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

Anonymous ACL submission

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 UQ framework specialized for clinical summarization, which estimates token-level uncertainty directly from pre-existing labeled summaries. Rather than sampling new outputs, Reverse Probing treats the text as a probe into the model's internal state, extracting uncertainty signals from four categories of internal activations. We evaluate on two expert-annotated clinical datasets and outperform eight adapted baselines on all metrics, achieving up to $4\times$ higher AUPRC while reducing inference time and computational costs. Feature analysis reveals that delta energy and 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 sequence 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 correct, 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) formalize this distinction, redefining hallucination as a confident error and arguing that models should express honest uncertainty rather than optimize 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 masked language models like BERT (Devlin et al., 2019), which evaluate confidence by predicting a masked token within a bidirectional context. Decoder-only LLMs operate autoregressively and do not natively support such token-masking evaluation. We bridge this architectural gap with a prompt-based masking strategy. We sequentially replace each target token with a mask placeholder and prompt the model to predict it from the context. The name *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 map 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 uncertainty 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 Probing, 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 in clinical summaries. We conduct a systematic evaluation comparing Reverse Probing against eight baselines across six LLMs on two expert-annotated datasets.
- We provide interpretable analysis of uncertainty sources, finding that delta energy and neighborhood context are the most universal predictors across all models and 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 quantified from the variance or distribution of the 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.

Liu et al. (2024) show that hidden activations encode uncertainty information, and Vazhentsev et al. (2025) show that some attention heads are correlated to incorrect generations. Guo et al. (2017) further motivates this direction by showing that output probabilities alone are unreliable due to systematic overconfidence. Despite these advances, all existing methods are designed and evaluated on general-domain tasks. No method addresses uncertainty in long-form medical summaries, where errors are sparse, terminology is specialized, and the cost of multi-sampling is high.

2.2 Clinical Explainability and Faithfulness

Explainability methods have been widely applied to clinical decision support, but rarely to 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 judgement (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 an uncertainty 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 vast majority of tokens are factually supported, while unsupported content appears only in sparse, scattered spans. This makes fine-grained UQ fundamentally different from task-level evaluation, where standard metrics are too coarse to capture localized errors.

Our approach is inspired by masked language model probing. BERT-style models support token masking, but decoder-only LLMs are forwarding models by design. We bridge this gap with a prompt-based masking strategy: we replace each target token with a mask placeholder and prompt the model to predict it from context, turning a generative model into a token-level probe without modifying its weights.

The key insight behind Reverse Probing is that the BHC serves as patient-specific evidence. When a summary token is well-supported, the model's attention anchors to the BHC. When a token is unsupported, this shift is absent or misdirected. By comparing activations with and without the BHC, we use the absence of grounding as one of the most important signals. This is precisely where the "reverse" in Reverse Probing comes from. While traditional methods generate new text to check for

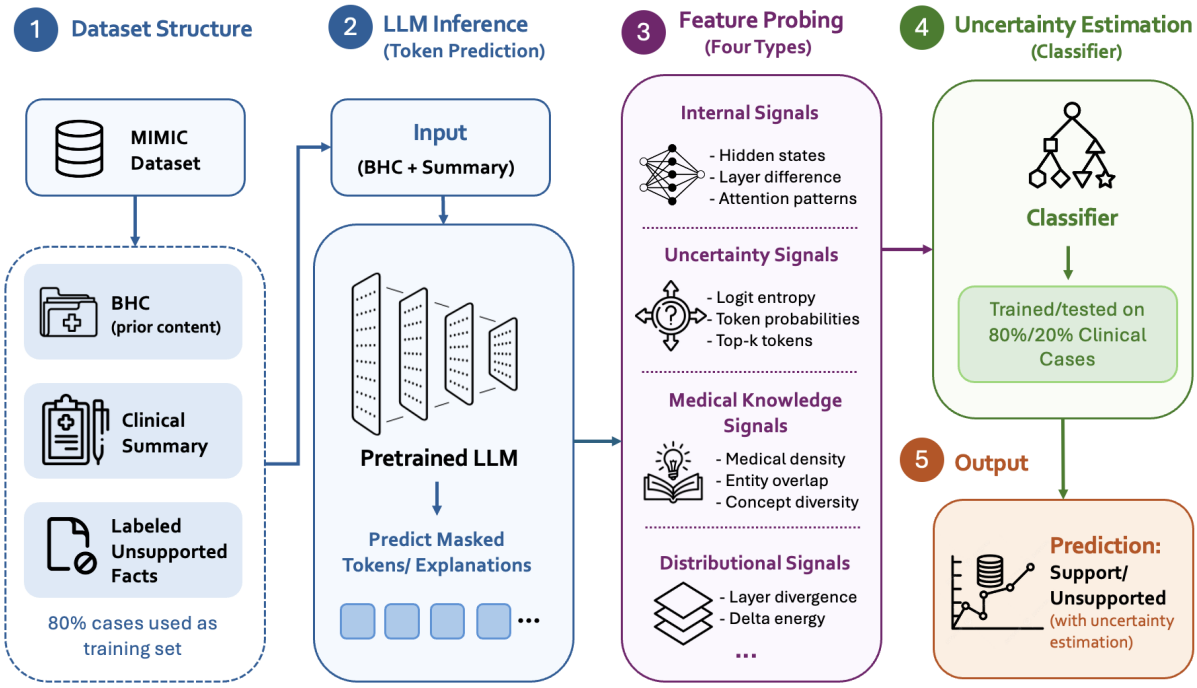

Figure 1: Overview of the Reverse Probing framework. The BHC and clinical summary are fed into a frozen LLM, which extracts four categories of features from its internal representations. A supervised classifier is then trained to predict token-level uncertainty.

factual accuracy, our approach does the opposite: it uses the existing text as a probe to 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 Brief Hospital Course (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 Hegselmann 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 sample consists of a Brief Hospital Course (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 LLM-generated summaries (Hegselmann et al., 2024), while the BHC from both MIMIC-DI and Generated-DI are originated 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. The low ratio of unsupported spans in both datasets reflects severe class imbalance at the token level, which adds considerable difficulty to the classification task.

Both 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 1.14 spans per summary.

3.2 Feature Extraction

We propose four types of sources that potentially affect uncertainty inside LLM's internal representations, output distributions, domain-specific knowledge, and behavior under context perturbation. To study how feature granularity affects performance, we define four configurations of increasing detail: 93, 120, 204, and up to 454 or 886 features. All models are evaluated on the 93, 120, and 204 con-

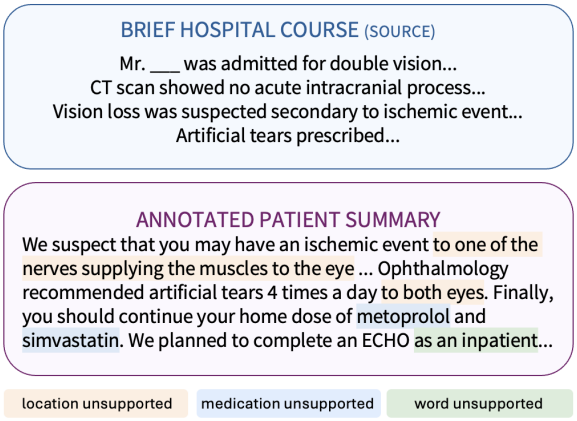

Figure 2: An example from the dataset Hallucinations-MIMIC-DI. Highlighted spans indicate unsupported facts, categorized by label type. The dataset structure for Hallucinations-Generated-DI is exact the same as the Hallucinations-MIMIC-DI.

figurations. The difference in the largest configuration reflects model architecture: we collect 454 as the maximum features for 7B and 8B models, while the 70B model used 886 due to its 80 layers. Complete details of all features are provided in Section C.

Internal Signals. For each token at position i, we extract the ℓ_2 norm, mean, and standard deviation of the hidden state $\mathbf{h}_i^{(l)}$ at selected layers l. We measure layer-to-layer change using L2 distance and cosine similarity between consecutive hidden states.

Layer-to-layer change is measured by:

$$\text{LayerChange}_{l\to l'} = \|\mathbf{h}_i^{(l')} - \mathbf{h}_i^{(l)}\|_2,$$
$$\text{LayerCosine}_{l\to l'} = \frac{\mathbf{h}_i^{(l)} \cdot \mathbf{h}_i^{(l')}}{\|\mathbf{h}_i^{(l)}\|\|\mathbf{h}_i^{(l')}\|} \quad (1)$$

For attention, we compute per-head entropy and the total attention weight directed toward the BHC prefix at selected layer-head pairs. We also measure inter-layer attention drift as the KL divergence between the same head's attention distribution at adjacent layers. Attention rollout accumulates attention matrices with residual connections across layers and takes snapshots at shallow, middle, and deep checkpoints to track how BHC information propagates through the network.

Uncertainty Signals. These features summarize the shape of $P(t_i \mid \text{BHC}, t_{<i})$ from the with-context forward pass. Given the logit vector $\mathbf{z}$ and probability distribution $p = \text{softmax}(\mathbf{z})$, we compute Shannon entropy $H = -\sum_v p_v \log p_v$, the margin between the top two candidates $p_{(1)} - p_{(2)}$, and the Gini coefficient of the probability distribution. We also compute a free-energy score $\mathcal{E} = -\log \sum_v \exp(z_v)$, which captures overall model confidence without softmax normalization. Together, these features reflect how concentrated or diffuse the model's prediction is at each position.

Medical Knowledge Signals. These features encode domain-specific context around each token. We run named entity recognition offline using scispaCy (Neumann et al., 2019) with UMLS/MedDRA (Bodenreider, 2004) vocabulary, which yields entity type labels and confidence scores. Medical density counts the fraction of surrounding tokens identified as medical entities within multiple window sizes. Corpus-level token frequency and IDF come from the training set. We also compute conditional PMI (CPMI) by normalizing PMI within entity type groups, which corrects for the tendency of certain entity types such as chemicals to carry systematically higher raw PMI values.

Distributional Signals. We run two forward passes per token: one with the BHC prepended and one without. Let p^+ and p^- denote the two resulting distributions. We compute the change in token probability $\Delta p = p^+_{t_i} - p^-_{t_i}$, entropy $\Delta H = H^+ - H^-$, and free energy:

$$\Delta\mathcal{E} = \mathcal{E}^+ - \mathcal{E}^-, \quad \mathcal{E}^\pm = -\log \sum_v \exp(z_v^\pm) \quad (2)$$

A large positive $\Delta\mathcal{E}$ means the BHC makes the output distribution less peaked, so the token is less confidently supported by the clinical record. We further compute KL and Jensen-Shannon divergence between the two distributions. We compute PMI as $\text{PMI} = \log p^+_{t_i} - \log p^-_{t_i}$, and a normalized variant $\text{nPMI} = \text{PMI}/-\log p^+_{t_i}$, where we use $-\log p^+_{t_i}$ as the normalizer rather than the joint probability, as the latter is unavailable in this setting. We also compute BHC information gain $\Delta H_{\text{BHC}} = H^- - H^+$, which measures how much the clinical record reduces entropy at each position. Note that this runs in the opposite direction to $\Delta H = H^+ - H^-$: a positive ΔH_{BHC} means the BHC lowers entropy, so the token is more confidently supported. Finally, neighborhood features collect token probabilities within sliding windows of multiple sizes around position i. Tokens within the same unsupported span tend to share uncertainty patterns, so these signals also help detect whether a suspicious token is part of a broader problematic region.

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, train two optimized classifiers: XGBoost (Chen and Guestrin, 2016) and CatBoost (Prokhorenkova et al., 2018) on extracted features to predict uncertainty source according to the change of contexts.

Data is split at the document level, treating each BHC–summary pair as one instance, 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 Sankarasubbu, 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 adapt eight general-domain baselines to our setting: Token Entropy (Ma et al., 2025), Semantic Energy (Ma et al., 2025), Uncertainty Aware Heads (Vazhentsev et al., 2025), Attention Chain (Li et al., 2025), SE 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 (Semantic Entropy) Probe 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 micro F1, AUCROC, and AUPRC for unsupported tokens across all documents. 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. Rather than evaluating whether an output is factually incorrect, we measure how models' internal representations activate when encountering unsupported facts.

Table 1 reports AUCROC and AUPRC for all methods on both datasets. Reverse Probing outperforms all baselines by a large margin on both metrics. On MIMIC-DI (Human-Written), 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 class imbalance is more severe, and all methods score lower on AUPRC. Our two classifiers, XGBoost and CatBoost perform comparably.

Although our AUCROC values are high (up to 0.89 on MIMIC-DI and 0.94 on Generated-DI), this partly reflects the severe class imbalance in both datasets: only 7.3% of tokens in MIMIC-DI

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	0.5361	0.1050
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)	0.8691	0.5178	**0.8848**	**0.5029**	0.8435	**0.2805**	0.9144	**0.2895**

Table 1: Main results on MIMIC-DI and Generated-DI. Reverse Probing uses the max feature configuration (454 features for 7–8B models, 886 for 70B models). Baselines are adapted from general-domain methods as described in Section 3.4. Best results per column are **bold**; second best are underlined.

are annotated as unsupported, and just 2.05% in Generated-DI. AUCROC inflates because the true negatives dominate the score. AUPRC is more informative here, and our results show consistent improvement on this metric across all models and configurations.

Notably, even the adapted baselines achieve non-trivial scores despite being designed for generated answers for the general domain. This suggests that pre-existing clinical text can also be used to access uncertainty signals, and the core mechanisms of general-domain UQ methods remain partially effective when adapted to this setting. 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.

Reverse Probing also requires only two forward passes per token: with and without BHC. This method executes 100 samples in under 10 minutes on a single B200 GPU, compared to average 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.

Figure 3 extends this comparison to all six base LLMs. Reverse Probing maintains a clear and consistent advantage across models of varying size and biomedical pretraining, while baseline performance varies substantially across LLMs. This suggests that Reverse Probing is robust to different base models.

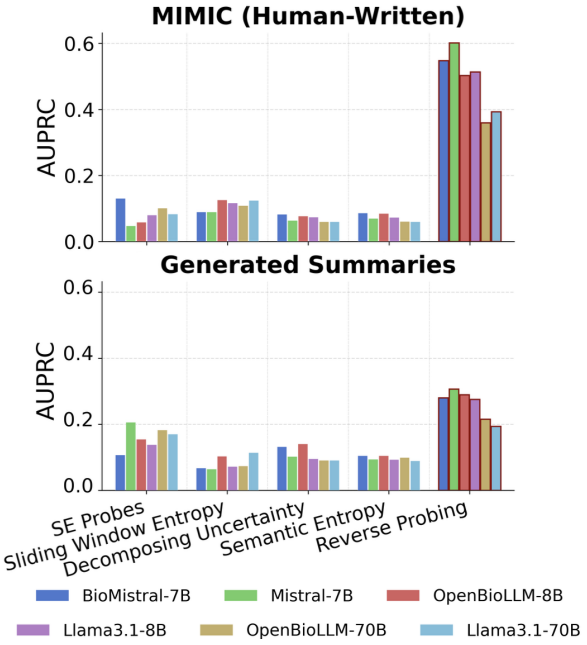

Figure 3: AUPRC across base LLMs for four selected baselines and Reverse Probing. Each bar represents one base LLM using the max feature configuration (454 for 7-8B models and 886 for 70B models).

Table 2 reports Reverse Probing AUPRC across model scale and pretraining condition. Smaller 7-8B models consistently outperform 70B models on both datasets, with Mistral-7B achieving the highest AUPRC of 0.6022 on MIMIC-DI and 0.3070 on Generated-DI.

An unexpected finding is that general-purpose models such as Mistral-7B and Llama3.1-8B sometimes match or outperform biomedical counterparts such as OpenBioLLM-8B. A likely explanation is that biomedical pretraining makes models more

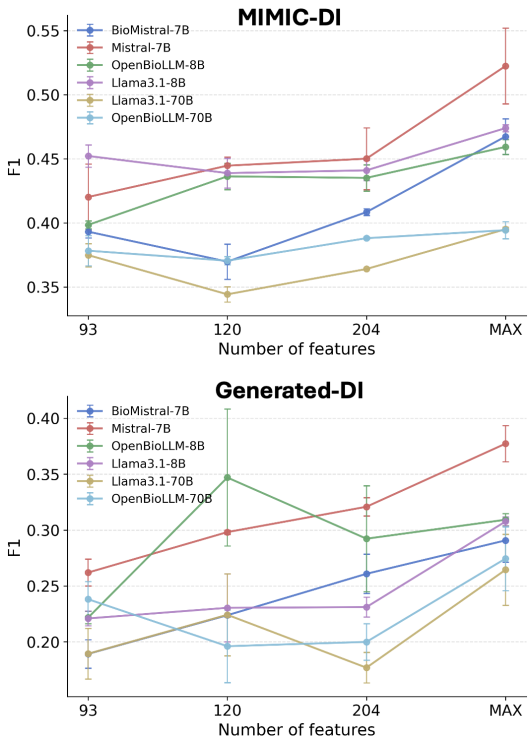

Figure 4: F1 vs. number of features on MIMIC-DI and Generated-DI. Each point is the mean over XGBoost and CatBoost.

Model	MIMIC-DI	Generated-DI
Mistral-7B	**0.6022**	**0.3070**
BioMistral-7B	0.5478	0.2805
Llama3.1-8B	0.5140	0.2761
OpenBioLLM-8B	0.5029	0.2895
Llama3.1-70B	0.3934	0.1939
OpenBioLLM-70B	0.3597	0.2073

Table 2: Reverse Probing AUPRC across model scale and pretraining condition (best classifier, best feature configuration).

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 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. Our current feature set samples a fixed number of layers, which may not capture enough of this spread in 70B models compared to their smaller counterparts. 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 4 shows how F1 changes with feature count on MIMIC-DI and Generated-DI. Performance generally increases with more features, with the largest gain between 204 and the maximum configuration. Compared to the 93-feature configuration, the 120-feature set removes the widest neighbourhood window ($w = 9$). These changes temporarily reduce the signal available to the classifier, and the additional hidden state features introduced at this stage are not yet sufficient to compensate. Performance recovers at 204 and beyond, once denser layer sampling and wider context windows restore and surpass the original signal coverage. Generated-DI consistently yield lower F1 than MIMIC-DI across all configurations. This is due to the more severe class imbalance and fewer labels.

Figure 5 shows the most consistently important features across models and configurations. Despite human-defined four feature categories, only two dominate in practice: Distributional signals and Medical Knowledge signals. Features such as `delta_energy`, `neighbor_avg`, and `delta_entropy` appear reliably across all models and configurations. Neighborhood features are particularly stable, which aligns with prior work showing that sliding window entropy improves uncertainty estimation by incorporating local context (Sriramanan et al., 2024). This is especially relevant in clinical text, where medical concepts are typically expressed as multi-token spans: the uncertainty of a single token is rarely meaningful in isolation, and neighborhood signals help the classifier detect whether a token belongs to a broader uncertain region. The two datasets differ subtly in emphasis: in MIMIC-DI, neighborhood features such as `neighbor_avg_w5` and `neighbor_std_w7` carry the most weight, suggesting that uncertainty in human-written notes clusters within local spans; in Generated-DI, distributional contrast features `delta_energy` and `delta_prob` become more prominent. This reflects that model-generated errors are more strongly tied to how the BHC shifts the output distribution than to local token patterns.

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

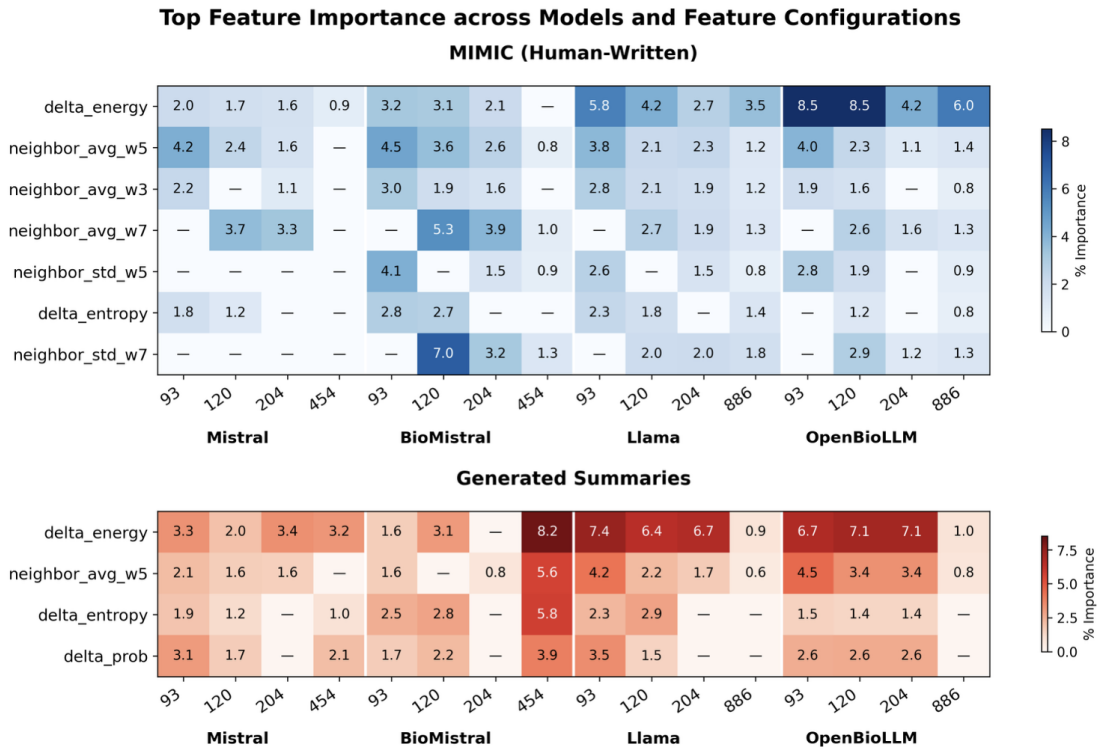

Figure 5: Top recurring features by importance across models and feature configurations. Blue cells indicate MIMIC-DI (Human-Written); red cells indicate Generated-DI. Darker color indicates higher importance (% of total). Dashes indicate the feature did not appear in the top 20 for that configuration.

Token Entropy and Semantic Energy measure the output distribution at the final layer. However, they missed the richer information in hidden states, layer transitions, and attention patterns. Our results suggest that these deeper internal signals carry most of the useful uncertainty information in clinical text.

5 Conclusion

We introduced Reverse Probing, the first 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 10 minutes, and making uncertainty estimation practical where evaluating full generated summaries is infeasible. Reverse Probing consistently outperforms all adapted baselines, achieving $4\times$ improvement over the strongest.

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, a signal more sensitive than token probability alone. 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. By tracing predictions back to specific internal signals, Reverse Probing turns uncertainty estimation into an interpretable process, taking a step toward explainable AI in clinical settings.

Limitations

Model coverage. The number of models evaluated is mainly based on Mistral and Llama, and we haven't validated Reverse Probing across a broader range of architectures and training recipes.

Large model adaptability. Our framework is better suited to 7-8B scale 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.

Low-resource Data. Expert-annotated clinical

datasets of this kind are inherently scarce, as labeling unsupported spans in medical text requires trained clinicians and is difficult to scale.

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 Hegselmann et al. (2025). Both datasets are derived from MIMIC-IV-Note (Johnson et al., 2023), a de-identified clinical dataset. Access to MIMIC-IV-Note requires credentialed registration, completion of CITI Data or Specimens Only Research training, and signing a data use agreement through PhysioNet. All authors fulfilled these requirements prior to all data access, in compliance with the PhysioNet data use policy. No new patient data was collected. Our framework is intended as a research tool for evaluating model uncertainty on existing labeled clinical summaries.

References

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

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, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA. Curran Associates Inc.

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.

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, Zelalem 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.

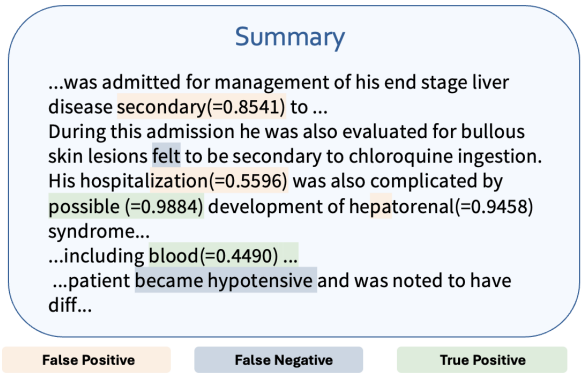

Figure 6: The ground truth labels and uncertainty quantification results for a sample clinical summary.

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. Reverse Probing requires only two forward passes per summary and processes 100 clinical summaries in approximately 10 minutes. Methods requiring multiple sampling passes are substantially 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 trail significantly on predictive performance.

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 Features Details

We extract token-level features from each model in four progressively richer configurations (93, 120, 204, and max: 454/886 features). Each version is built on the previous one. All 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, neighbourhood smoothness, medical entity detection, and lexical baseline statistics. Later versions expand these groups and introduce new ones: PMI-based grounding signals, entropy decomposition, hidden-state trajectory features across all transformer layers, and attention drift and rollout features. Tables 3–6 detail each configuration.

D Full Experimental Results

This appendix reports complete results for all experiments. Table 7 through Table 12 present Reverse Probing and all baseline results for all six base LLMs, covering every feature configuration and three classifiers (XGBoost, LightGBM, CatBoost) on both MIMIC-DI (Human-Written) and Generated-DI.

Table 1 in the main paper reports only the best classifier and best feature configuration per model; the full tables here allow direct comparison across configurations and classifiers. We also evaluate LightGBM (Ke et al., 2017) as an additional classifier, but its performance falls consistently below XGBoost and CatBoost across all settings, so we exclude it from the main results for brevity.

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	$-\big(\max_v \ell_v + \log \sum_v e^{\ell_v - \max \ell}\big)$	
Context Contrast (Δ BHC)	delta_prob	$p_i - p_i^-$	3
	delta_entropy	$H(\mathbf{p}) - H(\mathbf{p}^-)$	
	delta_energy	$\mathrm{energy}(\mathbf{p}) - \mathrm{energy}(\mathbf{p}^-)$	
Ranking & Semantic Similarity ($k \in \{5, 10, 20\}$)	rank_top$k$	Rank of token i among top-k predictions (0 if absent)	21
	in_top$k$	$\mathbf{1}[\mathrm{rank} > 0]$	
	max_sim_top$k$	$\max_j \cos(\mathbf{e}_i, \mathbf{e}_{(j)})$ over top-k tokens	
	avg_sim_top$k$	Mean cosine similarity to top-k tokens	
	top3_sim_top$k$	Mean cosine similarity to top-3 tokens	
	sim_std_top$k$	Std. dev. of cosine similarities to top-k	
	semantic_rank_top$k$	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	$p_i - \bar{p}_{\mathrm{nbr}}$	
	relative_isolation_ww	$(p_i - \bar{p}_{\mathrm{nbr}}) \,/\, (\sigma_{\mathrm{nbr}} + \varepsilon)$	
	medical_density_ww	Fraction of neighbours 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 training corpus	7
	freq_normalized	freq / total tokens	
	freq_log	$\log(\mathrm{freq} + 1)$	
	idf	$\log(N \,/\, (df + 1))$, N = number of documents	
	rarity	$1 \,/\, (\mathrm{freq} + 1)$	
	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_lℓ	$\lVert h_\ell \rVert_2$	12
	hidden_mean_lℓ	$\mathrm{mean}(h_\ell)$	
	hidden_std_lℓ	$\mathrm{std}(h_\ell)$	
	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_lℓ_1_to_lℓ_2	$\lVert h_{\ell_2} - h_{\ell_1} \rVert_2$ between consecutive sampled layers	6
	layer_cosine_lℓ_1_to_lℓ_2	$\cos(h_{\ell_1}, h_{\ell_2})$	
	3 consecutive pairs from the 4 sampled layers above.		
Attention Snapshots (6 heads)	attn_entropy_lℓ_hh	$H(\alpha_{\ell,h,i,\cdot})$, entropy of the attention row	18
	attn_to_bhc_lℓ_hh	$\sum_{j \le L_{\mathrm{BHC}}} \alpha_{\ell,h,i,j}$, total attention weight to BHC tokens	
	attn_max_lℓ_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 (6 pairs) is identical; layer indices are scaled proportionally to model depth.		
		Total	**93**

Table 3: Token-level features (93 total) extracted from both **Llama-3.1-8B-Instruct** (32 layers, 32 attention heads) and **Llama3-OpenBioLLM-70B** (80 layers, 64 attention heads). The feature set is identical across both models; 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 3.*		10 / 10
Context Contrast (Δ BHC)	*Same 3 features as Table 3.*		3 / 3
Ranking & Semantic Similarity ($k \in \{5, 10, 20\}$)	*Same 21 features as Table 3.*		21 / 21
Neighbourhood ($w \in \{2, 3, 5, 7\}$)	`neighbor_avg_w`w `neighbor_std_w`w `isolation_w`w `relative_isolation_w`w `medical_density_w`w	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 3.*		7 / 7
Hidden States (8 layers)	`hidden_norm_l`ℓ `hidden_mean_l`ℓ `hidden_std_l`ℓ	$\lVert h_\ell \rVert_2$ $\text{mean}(h_\ell)$ $\text{std}(h_\ell)$	24 / 24
	8 layers sampled at uniform intervals (≈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_l`ℓ_1`_to_l`ℓ_2 `layer_cosine_l`ℓ_1`_to_l`ℓ_2	$\lVert h_{\ell_2} - h_{\ell_1} \rVert_2$ 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_l`ℓ`_h`h `attn_to_bhc_l`ℓ`_h`h `attn_max_l`ℓ`_h`h	$H(\alpha_{\ell,h,i,\cdot})$, entropy of the attention row $\sum_{j \le L_{\text{BHC}}} \alpha_{\ell,h,i,j}$, 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:* $(7, 16), (17, 24), (57, 0), (57, 9), (57, 28), (77, 8)$. *Layer indices scaled proportionally; head count identical.*		
		Total	**120 / 120**

Table 4: 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 3 are abbreviated; only modified or new groups are shown in full. Compared to the 93-feature set: (1) neighbourhood windows reduced from $\{2, 3, 5, 7, 9\}$ to $\{2, 3, 5, 7\}$; (2) medical detection upgraded from keyword matching to scispaCy NER with 3 output features (dropping the one-hot entity indicators); (3) hidden-state sampling expanded to 8 layers with 7 inter-layer change pairs. Notation as in Table 3; layer/head indices that differ between models are listed in the notes rows.

Group	Feature	Description / Formula	#
Logit / Contrast / Ranking / Neighbourhood	*Same as Table 4: 10 logit + 3 context + 21 ranking & semantic + 20 neighbourhood.*		54 / 54
Medical NER (scispaCy + UMLS/MedDRA)	`is_medical`	Binary: token detected by NER or vocabulary	10 / 10
	`ner_entity_type`	Integer-encoded entity type (12 classes)	
	`medical_confidence`	Source confidence: NER+vocab = 1.0, NER = 0.85, vocab = 0.6	
	`ner_is_chemical`	One-hot entity-type flags	
	`ner_is_disease`		
	`ner_is_gene`		
	`ner_is_cancer`		
	`ner_is_anatomy`		
	`ner_is_pathology`		
	`ner_is_high_risk`	High-risk flag: chemical, drug, or cancer entity	
Lexical & Baseline	*Same 7 features as Table 4.*		7 / 7
PMI (*new*)	`pmi`	$\log p_i^{\text{with}} - \log p_i^{\text{no}}$; high PMI indicates BHC-grounded token	3 / 3
	`pmi_vs_prior`	$\log p_i^{\text{with}} - \log p_i^{\text{prior}}$; BHC+context vs. language prior	
	`npmi`	$\text{PMI} / (-\log p_i^{\text{with}} + \varepsilon)$, normalised PMI $\in [-1, 1]$	
Entropy Decomposition (*new*)	`bhc_info_gain`	$H(\mathbf{p}^-) - H(\mathbf{p})$; uncertainty reduced by BHC	5 / 5
	`ctx_info_gain`	$H(\mathbf{p}^{\text{prior}}) - H(\mathbf{p}^-)$; uncertainty reduced by context (non-BHC)	
	`total_info_gain`	$H(\mathbf{p}^{\text{prior}}) - H(\mathbf{p})$; total uncertainty reduction	
	`bhc_contribution_ratio`	$\Delta H_{\text{BHC}} / (\Delta H_{\text{total}} + \varepsilon)$; fraction of gain attributable to BHC	
	`bhc_info_gain_norm`	$\Delta H_{\text{BHC}} / (H_{\text{prior}} + \varepsilon)$; gain normalised by prior entropy	
Medical Prior Decomposition (*new*)	`halluc_risk_ratio`	$p_i^{\text{with}} / (p_i^{\text{prior}} + \varepsilon)$; ratio near 1 indicates reliance on prior	5 / 5
	`patient_specificity`	$p_i^{\text{with}} - p_i^{\text{prior}}$; absolute probability gain from BHC	
	`context_reliance_score`	$(p_i^{\text{with}} - p_i^-) / (p_i^- + \varepsilon)$; relative BHC contribution	
	`prob_dominance_order`	$\arg\max(p^{\text{prior}}, p^-, p^{\text{with}})$; ordinal encoding of probability ranking	
	`patient_specificity_norm`	$(p_i^{\text{with}} - p_i^{\text{prior}}) / (p_i^{\text{prior}} + \varepsilon)$	
Hidden States / Change / Attn. Snapshots	*Same as Table 4: 24 hidden states + 14 hidden state change + 18 attention snapshots.*		56 / 56 33 / 44
Attention Drift (4 heads, (*new*)	`attn_drift_l`ℓ_1`_l`ℓ_2`_h`h	$D_{\text{KL}}(\tilde{\alpha}_{\ell_1,h,i} \,\|\, \tilde{\alpha}_{\ell_2,h,i})$ between adjacent-layer attention rows; captures sudden distributional shifts	
	3 layer bands: shallow, mid, deep. 8B: layers $\{0\text{–}3\} \cup \{14\text{–}17\} \cup \{28\text{–}31\}$, *11 pairs* $\times$ *4 heads* $h \in \{0, 8, 16, 24\} = 33$*; 70B:* $\{0\text{–}7\} \cup \{35\text{–}42\} \cup \{70\text{–}79\}$, *11 pairs* $\times$ *4 heads* $h \in \{0, 16, 32, 48\} = 44$.		
Attention Rollout (3 checkpoints, (*new*)	`rollout_to_bhc_l`ℓ	$\sum_{j<L_{\text{BHC}}} [\mathbf{R}_\ell]_{i,j}$, cumulative attention to BHC	9 / 9
	`rollout_entropy_l`ℓ	$H(\mathbf{R}_\ell[i, \cdot])$	
	`rollout_max_weight_l`ℓ	$\max_j [\mathbf{R}_\ell]_{i,j}$	
	$\mathbf{R}_\ell = \prod_{k=0}^{\ell}(0.5\,\mathbf{A}_k + 0.5\,\mathbf{I})$*; checkpoints at end of each band:* $\ell \in \{3, 17, 31\}$ *(both models).*		
		Total	**204 / 204**

Table 5: Token-level features (204 total) extracted using **Llama-3.1-8B-Instruct** and **Llama3-OpenBioLLM-70B**. Groups unchanged from Table 4 are abbreviated. Compared to the 120-feature set, four changes are made: (1) Medical NER expanded from 3 to 10 features (one-hot entity-type indicators restored); (2) three PMI-based features added to quantify BHC grounding; (3) five entropy-decomposition features added to separate BHC vs. context contributions; (4) five medical-prior decomposition features added; (5) attention drift and rollout groups added, covering three depth bands. Notation as in Table 3; layer/head indices that differ between models are listed in the notes rows.

Group	Feature	Description / Formula	#
Logit / Contrast / Ranking / Neighbourhood / Medical NER / Lexical / PMI / Entropy Decomp. / Medical Prior Decomp.	*Same as Table 5: 54 + 10 + 7 + 3 + 5 + 5 = 84 features.*		84 / 84
Hidden States (all L layers, expanded)	`hidden_norm_l`ℓ `hidden_mean_l`ℓ `hidden_std_l`ℓ	$\lVert h_\ell \rVert_2,\ \ell = 0, \ldots, L{-}1$ $\mathrm{mean}(h_\ell)$ $\mathrm{std}(h_\ell)$	96 / 240
	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_l`ℓ`_to_l`$(\ell{+}1)$ `layer_cosine_l`ℓ`_to_l`$(\ell{+}1)$	$\lVert h_{\ell+1} - h_\ell \rVert_2$ $\cos(h_\ell,\ h_{\ell+1})$	62 / 158
	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 5.*		18 / 18 124 / 316
Attention Drift (4 heads, all $L{-}1$ pairs, expanded)	`attn_drift_l`ℓ`_l`$(\ell{+}1)$`_h`h	$D_{\mathrm{KL}}(\tilde{\alpha}_{\ell,h,i} \,\lVert\, \tilde{\alpha}_{\ell+1,h,i})$, 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_l`ℓ `rollout_entropy_l`ℓ `rollout_max_weight_l`ℓ	$\sum_{j<L_{\mathrm{BHC}}} [\mathbf{R}_\ell]_{i,j}$, cumulative attention to BHC $H(\mathbf{R}_\ell[i,\cdot])$ $\max_j [\mathbf{R}_\ell]_{i,j}$	24 / 24
	$\mathbf{R}_\ell = \prod_{k=0}^{\ell}(0.5\,\mathbf{A}_k + 0.5\,\mathbf{I})$*; expanded from 3 to 8 uniformly spaced checkpoints (8B: every 4 layers; 70B: every 10 layers).*		
		Total	**454 / 886**

Table 6: 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 5 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 3.

Classifier	F1	AUCROC	AUPRC
MIMIC-DI / 93			
XGBoost	0.3882	0.8293	0.4297
LightGBM	0.3823	0.8258	0.4188
CatBoost	0.3980	0.8216	0.4126
MIMIC-DI / 120			
XGBoost	0.3834	0.8482	0.4366
LightGBM	0.3730	0.8420	0.4292
CatBoost	0.3560	0.8328	0.4197
MIMIC-DI / 204			
XGBoost	0.4057	0.8615	0.4562
LightGBM	0.3967	0.8613	0.4512
CatBoost	0.4110	0.8512	0.4543
MIMIC-DI / MAX (454)			
XGBoost	0.4812	0.8911	0.5478
LightGBM	0.5016	0.8806	0.5310
CatBoost	0.4535	0.8691	0.5178
Generated-DI / 93			
XGBoost	0.2020	0.8192	0.1170
LightGBM	0.1724	0.7975	0.1277
CatBoost	0.1765	0.7856	0.1254
Generated-DI / 120			
XGBoost	0.2222	0.8364	0.1997
LightGBM	0.1905	0.7343	0.1574
CatBoost	0.2254	0.7979	0.1623
Generated-DI / 204			
XGBoost	0.2785	0.8318	0.1833
LightGBM	0.2588	0.7922	0.1798
CatBoost	0.2432	0.8062	0.1442
Generated-DI / MAX (454)			
XGBoost	0.3103	0.8819	0.2316
LightGBM	0.2540	0.8269	0.1975
CatBoost	0.2712	0.8435	0.2805

Table 7: Reverse Probing full results: BioMistral-7B. All feature configurations and classifiers.

Classifier	F1	AUCROC	AUPRC
MIMIC-DI / 93			
XGBoost	0.4459	0.8480	0.4654
LightGBM	0.3877	0.8373	0.4428
CatBoost	0.3944	0.8260	0.4328
MIMIC-DI / 120			
XGBoost	0.4514	0.8598	0.4818
LightGBM	0.4258	0.8351	0.4612
CatBoost	0.4378	0.8372	0.4545
MIMIC-DI / 204			
XGBoost	0.4742	0.8594	0.4894
LightGBM	0.4464	0.8545	0.4838
CatBoost	0.4259	0.8493	0.4706
MIMIC-DI / MAX (454)			
XGBoost	0.5519	0.9023	0.6022
LightGBM	0.5375	0.8902	0.5828
CatBoost	0.4928	0.8937	0.5737
Generated-DI / 93			
XGBoost	0.2500	0.8069	0.1938
LightGBM	0.3077	0.7836	0.2225
CatBoost	0.2740	0.8178	0.1799
Generated-DI / 120			
XGBoost	0.3000	0.8162	0.1878
LightGBM	0.2588	0.7851	0.1824
CatBoost	0.2963	0.8397	0.1960
Generated-DI / 204			
XGBoost	0.3125	0.8164	0.2005
LightGBM	0.2791	0.7960	0.2271
CatBoost	0.3291	0.8030	0.1932
Generated-DI / MAX (454)			
XGBoost	0.3611	0.8476	0.2373
LightGBM	0.2333	0.8429	0.2131
CatBoost	0.3934	0.8539	0.3070

Table 8: Reverse Probing full results: Mistral-7B. All feature configurations and classifiers.

Classifier	F1	AUCROC	AUPRC
MIMIC-DI / 93			
XGBoost	0.3953	0.8598	0.4418
LightGBM	0.4049	0.8832	0.4708
CatBoost	0.4015	0.8645	0.4485
MIMIC-DI / 120			
XGBoost	0.4466	0.8781	0.4615
LightGBM	0.4156	0.8640	0.4358
CatBoost	0.4259	0.8643	0.4445
MIMIC-DI / 204			
XGBoost	0.4454	0.8774	0.4600
LightGBM	0.4260	0.8841	0.4733
CatBoost	0.4247	0.8684	0.4525
MIMIC-DI / MAX (454)			
XGBoost	0.4651	0.8833	0.4976
LightGBM	0.4747	0.8761	0.4996
CatBoost	0.4534	0.8848	0.5029
Generated-DI / 93			
XGBoost	0.2162	0.8714	0.1719
LightGBM	0.2000	0.8137	0.1689
CatBoost	0.2273	0.8818	0.2194
Generated-DI / 120			
XGBoost	0.2857	0.9046	0.2421
LightGBM	0.2623	0.9110	0.2359
CatBoost	0.4082	0.8739	0.2588
Generated-DI / 204			
XGBoost	0.2449	0.8983	0.1897
LightGBM	0.2609	0.8634	0.1869
CatBoost	0.3396	0.8773	0.2277
Generated-DI / MAX (454)			
XGBoost	0.3038	0.9364	0.2892
LightGBM	0.3200	0.9080	0.2079
CatBoost	0.3146	0.9144	0.2895

Table 9: Reverse Probing full results: OpenBioLLM-8B. All feature configurations and classifiers.

Classifier	F1	AUCROC	AUPRC
MIMIC-DI / 93			
XGBoost	0.4608	0.8657	0.4742
LightGBM	0.4257	0.8549	0.4512
CatBoost	0.4435	0.8527	0.4842
MIMIC-DI / 120			
XGBoost	0.4272	0.8814	0.4732
LightGBM	0.4235	0.8572	0.4491
CatBoost	0.4504	0.8659	0.4808
MIMIC-DI / 204			
XGBoost	0.4488	0.8747	0.4763
LightGBM	0.4463	0.8787	0.4900
CatBoost	0.4331	0.8724	0.4858
MIMIC-DI / MAX (454)			
XGBoost	0.4712	0.8967	0.5110
LightGBM	0.4500	0.8870	0.5080
CatBoost	0.4766	0.8755	0.5140
Generated-DI / 93			
XGBoost	0.2276	0.9133	0.1496
LightGBM	0.1562	0.8581	0.1216
CatBoost	0.2143	0.8481	0.1956
Generated-DI / 120			
XGBoost	0.2000	0.9228	0.1854
LightGBM	0.1852	0.8346	0.1254
CatBoost	0.2609	0.9136	0.2033
Generated-DI / 204			
XGBoost	0.2222	0.9229	0.1479
LightGBM	0.1364	0.8490	0.1361
CatBoost	0.2400	0.9130	0.2409
Generated-DI / MAX (454)			
XGBoost	0.3117	0.9365	0.2121
LightGBM	0.2222	0.8890	0.2761
CatBoost	0.3036	0.9245	0.2253

Table 10: Reverse Probing full results: Llama3.1-8B. All feature configurations and classifiers.

Classifier	F1	AUCROC	AUPRC
MIMIC-DI / 93			
XGBoost	0.3665	0.8545	0.3023
LightGBM	0.3668	0.8562	0.2968
CatBoost	0.3900	0.8658	0.3260
MIMIC-DI / 120			
XGBoost	0.3672	0.8512	0.2991
LightGBM	0.3629	0.8361	0.2660
CatBoost	0.3737	0.8562	0.3023
MIMIC-DI / 204			
XGBoost	0.3890	0.8639	0.3125
LightGBM	0.3695	0.8525	0.3072
CatBoost	0.3870	0.8585	0.3281
MIMIC-DI / MAX (886)			
XGBoost	0.3876	0.8710	0.3500
LightGBM	0.3731	0.8549	0.3408
CatBoost	0.4009	0.8706	0.3597
Generated-DI / 93			
XGBoost	0.2222	0.8089	0.1249
LightGBM	0.1111	0.7940	0.0848
CatBoost	0.2540	0.8262	0.2008
Generated-DI / 120			
XGBoost	0.1636	0.8130	0.1208
LightGBM	0.1944	0.7625	0.0795
CatBoost	0.2286	0.8218	0.1725
Generated-DI / 204			
XGBoost	0.1837	0.8141	0.1277
LightGBM	0.1690	0.7320	0.1182
CatBoost	0.2162	0.7998	0.1548
Generated-DI / MAX (886)			
XGBoost	0.3030	0.7985	0.2162
LightGBM	0.1622	0.7791	0.2161
CatBoost	0.2459	0.8117	0.2073

Table 11: Reverse Probing full results: OpenBioLLM-70B. All feature configurations and classifiers.

Classifier	F1	AUCROC	AUPRC
MIMIC-DI / 93			
XGBoost	0.3657	0.8407	0.2965
LightGBM	0.3803	0.8341	0.3074
CatBoost	0.3839	0.8494	0.3528
MIMIC-DI / 120			
XGBoost	0.3382	0.8408	0.3057
LightGBM	0.3563	0.8235	0.2747
CatBoost	0.3502	0.8337	0.2964
MIMIC-DI / 204			
XGBoost	0.3643	0.8524	0.3233
LightGBM	0.3850	0.8435	0.3093
CatBoost	0.3636	0.8533	0.3268
MIMIC-DI / MAX (886)			
XGBoost	0.3932	0.8594	0.3934
LightGBM	0.3865	0.8408	0.3552
CatBoost	0.3968	0.8563	0.3835
Generated-DI / 93			
XGBoost	0.1667	0.8344	0.1252
LightGBM	0.2105	0.7888	0.1353
CatBoost	0.2121	0.8055	0.1679
Generated-DI / 120			
XGBoost	0.2609	0.8226	0.1486
LightGBM	0.1923	0.7736	0.1345
CatBoost	0.1875	0.7657	0.1322
Generated-DI / 204			
XGBoost	0.1633	0.7877	0.1247
LightGBM	0.1538	0.7730	0.1062
CatBoost	0.1905	0.8179	0.1378
Generated-DI / MAX (886)			
XGBoost	0.2963	0.8132	0.1697
LightGBM	0.1463	0.8097	0.1297
CatBoost	0.2326	0.7999	0.1939

Table 12: Reverse Probing full results: Llama3.1-70B. All feature configurations and classifiers.