries in under 20 minutes on a single B200 GPU, compared to an average of 7 hours for multi-sampling-based baselines such as Semantic Entropy and SE Probes. This efficiency comes from operating directly on pre-existing text with only predicting one masked token each time, making it a practical candidate for real-time clinical deployment. Table 3 in Appendix A reports the full comparison.

An unexpected finding is that general-purpose models such as Mistral-7B and Llama3.1-8B sometimes outperform biomedical models such as OpenBioLLM-8B. The best result of MIMIC-DI and Generated-DI also comes from Mistral-7B, which is a smaller general domain model. An explanation is that biomedical pretraining makes models more confident in medical content, suppressing the uncertainty signals that our features rely on. An overconfident model produces flatter entropy differences and smaller delta energy, making uncertain tokens harder to distinguish from supported ones.

The results of 70B models are also lower than those of smaller models. We hypothesize that larger

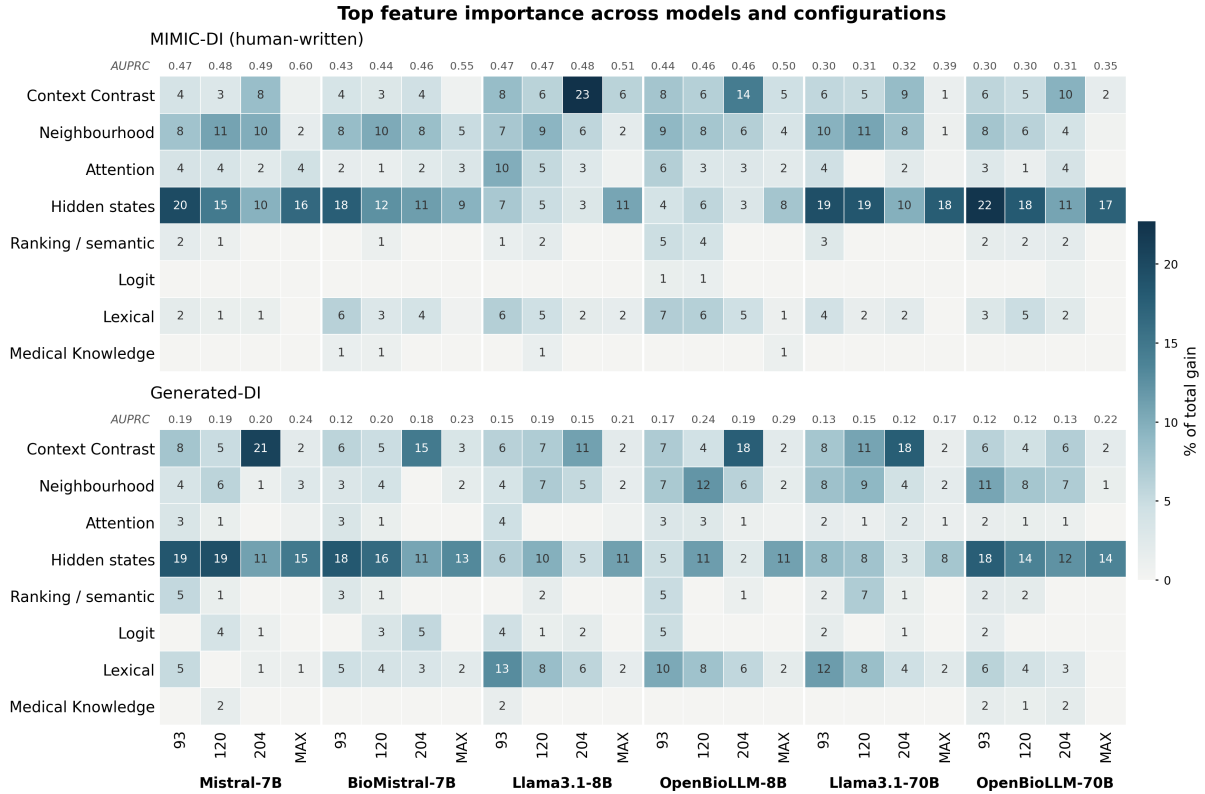


Figure 4: Feature importance across models and feature configurations, using XGBoost. Rows group features into the eight families of Section 3.2, which map onto four signal categories: distributional (context contrast, neighborhood), semantic (logit, ranking & semantic), internal (hidden states, attention), and medical knowledge (medical NER, lexical). Each column is one (model, configuration) pair, and each cell gives the share of total gain carried by that family’s features among the top 20. Darker cells carry a larger share. AUPRC for each configuration is shown above the column. MAX denotes 454 features for 7/8B models and 886 for 70B models.

models tend to produce overconfident predictions across all contexts, compressing the dynamic range of uncertainty features and making it harder for the classifier to separate unsupported from supported tokens. Additionally, larger models have more layers, and uncertainty signals may be more diffusely distributed across their representations. However, with a larger number of model layers, the sampling is effectively sparser for large models when the number of sampled features remains the same. Our current feature set samples a fixed number of layers, which may not capture enough information in 70B models compared to smaller models. This is consistent with findings in the literature showing that model scale does not reliably improve uncertainty estimation (Guo et al., 2017).

4.2 Uncertainty Source

Figure 3 shows how F1 changes with collected feature number increased on MIMIC-DI and Generated-DI. Performance generally increases with more features. From 93 to 120 features, we

added more NER-derived features with scispaCy, and add a wider neighbourhood window ($w = 7$). Despite the larger feature count, F1 drops slightly for several models at this stage, suggesting that the NER-derived entity-type and confidence signals may be noisier. Performance recovers at 204 features, while the additional context-contrast, PMI, entropy-decomposition, and medical-prior-decomposition features are added. Given that more features enhance the final result, this aligns with our hypothesis that uncertainty exists in a discrete form within the internal representations of models.

Figure 4 shows the most consistently important features across models and configurations from XGBoost. Of the four categories we defined, only two carry substantial weight: distributional signals and internal signals. Under XGBoost, `delta_energy` enters the top 20 in 47 of the 48 configurations and averages 2.65% of total gain, the highest of any single feature. The second is `neighbor_avg_w5` at 1.69%. CatBoost gives the similar result: `delta_energy` appears in all 48 con-

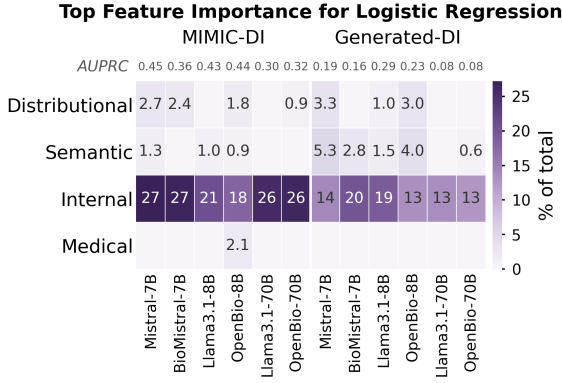


Figure 5: Top 20 recurring features by importance across models on MAX features, using Logistic Regression classifier, with contribution of the top 20 features and AUPRC above each column. AUPRC is shown above each column.

figurations with a mean share of 5.12%, close to twice the second-ranked feature `neighbor_avg_w5` (2.67%). The BHC context difference therefore holds up as the most consistent predictor across two classifiers and both datasets.

Neighborhood features rank second and stay stable across configurations. This matches prior work showing that sliding window entropy improves uncertainty estimation by incorporating local context (Sriramanan et al., 2024), and it fits the structure of clinical text, where medical concepts span several tokens. The uncertainty of a single token rarely means much on its own, and neighborhood signals let the classifier judge whether a token sits inside a broader uncertain region.

The two datasets differ in emphasis rather than in kind. Neighborhood features weigh more heavily in MIMIC-DI. This suggests that uncertainty in human-written notes clusters within local spans. Context contrast features like delta energy weigh more heavily in Generated-DI. This indicates generated content tracks how the BHC shifts the output distribution more closely.

Only the tree classifiers utilized the distributional signals. In Figure 5, Logistic regression places almost none of its weight on distributional signals, and it concentrates on internal signals. Within that group the weight concentrates on layer-wise norms and layer-to-layer change. This shows capturing the distributional signals requires a model that can represent interactions between features rather than weight them independently.

This finding also helps explain why traditional UQ methods underperformed. Methods such as

Token Entropy and Semantic Energy measure the output distribution at the final layer, which is categorized as Semantic Signals in our feature extraction process. However, in our final result, Semantic Signals contribute the least among all classifiers. Traditional methods missed the shift induced by removing the BHC, and the local structure that ties neighboring tokens together. Our results suggest that these deeper internal signals carry most of the useful uncertainty information in clinical text.

5 Conclusion

We introduce Reverse Probing, the first token-level uncertainty quantification framework for clinical text. Unlike traditional UQ methods that rely on generating new text, we show that pre-existing clinical summaries can serve as probes into the model’s internal state. By feeding existing text into a frozen LLM and taking snapshots of its internal representations, we train a classifier to detect whether hidden states and output distributions react to unsupported content. This eliminates text generation entirely, reducing inference time from 7-8 hours to under 20 minutes, and making uncertainty estimation practical where evaluating full generated summaries is infeasible. Reverse Probing consistently outperforms all adapted baselines, achieving approximately 2.9 to 6.6 times AUPRC improvement over all strongest baselines.

Feature analysis reveals that delta energy and neighborhood signals are the most stable uncertainty predictors. Delta energy captures how much BHC context shifts the entire output distribution. Neighborhood features reflect a fundamental property of clinical text: medical concepts span multiple tokens, and uncertainty rarely appears in isolation. These signals suggest that LLMs internally encode uncertainty in response to unsupported content. However, the uncertainty is distributed diffusely and interactively. By tracing predictions back to specific internal representations, Reverse Probing turns uncertainty estimation into an interpretable process, taking a step toward explainable AI in clinical settings.

Limitations

Low-resource data. Our evaluation is limited to two datasets, both derived from MIMIC-IV. To our knowledge, these are the only publicly available clinical summarization corpora with span-level annotations of unsupported facts. This kind of annota-

tion requires trained clinicians to verify each claim against the source record, which is slow, costly, and difficult to scale.

Large model adaptability. Our framework performs better on 7-8B models than to 70B models. Larger models have more layers and more diffuse internal representations, and we do not design a targeted feature extraction strategy for this scale. Adaptive layer sampling for large models is a natural direction for future work.

Single task generalization. We evaluate exclusively on discharge summaries, and it remains to be seen whether Reverse Probing generalizes to other clinical text types such as radiology reports or progress notes.

Uncertainty categorization. While our feature analysis traces uncertainty back to its source signals, we do not categorize uncertainty by type. The original datasets provide eleven fine-grained labels for unsupported facts, and it remains an open question whether different uncertainty sources systematically correspond to different error types. Linking feature-level signals to semantic categories of unsupported content would be a meaningful step toward deeper interpretability.

Ethical Considerations

This work uses Hallucinations-MIMIC-DI and Hallucinations-Generated-DI introduced by Hagselmann et al. (2025). Both datasets derive from MIMIC-IV-Note (Johnson et al., 2023), a de-identified clinical corpus. Access to MIMIC-IV-Note requires credentialed registration, completion of CITI Data or Specimens Only Research training, and a signed data use agreement through PhysioNet. All authors met these requirements before accessing any data, in compliance with the PhysioNet data use policy. No new patient data were collected. Our framework is a research tool for evaluating model uncertainty against existing expert annotations.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4190–4197.

Griffin Adams, Jason Zuckerg, and Noémie Elhadad. 2023. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. In *Ma-*

chine Learning for Healthcare Conference, pages 2–30. PMLR.

Olivier Bodenreider. 2004. *The unified medical language system (UMLS): integrating biomedical terminology*. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Tianqi Chen and Carlos Guestrin. 2016. *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1):640.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Akshat Dubey, Aleksandar Anžel, Bahar İlgen, and Georges Hattab. 2025. Ubiqtree: Uncertainty quantification in xai with tree ensembles. *Patterns*.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Stefan Hagselmann, Shannon Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. 2025. *Medical Expert Annotations of Unsupported Facts in Doctor-Written and LLM-Generated Patient Summaries*. *PhysioNet*. Version 1.0.1.

Stefan Hagselmann, Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. 2024. A data-centric approach to generate faithful and high quality patient summaries with large language models. In *Conference on Health, Inference, and Learning*, pages 339–379. PMLR.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the*

- 41st International Conference on Machine Learning, ICML'24. JMLR.org.
- Tao Huang, Rui Wang, Xiaofei Liu, Yi Qin, Li Duan, and Liping Jing. 2026. [Detecting misbehaviors of large vision-language models by evidential uncertainty quantification](#). In *The Fourteenth International Conference on Learning Representations*.
- Sujeong Hur, Yura Lee, Joongheum Park, Yeong Jeong Jeon, Jong Ho Cho, Duck Cho, Dobin Lim, Wonil Hwang, Won Chul Cha, and Junsang Yoo. 2025. Comparison of shap and clinician friendly explanations reveals effects on clinical decision behaviour. *NPJ Digital Medicine*, 8(1):578.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. [MIMIC-IV-Note: Deidentified free-text clinical notes](#). *PhysioNet*. Version 2.2.
- Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *Proceedings of the International Conference on Machine Learning (ICML)*. ArXiv:2406.15927.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations*.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [BioMistral: A collection of open-source pretrained large language models for medical domains](#). *Preprint*, arXiv:2402.10373.
- Yinghao Li, Rushi Qiang, Lama Moukheiber, and Chao Zhang. 2025. Language model uncertainty quantification with attention chain. In *Proceedings of the Conference on Language Modeling (COLM)*. ArXiv:2503.19168.
- Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. [Uncertainty estimation and quantification for llms: A simple supervised approach](#). *Preprint*, arXiv:2404.15993.
- Yingzhou Lu, Tianyi Chen, Nan Hao, Capucine Van Rechem, Jintai Chen, and Tianfan Fu. 2024. Uncertainty quantification and interpretability for clinical trial approval prediction. *Health Data Science*, 4:0126.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Hao Ma, Jiacheng Pan, Jiafei Liu, Yuchen Chen, Joey Tianyi Zhou, Gang Wang, and 1 others. 2025. Semantic energy: Detecting LLM hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. [ScispaCy: Fast and robust models for biomedical natural language processing](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.
- Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.
- Fabian Pedregosa, Ga  l Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pages 6639–6649.
- M Atif Qureshi, Abdul Aziz Noor, Awais Manzoor, Muhammad Deedahwar Mazhar Qureshi, Arjumand Younus, and Wael Rashwan. 2025. Explainability in action: A metric-driven assessment of local explanations for healthcare tabular models. *medRxiv*, pages 2025–05.
- Massimo Salvi, Silvia Seoni, Andrea Campagner, Arkadiusz Gertych, U Rajendra Acharya, Filippo Molinari, and Federico Cabitza. 2025. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics*, 197:105846.
- Ganesh Sriramanan, Sumit Bharti, Vinu Sankar Sadasivan, Srivatsa Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216.
- Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander

Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. 2025. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for LLMs. *arXiv preprint arXiv:2505.20045*.

Chenyu Wang, Weichao Zhou, Shantanu Ghosh, Kayhan Batmanghelich, and Wenchao Li. 2025. Semantic consistency-based uncertainty quantification for factuality in radiology report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1739–1754.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. Doclens: Multi-aspect fine-grained medical text evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679.

Gal Yona, Mor Geva, and Yossi Matias. 2026. [Hallucinations undermine trust; metacognition is a way forward](#). *Preprint*, arXiv:2605.01428.

A Implementation Details

All experiments are conducted on a single NVIDIA B200 GPU. Table 3 breaks down the cost. Reverse Probing requires only two forward passes per summary and processes 100 clinical summaries in less than 20 minutes. Methods requiring multiple sampling passes are more expensive: Semantic Entropy, Decomposing Uncertainty, and SE Probes each require 7-8 hours for the same 100 samples, and Uncertainty Aware Heads takes approximately 1.5 hours. Token Entropy, Semantic Energy, Sliding Window Entropy, and Attention Chain complete within a few minutes, comparable to our method, but performance is much lower.

B Example Result

Figure 6 shows a sample output of Reverse Probing on a clinical summary. Tokens flagged with uncertainty scores are highlighted in three colors: true positives (green) correctly identify tokens that are factually unsupported, false positives (red) are tokens flagged but not annotated as uncertain, and false negatives (blue) are annotated uncertain tokens that the model missed. The uncertainty score shown next to each flagged token reflects the classifier’s estimation that the model is not sure about this token.

C CatBoost Feature Contributions

Figure 7 repeats the analysis of Figure 4 using CatBoost instead of XGBoost. The two classifiers use

Method	Fwd. passes	Time
<i>Sampling-based</i>		
Semantic Entropy	$S \times N_s$	7–8 h
Decomposing Uncertainty	$S \times N_s$	7–8 h
SE Probes	$S \times N_s$	7–8 h
<i>Per-token prompting</i>		
Uncertainty Aware Heads	W	1.5 h
<i>Inference-only</i>		
Token Entropy	1	5 min
Semantic Energy	1	5 min
Sliding Window Entropy	1	10 min
Attention Chain	1	10 min
Reverse Probing	2	15+2 min

Table 3: Processing time of 100 clinical summaries on a single NVIDIA B200. N_s is the number of sampled generations per sentence, S is the number of sentences per summary and W is the number of words per summary. The 15 minutes for Reverse Probing denotes the time cost for feature collection, and the 2 minutes denotes the time for training and testing.

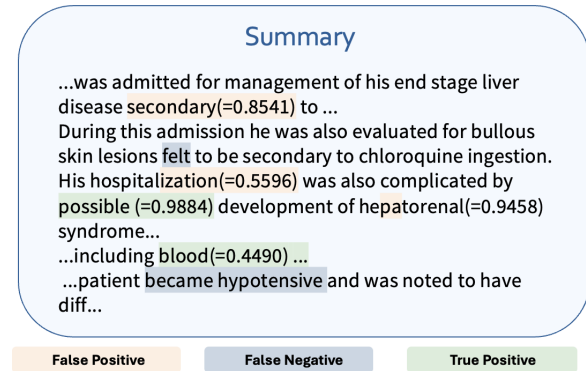


Figure 6: The ground truth labels and uncertainty quantification results for a sample clinical summary.

different definitions of importance: XGBoost reports gain, the average reduction in loss attributable to a feature across all splits, while CatBoost reports PredictionValuesChange.

The feature contribution is similar to the result from XGBoost. Distributional and internal signals carry most of the weight, semantic features contribute little, and medical NER features never enter the top 20 in any of the 48 configurations. `delta_energy` ranks first again, with a mean share of 5.12% against 2.67% for `neighbor_avg_w5`.

D Features Details

We extract token-level features from each model in four progressively richer configurations (93, 120, 204, and max: 454/886 features). To save space, each version is based on the previous one. All

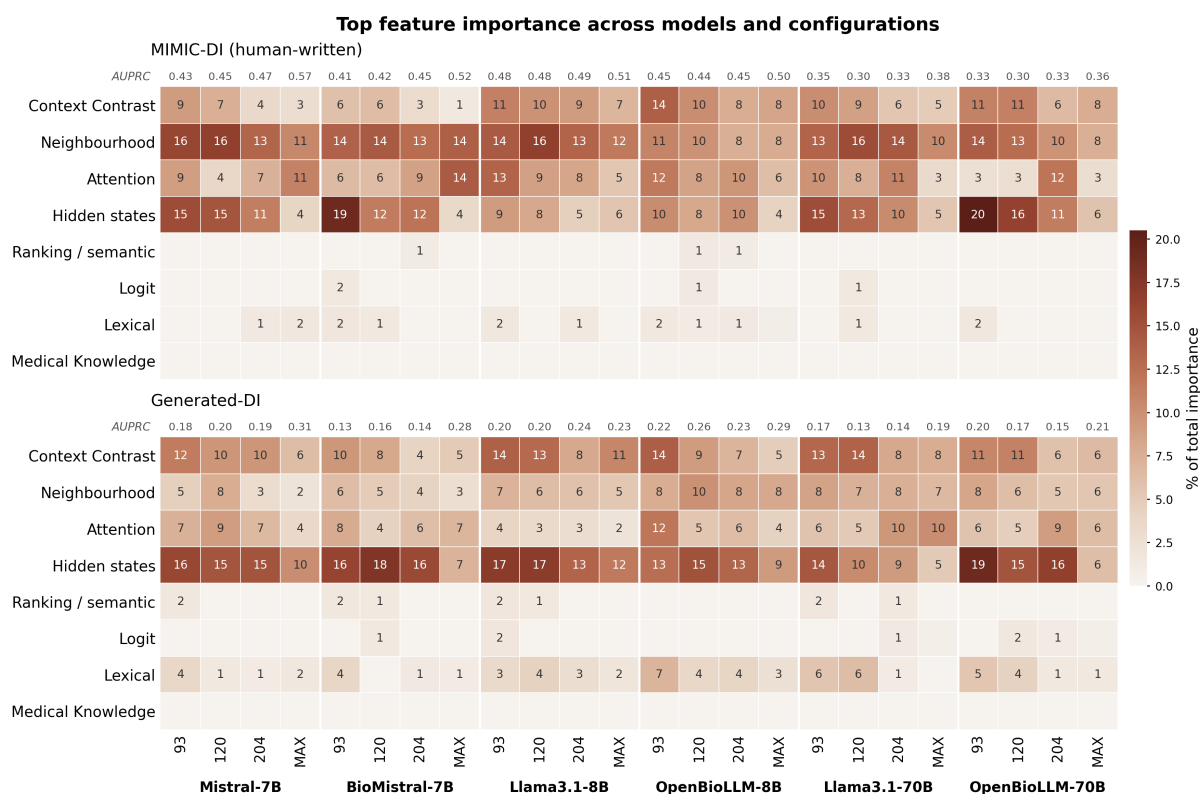


Figure 7: CatBoost Feature importance across models and feature configurations, with importance measured by PredictionValuesChange. Cells give the share of total importance carried by each family among the top 20 features. AUPRC is shown above each column. MAX denotes 454 features for 7/8B models and 886 for 70B models.

configurations share the same core groups: logit-based uncertainty signals, context-contrast features comparing distributions with and without the BHC, semantic similarity to top- k predictions, neighborhood smoothness, medical entity detection, and lexical baseline statistics. Later versions expand these groups and gradually introduce new ones. Tables 4–7 record details of each configuration.

E Full Experimental Results

This appendix reports complete results for all experiments. Table 8 presents Reverse Probing results for all six base LLMs and all configurations.

Group	Feature	Description / Formula	#
Logit (with BHC)	entropy	$H = -\sum_v p_v \log p_v$	10
	normalized_entropy	$H / \log V$	
	max_prob	$\max_v p_v$	
	current_prob	p_i	
	margin	$p_{(1)} - p_{(2)}$, top-2 probability gap	
	ratio	$p_{(1)} / p_{(2)}$	
	topk_cumulative	$\sum_{j=1}^{10} p_{(j)}$	
	gini	$\frac{2 \sum_j j p_{(j)}}{V \sum_j p_{(j)}} - \frac{V+1}{V}$, probabilities sorted descending	
	perplexity	$\exp(-\log p_i)$	
energy	$-(\max_v \ell_v + \log \sum_v e^{\ell_v - \max \ell})$		
Context Contrast (Δ BHC)	delta_prob	$p_i - p_i^-$	3
	delta_entropy	$H(\mathbf{p}) - H(\mathbf{p}^-)$	
	delta_energy	$\text{energy}(\mathbf{p}) - \text{energy}(\mathbf{p}^-)$	
Ranking & Semantic Similarity ($k \in \{5, 10, 20\}$)	rank_topk	Rank of token i among top- k predictions (0 if absent)	21
	in_topk	$1[\text{rank} > 0]$	
	max_sim_topk	$\max_j \cos(\mathbf{e}_i, \mathbf{e}_{(j)})$ over top- k tokens	
	avg_sim_topk	Mean cosine similarity to top- k tokens	
	top3_sim_topk	Mean cosine similarity to top-3 tokens	
	sim_std_topk	Std. dev. of cosine similarities to top- k	
semantic_rank_topk	First rank j with $\cos(\mathbf{e}_i, \mathbf{e}_{(j)}) > 0.7$		
Neighbourhood ($w \in \{2, 3, 5\}$)	neighbor_avg_ww	Mean p_j of the $2w$ surrounding tokens	15
	neighbor_std_ww	Std. dev. of surrounding token probabilities	
	isolation_ww	$p_i - \bar{p}_{\text{nbr}}$	
	relative_isolation_ww	$(p_i - \bar{p}_{\text{nbr}}) / (\sigma_{\text{nbr}} + \varepsilon)$	
medical_density_ww	Fraction of neighbors matching a medical keyword list		
Medical Keyword	is_medical	Binary flag: token string matches a curated medical keyword list	1
Lexical & Baseline	freq	Raw token frequency in the corpus	7
	freq_normalized	freq / total tokens	
	freq_log	$\log(\text{freq} + 1)$	
	idf	$\log(N / (df + 1))$, N = number of documents	
	rarity	$1 / (\text{freq} + 1)$	
	baseline_prob	p_i computed from a minimal BOS-only context	
baseline_entropy	H computed from a minimal BOS-only context		
Hidden States (4 layers)	hidden_norm_l ℓ	$\ h_\ell\ _2$	12
	hidden_mean_l ℓ	$\text{mean}(h_\ell)$	
	hidden_std_l ℓ	$\text{std}(h_\ell)$	
	4 layers sampled at uniform intervals. 8B: layers {8, 16, 24, 31} (every 8 layers); 70B: layers {20, 40, 60, 79} (every 20 layers).		
Hidden State Change	layer_change_l ℓ_1 _to_l ℓ_2	$\ h_{\ell_2} - h_{\ell_1}\ _2$ between consecutive sampled layers	6
	layer_cosine_l ℓ_1 _to_l ℓ_2	$\cos(h_{\ell_1}, h_{\ell_2})$	
	3 consecutive pairs from the 4 sampled layers above.		
Attention Snapshots (6 heads)	attn_entropy_l ℓ _hh	$H(\alpha_{\ell,h,i,\cdot})$, entropy of the attention row	18
	attn_to_bhc_l ℓ _hh	$\sum_{j \leq L_{\text{BHC}}} \alpha_{\ell,h,i,j}$, total attention weight to BHC tokens	
	attn_max_l ℓ _hh	$\max_j \alpha_{\ell,h,i,j}$	
	6 (layer, head) pairs sampled to cover shallow, mid, and deep layers. 8B: (7, 16), (15, 24), (23, 0), (23, 9), (23, 28), (31, 8); 70B: (7, 16), (17, 24), (57, 0), (57, 9), (57, 28), (77, 8). The sampling count is identical; indices are scaled to model depth and head count.		
Total			93

Table 4: 93 token-level features. Parameters that differ between models are noted in the Description column. Notation: p_i = softmax probability of the actual token i ; $\mathbf{p}$ = full distribution over vocabulary V ; ℓ_i = logit of token i ; $\mathbf{p}^-$ = distribution computed *without* BHC context; h_ℓ = hidden state vector at layer ℓ ; $\alpha_{\ell,h}$ = attention weights at layer ℓ , head h .

Group	Feature	Description / Formula	#
Logit (with BHC)	Same 10 features as Table 4.		10 / 10
Context Contrast (Δ BHC)	Same 3 features as Table 4.		3 / 3
Ranking & Semantic Similarity ($k \in \{5, 10, 20\}$)	Same 21 features as Table 4.		21 / 21
Neighbourhood ($w \in \{2, 3, 5, 7\}$)	neighbor_avg_ww neighbor_std_ww isolation_ww relative_isolation_ww medical_density_ww	Mean p_j of the $2w$ surrounding tokens Std. dev. of surrounding token probabilities $p_i - \bar{p}_{\text{nbr}}$ $(p_i - \bar{p}_{\text{nbr}}) / (\sigma_{\text{nbr}} + \varepsilon)$ Fraction of neighbours detected as medical terms (NER)	20 / 20
Medical NER (scispaCy + UMLS/MedDRA)	is_medical ner_entity_type medical_confidence	Binary: token detected by NER or vocabulary Integer-encoded entity type (12 classes: O, CHEMICAL, DISEASE, ...) Source confidence: NER+vocab = 1.0, NER = 0.85, vocab = 0.6	3 / 3
Lexical & Baseline	Same 7 features as Table 4.		7 / 7
Hidden States (8 layers)	hidden_norm_ll hidden_mean_ll hidden_std_ll	$\ h_\ell\ _2$ $\text{mean}(h_\ell)$ $\text{std}(h_\ell)$	24 / 24
	8 layers sampled at uniform intervals ($\approx$ every 4 layers for 8B, every 10 for 70B). 8B: {2, 6, 10, 14, 18, 22, 26, 31}; 70B: {5, 15, 25, 35, 45, 55, 65, 79}.		
Hidden State Change (7 pairs)	layer_change_ll1_to_ll2 layer_cosine_ll1_to_ll2	$\ h_{\ell_2} - h_{\ell_1}\ _2$ between consecutive sampled layers $\cos(h_{\ell_1}, h_{\ell_2})$	14 / 14
	7 consecutive pairs from the 8 sampled layers above.		
Attention Snapshots (6 heads)	attn_entropy_ll_hh attn_to_bhc_ll_hh attn_max_ll_hh	$H(\alpha_{\ell,h,i,\cdot})$, entropy of the attention row $\sum_{j \leq L_{\text{BHC}}} \alpha_{\ell,h,i,j}$, total attention weight to BHC $\max_j \alpha_{\ell,h,i,j}$	18 / 18
	6 (layer, head) pairs covering shallow, mid, and deep layers. 8B: (7, 16), (15, 24), (23, 0), (23, 9), (23, 28), (31, 8); 70B: (37, 24), (17, 16), (57, 0), (57, 9), (57, 28), (77, 8). Layer indices scaled proportionally; head count identical.		
Total			120 / 120

Table 5: Token-level features (120 total) extracted using **Llama-3.1-8B-Instruct** (32 layers, 32 attention heads) and **Llama3-OpenBioLLM-70B** (80 layers, 64 attention heads). Groups identical to Table 4 are abbreviated; only modified or new groups are shown in full. Compared to the 93-feature set: (1) Increased window size 7; (2) medical detection upgraded from keyword matching to scispaCy NER with 3 output features; (3) hidden-state sampling expanded to 8 layers with 7 inter-layer change pairs. Notation as in Table 4; layer/head indices that differ between models are listed in the notes rows.

Group	Added / Changed	Description / Formula	120 / 204
<i>Unchanged from Table 5: logit (10), ranking & semantic (21), neighbourhood (20), lexical & baseline (7), hidden states (24), hidden state change (14), attention snapshots (18).</i>			114 / 114
Context Contrast	kl_with_vs_no, kl_no_vs_with, js_divergence, rank_with_bhc, rank_without_bhc, delta_rank, topk_overlap, topk_jaccard, delta_margin, prob_vs_no_max	Divergence, rank shift, and top- k overlap between $\mathbf{p}$ and $\mathbf{p}^-$, complementing the three Δ features carried over from Table 5	3 / 13
Medical NER	6 one-hot entity-type flags, ner_is_high_risk	Adds entity-type-specific flags (chemical, disease, gene, cancer, anatomy, pathology) and a high-risk composite flag on top of the is_medical / ner_entity_type / medical_confidence features already present in Table 5	3 / 10
PMI (<i>new</i>)	pmi, pmi_vs_prior, npmi, cpmi	$\log p_i^{\text{with}} - \log p_i^{\text{no}}$; variants against the language prior, length-normalised, and standardised within entity type	0 / 4
Entropy Decomposition (<i>new</i>)	bhc_info_gain, ctx_info_gain, total_info_gain, bhc_contribution_ratio, bhc_info_gain_norm	Separates the entropy reduction attributable to the BHC from that of the surrounding context, using $H(\mathbf{p}^{\text{prior}})$, $H(\mathbf{p}^-)$, $H(\mathbf{p})$	0 / 5
Medical Prior Decomposition (<i>new</i>)	halluc_risk_ratio, patient_specificity, context_reliance_score, prob_dominance_order, patient_specificity_norm	Distinguishes tokens supported by patient-specific evidence from those recoverable from medical prior alone	0 / 5
			0 / 44
Attention Drift (<i>new</i>)	attn_drift_l1_l2_hh	$D_{\text{KL}}(\tilde{\alpha}_{\ell_1, h, i} \parallel \tilde{\alpha}_{\ell_2, h, i})$ between adjacent-layer attention rows; captures sudden shifts in grounding	
<i>Three bands of 4 consecutive layers (shallow, mid, deep) give 12 layers and 11 adjacent pairs; $11 \times 4 \text{ heads} = 44$. 8B: $\{0-3\} \cup \{14-17\} \cup \{28-31\}$, $h \in \{0, 8, 16, 24\}$. 70B: $\{0-3\} \cup \{35-38\} \cup \{76-79\}$, $h \in \{0, 16, 32, 48\}$.</i>			0 / 9
Attention Rollout (<i>new</i>)	rollout_to_bhc_l ℓ , rollout_entropy_l ℓ , rollout_max_weight_l ℓ	$\mathbf{R}_\ell = \prod_{k=0}^{\ell} (0.5 \mathbf{A}_k + 0.5 \mathbf{I})$, row-normalised at each step; cumulative attention to BHC, its entropy, and its max weight	
<i>One checkpoint at the last layer of each band (3 checkpoints $\times$ 3 features). 8B: $\ell \in \{3, 17, 31\}$; 70B: $\ell \in \{3, 38, 79\}$.</i>			
Total			120 / 204

Table 6: The 204-feature set changed from the 120-feature set of Table 5. Groups marked (*new*) did not exist in the 120-feature set (hence the 0 in the left-hand count) and were introduced for the first time here. $\mathbf{p}$, $\mathbf{p}^-$, and $\mathbf{p}^{\text{prior}}$ denote the output distribution with the BHC, without the BHC, and from a minimal BOS-only context. Note that delta_entropy and bhc_info_gain are sign-inverted forms of the same quantity.

Group	Feature	Description / Formula	#
Logit / Contrast / Ranking / Neighbourhood / Medical NER / Lexical / PMI / Entropy Decomp. / Medical Prior Decomp.	Same as Table 6: $54 + 10 + 7 + 3 + 5 + 5 = 84$ features.		84 / 84
Hidden States (all L layers, expanded)	hidden_norm_1 ℓ	$\ h_\ell\ _2, \ell = 0, \dots, L-1$	96 / 240
	hidden_mean_1 ℓ	$\text{mean}(h_\ell)$	
	hidden_std_1 ℓ	$\text{std}(h_\ell)$	
Expanded from 8 sampled layers to all L layers. 8B: $32 \times 3 = 96$; 70B: $80 \times 3 = 240$.			
Hidden State Change (all $L-1$ pairs, expanded)	layer_change_1 ℓ _to_1($\ell+1$)	$\ h_{\ell+1} - h_\ell\ _2$	62 / 158
	layer_cosine_1 ℓ _to_1($\ell+1$)	$\cos(h_\ell, h_{\ell+1})$	
	Expanded from 7 sampled pairs to all consecutive pairs. 8B: $31 \times 2 = 62$; 70B: $79 \times 2 = 158$.		
Attention Snapshots (6 heads)	Same 18 features as Table 6.		18 / 18 124 / 316
Attention Drift (4 heads, all $L-1$ pairs, expanded)	attn_drift_1 ℓ _1($\ell+1$)_hh	$D_{\text{KL}}(\tilde{\alpha}_{\ell,h,i} \parallel \tilde{\alpha}_{\ell+1,h,i})$, KL divergence between adjacent-layer attention rows	
	Expanded from 3 bands to all adjacent layer pairs. 4 heads at uniform intervals. 8B: $h \in \{0, 8, 16, 24\}$, 31 pairs = 124; 70B: $h \in \{0, 16, 32, 48\}$, 79 pairs = 316.		
Attention Rollout (8 checkpoints, expanded)	rollout_to_bhc_1 ℓ	$\sum_{j < L_{\text{BHC}}} [\mathbf{R}_\ell]_{i,j}$, cumulative attention to BHC	24 / 24
	rollout_entropy_1 ℓ	$H(\mathbf{R}_\ell^{[i, \cdot]})$	
	rollout_max_weight_1 ℓ	$\max_j [\mathbf{R}_\ell]_{i,j}$	
$\mathbf{R}_\ell = \prod_{k=0}^{\ell} (0.5 \mathbf{A}_k + 0.5 \mathbf{I})$; expanded from 3 to 8 uniformly spaced checkpoints (8B: every 4 layers; 70B: every 10 layers).			
Total			454 / 886

Table 7: Token-level features extracted using **Llama-3.1-8B-Instruct (454 features, 32 layers, 32 attention heads)** and **Llama3-OpenBioLLM-70B (886 features, 80 layers, 64 attention heads)**. Groups unchanged from Table 6 are abbreviated. Compared to the 204-feature set, three groups are expanded to cover all model layers: (1) hidden-state features extended from 8 sampled layers to all L layers; (2) hidden-state change features extended from 7 sampled pairs to all $L-1$ consecutive pairs; (3) attention drift extended from 3 depth bands to all adjacent layer pairs; (4) rollout checkpoints increased from 3 to 8. The increase in total features (204→454/886) is entirely due to these layer-coverage expansions. Notation as in Table 4.

Config	XGBoost			CatBoost		
	F1	AUCROC	AUPRC	F1	AUCROC	AUPRC
OpenBioLLM-8B / MIMIC-DI						
93	0.3953	0.8598	0.4418	0.4015	0.8645	0.4485
120	0.4466	0.8781	0.4615	0.4259	0.8643	0.4445
204	0.4454	0.8774	0.4600	0.4247	0.8684	0.4525
MAX (454)	0.4651	0.8833	0.4976	0.4534	0.8848	0.5029
OpenBioLLM-8B / Generated-DI						
93	0.2162	0.8714	0.1719	0.2273	0.8818	0.2194
120	0.2857	0.9046	0.2421	0.4082	0.8739	0.2588
204	0.2449	0.8983	0.1897	0.3396	0.8773	0.2277
MAX (454)	0.3038	0.9364	0.2892	0.3146	0.9144	0.2895
Llama3.1-8B / MIMIC-DI						
93	0.4608	0.8657	0.4742	0.4435	0.8527	0.4842
120	0.4272	0.8814	0.4732	0.4504	0.8659	0.4808
204	0.4488	0.8747	0.4763	0.4331	0.8724	0.4858
MAX (454)	0.4712	0.8967	0.5110	0.4766	0.8755	0.5140
Llama3.1-8B / Generated-DI						
93	0.2276	0.9133	0.1496	0.2143	0.8481	0.1956
120	0.2000	0.9228	0.1854	0.2609	0.9136	0.2033
204	0.2222	0.9229	0.1479	0.2600	0.9340	0.2761
MAX (454)	0.3117	0.9365	0.2121	0.3036	0.9245	0.2253
OpenBioLLM-70B / MIMIC-DI						
93	0.3665	0.8545	0.3023	0.3900	0.8658	0.3260
120	0.3672	0.8512	0.2991	0.3737	0.8562	0.3023
204	0.3890	0.8639	0.3125	0.3870	0.8585	0.3281
MAX (886)	0.3876	0.8710	0.3500	0.4009	0.8706	0.3597
OpenBioLLM-70B / Generated-DI						
93	0.2222	0.8089	0.1249	0.2540	0.8262	0.2008
120	0.1636	0.8130	0.1208	0.2286	0.8218	0.1725
204	0.1837	0.8141	0.1277	0.2162	0.7998	0.1548
MAX (886)	0.3030	0.7985	0.2162	0.2459	0.8117	0.2073
Llama3.1-70B / MIMIC-DI						
93	0.3657	0.8407	0.2965	0.3839	0.8494	0.3528
120	0.3382	0.8408	0.3057	0.3502	0.8337	0.2964
204	0.3643	0.8524	0.3233	0.3636	0.8533	0.3268
MAX (886)	0.3932	0.8594	0.3934	0.3968	0.8563	0.3835
Llama3.1-70B / Generated-DI						
93	0.1667	0.8344	0.1252	0.2121	0.8055	0.1679
120	0.2609	0.8226	0.1486	0.1875	0.7657	0.1322
204	0.1633	0.7877	0.1247	0.1905	0.8179	0.1378
MAX (886)	0.2963	0.8132	0.1697	0.2326	0.7999	0.1939
BioMistral-7B / MIMIC-DI						
93	0.3882	0.8293	0.4297	0.3980	0.8216	0.4126
120	0.3834	0.8482	0.4366	0.3560	0.8328	0.4197
204	0.4057	0.8615	0.4562	0.4110	0.8512	0.4543
MAX (454)	0.4812	0.8911	0.5478	0.4535	0.8691	0.5178
BioMistral-7B / Generated-DI						
93	0.2020	0.8192	0.1170	0.1765	0.7856	0.1254
120	0.2222	0.8364	0.1997	0.2254	0.7979	0.1623
204	0.2785	0.8318	0.1833	0.2432	0.8062	0.1442
MAX (454)	0.3103	0.8819	0.2316	0.2712	0.8435	0.2805
Mistral-7B / MIMIC-DI						
93	0.4459	0.8480	0.4654	0.3944	0.8260	0.4328
120	0.4514	0.8598	0.4818	0.4378	0.8372	0.4545
204	0.4742	0.8594	0.4894	0.4259	0.8493	0.4706
MAX (454)	0.5519	0.9023	0.6022	0.4928	0.8937	0.5737
Mistral-7B / Generated-DI						
93	0.2500	0.8069	0.1938	0.2740	0.8178	0.1799
120	0.3000	0.8162	0.1878	0.2963	0.8397	0.1960
204	0.3125	0.8164	0.2005	0.3291	0.8030	0.1932
MAX (454)	0.3611	0.8476	0.2373	0.3934	0.8539	0.3070

Table 8: Reverse Probing full results across all models.