From Compact to Diffuse: How LLM Uncertainty Degrades Linear Probes

Anonymous ACL submission

Abstract

Probing has been widely used to decode knowledge representations in LLMs. Yet, existing paradigms predominantly rely on dataset-level metrics, which obscure instance-level anomalies and leave the underlying mechanisms of specific probing failures largely unexplored. To address this limitation, we propose a fine-grained evaluation framework that links internal probing behavior with external generative behavior. Our analysis reveals a strong negative correlation between probe performance and the LLM response uncertainty. To mechanistically explain this phenomenon, we introduce the Feature Diffusion Hypothesis, which posits that probing failures should not be interpreted as a binary absence of knowledge, but rather as evidence of structural degradation in the latent representation space. Specifically, confident generation relies on compactly encoded features, whereas high uncertainty is driven by a diffuse distribution of concept features. This dispersion dilutes the conceptual signal, simultaneously degrading probe performance. Comprehensive experiments across six datasets and six LLMs consistently validate the correlation and hypothesis. Ultimately, these findings provide new insights into efficient hallucination detection from internal representations.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Yang et al., 2024; Team et al., 2023) have demonstrated remarkable performance across a wide range of tasks (Park et al., 2023; Wang et al., 2024a; Qiu et al., 2024; Thirunavukarasu et al., 2023). However, LLMs frequently generate non-factual or fabricated content, so called "hallucinations" (Maynez et al., 2020; Ji et al., 2023), which poses significant risks for deployment in high-stakes and specialized domains such as law and healthcare. Understanding the internal mechanisms behind such failures (Singh et al., 2024; Zhao et al., 2024; Wang et al., 2024b) has therefore become a central step toward building LLMs that are not only capable, but also reliable, interpretable, and trustworthy in real-world applications.

To explain LLMs' behaviors, probing-based explanations have become a widely used approach for examining what information is encoded in LLM representations. These methods (Alain and Bengio, 2017; Kim et al., 2018; Belinkov, 2022; Gurnee and Tegmark, 2024; Mikolov et al., 2013a) typically train a simple auxiliary model, known as a probe, to predict concept attributes, such as syntactic properties or toxicity, from the representations of a frozen model, thereby revealing how such concepts are organized in its latent space. For example, recent studies show that LLM representations exhibit substantial alignment with human perceptual and conceptual systems in domains such as color and space (Patel and Pavlick, 2022; Abdou et al., 2021). Similarly, Nanda et al. (2023) demonstrate that OthelloGPT (Li et al., 2023) learns linearly probeable representations of board states, including turn-taking information. Gurnee and Tegmark (2024) further reveal that LLMs encode temporal and spatial information across multiple scales. Dong et al. (2025) point out that LLMs encode global attributes of their entire future responses such as response length and reasoning steps, and answer confidence.

Despite the success of probing methods in uncovering knowledge representations, existing studies predominantly rely on dataset-level aggregate metrics, such as overall accuracy or R^2 scores, to assess representational alignment. For instance, a probe achieving an accuracy or R^2 score of 80% (Patel and Pavlick, 2022; Gurnee and Tegmark, 2024; Dong et al., 2025) is empirically interpreted as evidence that the target concept is well encoded in the model's latent space. However, this macro-centric paradigm collapses critical instance-level anomalies (the remaining 20% of failing in-

stances) into global averages, thereby obscuring the fine-grained failure modes that emerge when concept representations are weakly, inconsistently, or loosely encoded. This raises a key question: when a probe performs well on average but fails on specific instances, how do such failures manifest in its external generative behavior, and what happens inside the model's representation space?

To understand probe performance and response behavior, we shift our analytical focus from traditional global statistics to instance-level generative behaviors. We design a novel instance-level evaluation framework to systematically map external uncertainty to internal probe performance (shown in Figure 1). Rather than training a single global probe, we stratify query samples by their generative uncertainty (e.g., measured by response variance or entropy across multiple inferences) and train separate probes for different uncertainty segments. We finally report the correlation efficient between the uncertainty score and probe performance across these segments. Our results reveal a striking phenomenon: when an LLM exhibits lower epistemic uncertainty, its latent representations can be highly accurately captured by a linear probe. Conversely, an instance-level probing failure consistently coincides with severe external generative uncertainty.

To unpack the mechanisms behind this correlation, we propose the Feature Diffuse Hypothesis, positing that an LLM's external hallucination and internal probing failure stem from the exact same structural breakdown. To empirically verify this, we introduce an intervention view based on feature attribution (AttnLRP (Achtibat et al., 2024)) to extract the feature-importance profiles underlying the LLM's responses. By progressively zeroing out dimensions based on importance scores, we observe the resulting degradation in probe performance. Crucially, our intervention analysis reveals that LLM uncertainty operates along a representational degradation continuum. Low-uncertainty responses stem from a compactly encoded set of features, enabling the probe to fit effortlessly. In contrast, for high-uncertainty instances, concepts are not entirely absent, but rather diffusely encoded with high structural heterogeneity. This structurally diffuse distribution limits the signal-to-noise ratio, simultaneously causing divergent generative outputs and hampering the linear probe's ability to decode the concept.

Our contributions are summarized as follows:

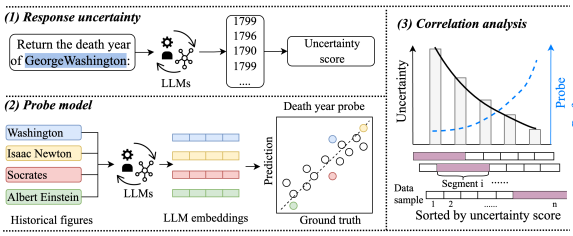


Figure 1: The proposed instance-level evaluation framework. By stratifying query samples based on their external generative uncertainty, we evaluate segment-specific linear probes trained on entity embeddings. The results demonstrate a strong negative correlation between LLM uncertainty and internal probe decodability.

- We shift the probing paradigm from global statistics to instance-level generative behaviors, uncovering a strong negative correlation between LLM response uncertainty and probe performance.
- Through feature intervention, we demonstrate that high response uncertainty stems from diffusely encoded features in the latent space, which inherently hampers probe decodability.
- We conduct extensive experiments on six datasets using six LLMs to validate our hypothesis and further reveal that LLM uncertainty operates along a representational degradation continuum.

2 Related Work

LLM uncertainty. LLMs often suffer from hallucination (Maynez et al., 2020; Ji et al., 2023), producing non-factual or fabricated responses that deviate from grounded knowledge. One source of hallucination arises when the answer is deterministic yet the model's predictions vary (Hüllermeier and Waegeman, 2021; Abbasi Yadkori et al., 2024; Shorinwa et al., 2025). Accordingly, response-level uncertainty has become a useful signal for hallucination detection where dispersion over multiple responses from the same query is measured.

Prior works map response uncertainty across multiple dimensions, spanning internal activations, token-level signals, and sentence representations. Representative approaches include internal activation covariance metrics (Chen et al.; Sriramanan et al., 2024), token log-likelihood (Kadavath et al., 2022), sentence semantic entropy (Farquhar et al., 2024; Kuhn et al., 2023; Kossen et al., 2025), and other uncertainty estimation techniques (Dwaracherla et al., 2023; Johnson et al., 2024; Wang et al., 2023). In our work, we forge a mechanistic link between internal representa-

tions—analyzed via probing—and external generative behaviors.

Probing (Alain and Bengio, 2017; Kim et al., 2018; Belinkov, 2022) is a simple yet effective interpretability technique for assessing whether a target concept is encoded in a model's representations. A lightweight probe, typically under the linear-representation hypothesis (Mikolov et al., 2013b), is trained on an annotated dataset to predict concept attributes (e.g., color, birth/death year) from the model's latent representations. Recent studies (Patel and Pavlick, 2022; Abdou et al., 2021) have shown that LLM representations exhibit strong alignment with human perceptual and conceptual systems, as evidenced through linear probing. Nanda et al. (2023) further demonstrate that OthelloGPT (Li et al., 2023) learns internal representations of board turn-taking states that can be linearly decoded. Similarly, Gurnee and Tegmark (2024) reveal that LLMs linearly encode continuous temporal and spatial information across multiple scales. While standard probing benchmarks successfully show that LLMs encode rich conceptual knowledge, their reliance on aggregate dataset-level metrics inherently overlooks fine-grained probing failures and how they influence the model's actual generative outputs

3 Methodology

3.1 Preliminary

Let an LLM f be decomposed into two modules $f = \varphi \circ \phi$, where ϕ maps a textual input x to a latent representation $\phi(x) \in \mathbb{R}^d$, and φ maps this latent representation $\phi(x)$ to the final generation $\varphi(\phi(x))$. A probing method (Alain and Bengio, 2017; Belinkov, 2022; Zhao et al., 2024) aims to train a simple probe g_W, parameterized by W, which predicts the presence or strength of a concept from the LLM latent representation $\phi(x)$. For example, consider a training dataset containing the names of historical figures and their death years, where the probe is trained to predict the death year from the embeddings of corresponding person names. Here, $\mathcal{D} \in \mathbb{R}^{n \times d}$ denotes the matrix of latent representations $\phi(x)$ for the n samples, and Y contains the corresponding concept values (presence or magnitude), taking either discrete (Kim et al., 2018; Alain and Bengio, 2017; Nanda et al., 2023) or continuous (Gurnee and Tegmark, 2024) values. The probe is trained by minimizing the loss with respect to its parameters W:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, g_W(D_i)\right) + \lambda ||W|_2^2, \quad (1)$$

where $\ell(\cdot)$ is a loss function (e.g., cross entropy or mean square error) measuring the difference between the prediction and its ground truth y_i. For classification tasks, we report test-set accuracy as the probe's performance metric; for regression tasks, we evaluate the test-set R^2 score and Spearman rank correlation. Higher probe performance indicates that the LLM more effectively encodes information relevant to the target concept.

3.2 Correlating Uncertainty with Probes

Previous works (Patel and Pavlick, 2022; Gurnee and Tegmark, 2024; Dong et al., 2025) often report the accuracy or R^2 score on the whole dataset. However, this macro-centric evaluation blinds us to minority failure cases, leaving the mechanistic link between their local representational anomalies and external generative behaviors unexplored.

To empirically validate correlation, we propose a novel framework, as illustrated in Figure 1. Given a probe dataset consisting of entity names paired with their target concept values, we query an instruction-tuned LLM multiple times (typically 20) using the same natural-language prompt to elicit the target concept, for example, "Return the death year of George Washington." Because LLMs adopt stochastic decoding in their recursive generation process, repeated sampling can yield different responses (e.g., "1799," "1796," and so on). We then apply uncertainty measurements to quantify the variability in these responses for each sample. After that, we rank all samples by descending response uncertainty, then partition this ordered list into overlapping segments using a sliding-window approach; each successive segment therefore contains progressively lower-uncertainty examples.

Instead of training a single probe for the entire dataset in previous approaches, we train a separate probe for each segment. For a segment with m samples, we initially obtain the LLM latent representation at a target layer l, following previous work (Gurnee and Tegmark, 2024). This yields a representation dataset $\{D_l, Y\}$, where D_l is the set of layer-l embeddings of m entities, Y contains the corresponding ground-truth labels (e.g., death year, birthplace). We split the segment into training and test sets with the K-fold cross validation (K=5

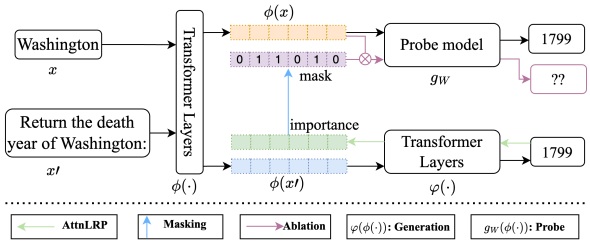


Figure 2: Pipeline for explaining the observed correlation. Step 1, obtaining feature-importance vectors w.r.t. LLM responses; Step 2, using these importance scores to mask entity representations; and Step 3, evaluating probe performance under progressive feature ablation.

in our experiments), and train a linear regression model $\hat{Y} = D_l W$ by optimizing the objective in Eqn (1). As the segments are ordered by decreasing uncertainty, we can readily observe the correlation probe performance and response uncertainty by calculating correlation coefficients such as Kendall's τ and Spearman's ρ across all segments.

3.3 Explaining the Correlation

To mechanistically unpack the observed correlation, we introduce an intervention framework based on feature attribution. We hypothesize that concept-related information is processed by shared internal circuits during the forward pass, whose representation jointly determines both external generative behavior and internal probe decodability. Under this hypothesis, perturbing features that are critical for text generation should simultaneously impair the linear decodability of the corresponding concept. Specifically, we test whether masks constructed from features that are important for LLM responses can causally modulate probe performance.

The detailed pipeline is illustrated in Figure 2. We first compute importance scores for the latent representation $\phi(x) \in \mathbb{R}^d$ with respect to the LLM's response using AttnLRP (Achtibat et al., 2024; Arras et al., 2025), an extension of the LRP framework (Bach et al., 2015) for transformer models. This method yields a feature-importance vector $\sigma(x) \in \mathbb{R}^d$, where each entry quantifies the contribution of the corresponding dimension of $\phi(x)$ to the generated output. A higher importance value indicates that ablating that feature in $\phi(x)$ would dramatically degrade response quality. After obtaining the feature importance vector with respect to the LLM's response, we use this vector as a binary mask to progressively zero out corresponding dimensions of the entity embeddings provided to the probe model, and then observe the resulting degradation in probe performance. We evaluate the degradation from two perspectives: (i) Freezing the probe and directly report the probe performance on masked test data (Wang et al., 2024c); (ii) Retraining the probe on masked training data, and reporting its performance on the masked test set, following the Remove and Retrain (RoAR) setting (Adebayo et al., 2018).

If the probe and the LLM's generation process rely on the same subset of representation, then restricting the probe to the high-importance dimensions identified from LLM generation should yield performance comparable to that achieved using the full dimension.

3.4 Feature Diffuse Hypothesis

Building on the above analysis, we introduce the *Feature Diffusion Hypothesis*. We argue that probing failures under high uncertainty do not necessarily imply the absence of knowledge. Instead, they suggest that concept-relevant features become geometrically dispersed in the latent space, reducing their linear decodability. In this view, confident responses are supported by compactly encoded concepts, whereas uncertain responses arise from diffuse feature distributions. We formalize this hypothesis by analyzing feature-importance profiles across uncertainty levels and showing how feature diffusion degrades probe performance.

Assume the set of important features for a given response r_i is denoted as $\hat{\phi}_i(x) \in \mathbb{R}^k \subseteq \phi(x)$, where $k \ll d$, following the sparse representation findings (Voita et al., 2019; Luo et al., 2025). When an LLM exhibits greater variability in its responses to the same query x, each response can be traced back to a different subset of important features in $\phi(x)$. In such cases, LLM will rely on a more diffuse set of feature dimensions in the representation space to answer user queries, i.e., the union of the activated feature subsets, $\bigcup_i \hat{\phi}_i(x)$.

According to the Lasso oracle inequality in compressed sensing (Bühlmann and Van De Geer, 2011), the gap between the empirical risk and the true risk of ridge regression (the probe) is upper-bounded by a sparsity term-the ratio of important features used. Higher uncertainty in LLM responses implies that probe performance is likely to deteriorate, since the relevant concept information resides in a more diffuse set of features, making it harder for the probe to learn a stable, low-dimensional mapping. The full analysis is demonstrated in Appendix section A.

Table 1: Correlation coefficients between probe performance and LLM uncertainty (measured by variance). τ and ρ denote the Kendall and Spearman rank correlation coefficients. '—' indicates that the LLM's uncertainty is uniformly low across all samples, making correlation measurement impossible; The underline symbol means many samples exhibit low uncertainty, so we compute correlations using only the 20,000 most uncertain samples.

Dataset	LLama 8B		LLama 70B		Qwen 14B		Qwen 72B		Mistral 24B		Gemma 27B	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
*Correlation coefficient: response uncertainty (**variance**) vs probe performance (R^2).*												
Figures	-0.89	-0.98	-0.76	-0.88	-0.71	-0.83	-0.79	-0.92	-0.86	-0.95	-0.96	-0.99
Artworks	-0.94	-0.99	-0.97	-0.93	-0.81	-0.93	-0.79	-0.94	-0.90	-0.97	-0.89	-0.97
News	-0.57	-0.71	-0.92	-0.98	-0.45	-0.67	-0.41	-0.68	-0.78	-0.90	-0.93	-0.99
NYC	-0.89	-0.96	-0.80	-0.90	-0.88	-0.96	-0.96	-0.99	-0.97	-0.99	-0.75	-0.88
USA	-0.91	-0.98	—	—	-0.86	-0.94	—	—	-0.72	-0.87	-0.91	-0.96
World	-0.80	-0.92	-0.89	-0.91	-0.69	-0.74	-0.89	-0.97	-0.92	-0.98	-0.96	-0.99
*Correlation coefficient: response uncertainty (**variance**) vs probe performance (**Spearman rank score**).*												
Figures	-0.67	-0.82	-0.77	-0.90	-0.82	-0.95	-0.55	-0.72	-0.86	-0.96	-0.89	-0.97
Artworks	-0.96	-0.99	-0.94	-0.87	-0.77	-0.91	-0.79	-0.93	-0.90	-0.97	-0.88	-0.97
News	-0.57	-0.72	-0.93	-0.99	-0.45	-0.66	-0.68	-0.85	-0.82	-0.93	-0.92	-0.98
NYC	-0.85	-0.91	-0.80	-0.90	-0.93	-0.97	-1.00	-1.00	-0.95	-0.99	-0.86	-0.94
USA	-0.87	-0.96	—	—	-0.88	-0.91	—	—	-0.71	-0.85	-0.74	-0.89
World	-0.84	-0.93	-0.89	-0.92	-0.77	-0.86	-0.83	-0.78	-0.89	-0.97	-0.86	-0.95

Table 2: Correlation coefficients between probe performance and LLM uncertainty (measured by entropy) on three temporal datasets. We exclude entropy for spatial datasets because even small numerical deviations in latitude or longitude can translate into substantial geographic distances, distorting the measurement.

Dataset	LLama 8B		LLama 70B		Qwen 14B		Qwen 72B		Mistral 24B		Gemma 27B	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
*Correlation coefficient: response uncertainty (**entropy**) vs probe performance (R^2).*												
Figures	-0.96	-0.87	-0.90	-0.77	-0.82	-0.76	-0.96	-0.85	-0.97	-0.89	-0.99	-0.95
Artworks	-0.99	-0.98	-0.99	-0.96	-0.97	-0.90	-0.87	-0.71	-0.98	-0.91	-0.88	-0.76
News	-0.99	-0.97	-0.99	-0.98	-0.94	-0.88	-0.91	-0.81	-0.96	-0.90	-0.98	-0.90
*Correlation coefficient: response uncertainty (**entropy**) vs probe performance (**Spearman rank score**).*												
Figures	-0.98	-0.99	-0.90	-0.97	-0.90	-0.83	-0.80	-0.93	-0.91	-0.97	-0.96	-0.99
Artworks	-0.97	-0.99	-0.98	-0.99	-0.88	-0.96	-0.80	-0.91	-0.91	-0.98	-0.90	-0.97
News	-0.95	-0.98	-0.98	-0.99	-0.89	-0.95	-0.81	-0.90	-0.96	-0.83	-0.93	-0.93

4 Experiments

4.1 Datasets

We evaluate our method on six datasets introduced in (Gurnee and Tegmark, 2024), with dataset sizes and representative examples introduced in Table 3. The first three datasets consist of historical figure names, artwork titles, and news event names, paired with corresponding temporal target concepts (e.g., year of death, release year). The last three datasets consist of location names in New York State, the United States, and worldwide, paired with their corresponding spatial concepts, namely latitude and longitude. We prompt the LLMs using a prompt template to extract target values, e.g., "Return the death year (or latitude and longitude) of {ENTITY}", and collect 20 responses per query. Prompts for each dataset are listed in Table 4, Appendix. For probe training, we fit a linear probe using only the "{ENTITY}"representations to predict the corresponding concept values.

4.2 Implementation Details

Our experiments are conducted on a series of instruction-tuned LLMs, ranging from 8 billion to 72 billion parameters, spanning multiple LLM families (including Llama-3.1 (8B, 70B), Qwen 2.5 (14B, 72B) (Yang et al., 2024), Mistral-Small

Table 3: Descriptions of the datasets.

Dataset	Count	Example
Figures	37,539	"Cleopatra", "Newton"
Artworks	31,321	"Stephen King's It"
News	28,389	"Dozens Rescued From Iraq ... "
NYC	19,838	"Borden Avenue Bridge"
USA	29,997	"Columbia University"
World	39,585	"Los Angeles", "Bei Jing"

(24B), and Gemma-2 (27B) (Team, 2025) series).

As shown in Gurnee and Tegmark (2024), probe performance plateaus since the middle layers of the LLM, a trend we also confirm in our own replications. Accordingly, we extract entity embeddings from the middle layers for remaining analysis. Unless otherwise noted, each sample is represented by the last-token activation from that layer (captured after the residual connection), following (Gurnee and Tegmark, 2024; Li et al., 2020). For larger LLMs with higher-dimensional embeddings, we enlarge the sliding-window segment size to include more samples and reduce overfitting in probing. We report the test-set R^2 *score and Spearman rank correlation* as the probe performance.

In LLM generation process, as the target concept are continuous, we adopt variance as our default measurement of LLM uncertainty (Bülte et al., 2025). For temporal concepts, we also experiment with entropy as an alternative uncertainty estimator. Note that entropy is not suited for measuring uncertainty in latitude and longitude predictions, as even small numerical differences can correspond to substantial geographic discrepancies.

Additional details are provided in supplementary section B, including the window size and stride used for segment division, probe training configurations, and the target probe layer for each LLM.

4.3 Main Results

We report the relationship between LLM uncertainty and probe performance across six datasets and six publicly LLMs in Tables 1 and 2. Overall, probe accuracy consistently increases as response uncertainty decreases, as illustrated in Figures 6 and 7 in the Appendix. Notable exceptions arise for stronger models, such as Llama 3.1 70B and Qwen 2.5 72B, which produce consistently correct responses on relatively simple tasks and exhibit negligible uncertainty. In such cases, the lack of response variability leaves insufficient signal to meaningfully assess the correlation between uncertainty and probe performance. Therefore, we exclude these cases from our evaluation. Accordingly, when uncertainty scores sharply collapse to zero after sorting, we restrict our evaluation to the top $20,000$ most uncertain samples.

The quantitative results in Table 1 corroborate a pronounced negative correlation: in most settings, Kendall's τ and Spearman's ρ between probe performance (measured by R^2 and Spearman's rank correlation coefficient) and LLM uncertainty fall below -0.5, often approaching -1. This effect is most evident when model uncertainty exhibits substantial variation across the test set. In addition, when response uncertainty is measured using entropy on the first three temporal datasets, the same trend persists, as shown in Table 2.

To assess whether this correlation is linear, we report Pearson correlation coefficients in Table 6 of Appendix. This results indicate that the linearity depends on the choice of uncertainty estimator. As entropy and variance are defined on different scales and exhibit different sensitivities to the numerical ranges of the ground-truth values, the strength of the observed correlation varies accordingly. Specifically, Pearson correlations are strongly negative when entropy is used as the uncertainty estimator, but they disappear when variance is employed. A thorough study of the conditions under which linearity holds is deferred to future work.

4.4 Hyperparameter Experiments

To assess the robustness of this correlation, we conduct sensitivity analyses with respect to generation temperature, sliding-window strategies, and prompt templates, as reported below.

Generation temperature. Temperature controls the creativity and randomness of LLM responses: lower values typically result in more deterministic outputs, while higher values lead to more stochastic responses. In this study, we generate responses using different temperatures, ranging from 0.1 to 1.5. We then report the correlation coefficients under different temperature settings on three temporal datasets for Llama 3.1 8B, as shown in Figure 3(a–c). The results show that the correlation remains stable across the temperature range of $[0.1, 1]$, which covers the default temperature settings for most LLMs. Similar results are observed using the Mistral-Small model in Figure 8.

Sliding window strategy. Window size and stride jointly control the sample number per segment and the total number of segments. Too few sam-

ples per segment can destabilize probe training, whereas too few segments can make the correlation estimates unreliable. In this study, we plot the Spearman correlation scores between LLM uncertainty (variance) and probe performance (Spearman rank coefficients) under different window and stride settings in Figure 3(d) on the *Historical Figures* dataset. Additional experiments using alternative uncertainty and probe measurements are reported in Figure 9 in the Appendix. Despite slight variations in the coefficient values, this correlation remains robust to the slide window parameters.

Prompt template. Different prompt template can affect the quality of generated responses, particularly when advanced prompting strategies such as chain-of-thought reasoning or in-context examples are employed. Here, we test several prompt templates that all have the same purpose: eliciting the LLM's answer to a specific target concept. Variants include role-playing instructions (e.g., "Act as a historian") and output-format constraints that guide the model's response. The prompt templates and the correlation scores, shown in Table 7 of Appendix, we can observe that the correlation persists regardless of the differences in prompt writing.

4.5 Correlation Analysis via Feature Ablation

We examine three temporal datasets to test our hypothesis that a shared circuit extracts concept information into subsets of embedding dimensions. We first compute feature-importance scores for a latent representation with respect to the LLM's outputs. Then, we progressively preserve features of entity embedding in descending order of importance.

Figure 4 (a–c) presents the results. The x-axis shows the fraction of top-important features retained (all others are zeroed), while the y-axis reports probe performance. The upper panels evaluate the probe model directly on masked representations; the lower panels follow the Remove-and-Retrain (RoAR) protocol, retraining the probe on masked training data and testing it on masked test data. We observe that keeping a small subset of highly important features achieves comparable probe performance as full features. Across all three datasets, retaining only the top 30% of features yields probe performance virtually identical to that obtained with the full embedding. Under RoAR, keeping the top 20% of features is enough to achieve a good fit on the Historical Figures and Artworks datasets. These results provide strong evidence for a shared, concept-related feature dimension that supports both the LLM's generation and the probe's predictions.

4.6 Verifying the hypothesis

In this study, we analyze how data subsets stratified by different uncertainty levels influence probe performance from an explainability perspective. In particular, we sample three equally sized subsets—5000 instances each—from the Historical Figures dataset, corresponding to low, medium, and high uncertainty regimes (hereafter Low, Mid, and High). For each subset, we compute feature-importance scores and progressively retain probe input features in descending order of importance. At each step, we record the performance increase of a probe trained on the full dataset.

The experimental results on Llama 3.1 (8B) are shown in Figure 4(d). A clear pattern emerges: the high-uncertainty subset must retain substantially more features to regain baseline probe performance, whereas the low-uncertainty subset achieves comparable accuracy with far fewer features. For instance, preserving the top 40% of important features approximately restores the probe performance for the high-uncertainty subset, while retaining only the top 20% is sufficient for the low-uncertainty subset, as shown in the upper figure. In the lower panel of Figure 4(d), the low-uncertainty probe reaches the similar original performance with just the top 5% of features, whereas the high-uncertainty probe still needs about 10% of features. These results provide evidence of a negative correlation, indicating that higher uncertainty is associated with lower probe performance.

This structural diffuse pattern is consistently observed on the Artworks dataset and when stratifying instances by response entropy (detailed in Appendix Figures 11 and 12).

4.7 Extension with Non-linear Representation

Here, we examine whether LLM embeddings are closely aligned with human prior knowledge in nonlinear representational settings when the models consistently produce correct answers. We consider two questions: (i) "What is the category of BRAND?" and (ii) "What hour comes after CLOCK?" For both tasks, LLMs produce highly consistent responses with uniformly low uncertainty, with entropy values of 0.12 and 0, respectively. We then visualize the layer-25 representations of BRAND and CLOCK HOUR using t-SNE.

As shown in Figure 5, the left panel visualizes

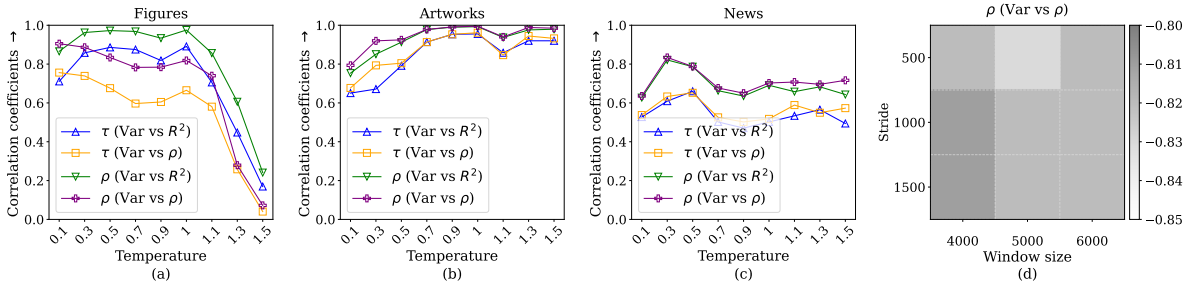


Figure 3: 'Var', τ, and ρ denote variance (uncertainty estimator), Kendall and Spearman rank correlation coefficients. Figure (a-c): The correlation under different temperatures for Llama 3.1 (8B). Here, we report absolute values for better visualization; Figure (d): the correlation under different settings in sliding window strategy.

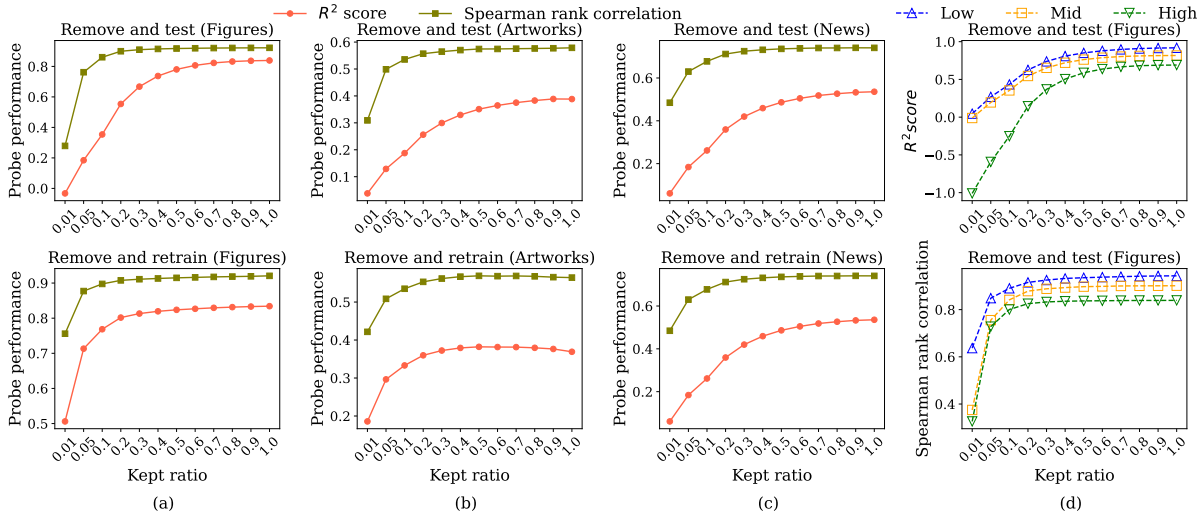
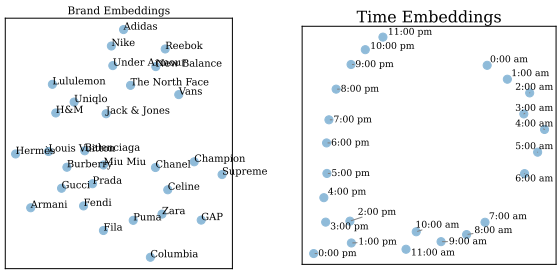


Figure 4: Correlation verification from lens of feature importance on Llama 3.1 (8B). x-axis represents the ratio of important features preserved, from 0.01 to 1 and y-axis denotes the resulting probe performance with *remove and test* and *remove and retrain* strategies. Figures (a-c), on three different temporal datasets; Figure (d), on three subsets with varied uncertainty level of Figures dataset.

embeddings of famous brands such as Nike, Adidas, and Armani. We observe that luxury brands (e.g., LV, Burberry) cluster together, while sports brands (e.g., Nike, Adidas, Puma) form a separate cluster. The right panel illustrates temporal representations spanning from 0 a.m. to 11 p.m., which are arranged in a circular pattern, mirroring the inherent periodic structure of clock time. Together, these examples suggest that when LLMs exhibit low response uncertainty, their latent representations are well-structured and closely aligned with human semantic priors.



(a) Brands (b) Time

Figure 5: T-SNE visualization of two examples that LLama 3.1 (8B) embeddings align with human knowledge, when LLMs exhibit lower response uncertainty.

5 Conclusion

In this paper, we propose a fine-grained evaluation framework to uncover the strong negative correlation between LLM generative uncertainty and linear probe decodability. To explain this link, we introduce the Feature Diffuse Hypothesis. Through feature ablation, we demonstrate that probing failures should not be merely interpreted as a binary absence of knowledge. Instead, high response uncertainty stems from a structurally diffuse distribution of concept features within the latent space. This representational degradation continuum fundamentally limits linear decodability, opening new avenues for efficient hallucination detection from internal representations.

Limitations

While this study establishes a strong mechanistic link between linear decodability and generative uncertainty, we acknowledge certain boundaries of our experimental scope, which outline promising directions for future work.

- The Ceiling Effect in Discrete Probing: Our study primarily focuses on continuous attributes (e.g., temporal and spatial regression) rather than discrete classification tasks. For simple discrete probing, modern LLMs often exhibit a ceiling effect, yielding highly accurate and consistently confident responses. This lack of variability collapses the uncertainty distribution, preventing the meaningful instance-level stratification required by our framework.
- Atomic Concepts vs. Open-ended QA: To cleanly observe representational mechanisms, our intervention framework isolates atomic concepts (e.g., extracting an entity's distinct attribute). However, real-world Question Answering (QA) scenarios often necessitate complex reasoning, multi-entity interactions, and compositional text generation. Investigating how the feature diffusion mechanism compounds under the complex uncertainty of open-ended QA remains an important avenue for future research.
- Entanglement of Confounding Factors: While the Feature Diffuse Hypothesis mechanically links structural dispersion with epistemic uncertainty, this relationship may be entangled with inherent confounders. Specifically, the observed diffusion could be jointly influenced by the intrinsic difficulty of the extraction task and the idiosyncratic encoding dynamics established during the LLM's pre-training phase. Disentangling dynamic generative uncertainty from static encoding artifacts remains a critical challenge for future analyses.

References

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty. *Advances in Neural Information Processing Systems*, 37:58077–58117.

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. Attnlrp: attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *The Fifth International Conference on Learning Representations*.

Leila Arras, Bruno Puri, Patrick Kahardipraja, Sebastian Lapuschkin, and Wojciech Samek. 2025. A close look at decomposition-based xai-methods for transformer language models. *arXiv preprint arXiv:2502.15886*.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

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

Peter Bühlmann and Sara Van De Geer. 2011. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Christopher Bülte, Yusuf Sale, Timo Löhr, Paul Hofman, Gitta Kutyniok, and Eyke Hüllermeier. 2025. An axiomatic assessment of entropy- and variance-based uncertainty quantification in regression. *Preprint*, arXiv:2504.18433.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Zhichen Dong, Zhanhui Zhou, Zhixuan Liu, Chao Yang, and Chaochao Lu. 2025. Emergent response planning in LLMs. In *Forty-second International Conference on Machine Learning*.

Vikranth Dwaracherla, Zheng Wen, Ian Osband, Xiuyuan Lu, Seyed Mohammad Asghari, and Benjamin Van Roy. 2023. Ensembles for uncertainty estimation: Benefits of prior functions and bootstrapping. *Transactions on Machine Learning Research*.

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

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Daniel D. Johnson, Daniel Tarlow, David Duvenaud, and Chris J. Maddison. 2024. Experts don't cheat: Learning what you don't know by predicting pairs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22406–22464. PMLR.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and 1 others. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. 2025. Semantic entropy probes: Robust and cheap hallucination detection in LLMs.

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*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Yuqi Luo, Chenyang Song, Xu Han, Yingfa Chen, Chaojun Xiao, Xiaojun Meng, Liqun Deng, Jiansheng Wei, Zhiyuan Liu, and Maosong Sun. 2025. Sparsing law: Towards large language models with greater activation sparsity. In *Forty-second International Conference on Machine Learning*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013a. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. 2024. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3).

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team. 2025. Gemma 3.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024a. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024b. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, Miami, Florida, USA. Association for Computational Linguistics.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. 2024c. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*.

Zi Wang, Alexander Ku, Jason Baldridge, Tom Griffiths, and Been Kim. 2023. Gaussian process probes (gpp) for uncertainty-aware probing. *Advances in neural information processing systems*, 36:63573–63594.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

A Formal Analysis of Section 3.4

Let us revisit the objective of explaining why response uncertainty is related to probe performance, thereby reinforcing the observed correlation between LLM uncertainty and probe accuracy. Given a textual sample x, we obtain the embedding of the target layer l, denoted as $\phi(x)$ where the layer index is omitted for simplicity. The final response is generated by forwarding $\phi(x)$ through the remaining layers, represented by $\varphi(\phi(x))$. Due to the stochastic nature of sampling during generation, the LLM may produce different responses for the same question x, denoted by $\{r_i, ..., r_m\}$. With the help of AttnLRP (Achtibat et al., 2024), we compute feature importance scores $\sigma_i(x)$ with respect to each response r_i, where a higher value indicates that the absence of the corresponding feature would significantly affect the final output.

Recent studies (Luo et al., 2025; Voita et al., 2019) have shown that only a small subset of features is highly important for model predictions. Let $\hat{\phi}_i(x) \in \mathbb{R}^k \subseteq \phi(x)$ denote the set of important features in $\phi(x)$ with respect to response r_i, where $k \ll d$. If the model produces varying responses, the feature set $\hat{\phi}_i(x)$ of each response is likely to differ, suggesting that diffuse features in $\phi(x)$ contribute to the final prediction. For instance, if the LLM gives two different death years for George Washington—1799 and 1796—the divergent outputs draw on a broader set of features in $\phi(x)$ than would be required for a single, consistent prediction like 1799.

Next we dive into the lower probe performance caused by high uncertainty response from Lasso oracle inequality in compressed sensing domain (Bühlmann and Van De Geer, 2011). For a linear regression model with an s-sparse true parameters, we have

$$\Pr\Big[L(\widehat{w}) - \widehat{L}(\widehat{w}) \;\leq\; C\,\sigma^2\sqrt{\frac{s\log d}{n}}\Big] \;\geq\; 1-\delta. \tag{2}$$

where,

$L(\widehat{w})$: Population (true) risk.

$\widehat{L}(\widehat{w})$: Empirical risk on n samples.

d : Total number of available features.

$s \ll d$: Number of *truly useful* (non-zero) features (sparsity).

σ^2 : Noise variance / loss Lipschitz constant.

C : Universal numerical constant.

δ : Confidence level of the high-probability bound.

With probability at least $1 - \delta$, the gap between the empirical risk and the population (true) risk is upper-bounded by a sparsity-controlled term (the ratio of important features used). For an LLM with higher response uncertainty, a diffuse number of features contribute to the diverse responses compared to a model with lower uncertainty, making it more difficult to achieve accurate predictions with a linear regression probe.

B Experiment Settings

B.1 Detailed Dataset Description

Our experiments are conducted on three spatial and three temporal datasets collected by Gurnee and Tegmark (2024). Each dataset contains the names of entities—such as celebrities, news headlines, artworks, or places—along with their associated spatial and temporal attributes. In (Gurnee and Tegmark, 2024), the authors propose learning a ridge regression probe to extract spatial or temporal information from entity embeddings. Table 3 summarizes the number of samples in each dataset and provides a representative example. Since we prompt LLMs to generate responses for the target concepts, we also list the corresponding prompt templates. Here we use simple prompts only, without employing advanced prompting strategies such as chain-of-thought or in-context examples. By default, the temperature for response generation is set to 1, which is the default setting in the HuggingFace pipeline generator.

B.2 Implementation Details

Here, we provide additional details about our experimental setup. Table 5 summarizes the settings used for training the probe model. Due to the higher latent dimensionality in large LLMs such as LLama 3.1 (70B), a larger number of training samples is required for the probe. To manage this, we set different window sizes and strides. For each sample, we extract the representation after the middle layer, using the hidden state of the last token.

In probe training, we randomly split the data whether a segment, a subset in Section 4.6, or the entire dataset in Section 4.5 into training and test sets at a $4:1$ ratio. We report both the R^2 score and the Spearman rank correlation coefficient on the test set to gauge probe performance. When experimenting on the full datasets, we further divided the training dataset into two parts: one for training the probe and the other for tuning the trade-off parameter λ. For all remaining experiments, we fix λ to the selected value without further tuning.

By default, we use the prompt template shown in Table 3 to generate multiple responses for each sample. We also experiment with alternative prompt formats on Historical Figures dataset with Llama 3.1 8B model, as listed in Table 7.

C Additional Experiment Results

C.1 Segment-wise Report of Uncertainty and Probe Performance

Figures 6 and 7 display segment-wise response uncertainty alongside probe accuracy for six datasets on Llama 3.1 8B. We can see that segments with high uncertainty consistently yield weaker probes, whereas low-uncertainty segments permit a much better fit. Segments are ordered from highest to lowest uncertainty from left to right, and probe accuracy increases monotonically as uncertainty decreases, indicating a strong negative correlation. As shown in Figure 7, we observe this negative correlation consistently holds regardless of whether uncertainty is quantified by variance or by entropy.

C.2 The linearity of correlation

Is the relationship between response uncertainty and probe performance linear? To verify this, we compute Pearson correlations on three temporal datasets (Table 6) using two uncertainty metrics—variance and entropy. As entropy and variance are defined on different scales and exhibit different sensitivities to the numerical ranges of the ground-truth values, the strength of the observed correlation varies accordingly. Entropy, in particular, yields the highest and most consistent Pearson correlations, which may help to explain why a recent work (Kossen et al., 2025) successfully trained a linear model to predict semantic entropy directly from hidden representations.

Table 4: Detailed descriptions of the datasets and prompts used to generate LLM responses.

Dataset	Count	Example	Prompt Template
Figures	37,539	"Cleopatra"	Return the death year of { }
Artworks	31,321	"Stephen King's It"	Return the release year of { }
News	28,389	"Dozens Rescued From ... "	Return the publish year of { }
NYC	19,838	"Borden Avenue Bridge"	Return the latitude and longitude of {} in NYC
USA	29,997	"Columbia University"	Return the latitude and longitude of {} in USA
World	39,585	"Los Angeles"	Return the latitude and longitude of {}

C.3 Self-entropy as uncertainty estimator

Table 8: Correlation between probe performance (measured by R^2 and Spearman rank score) and token-level uncertainty (self-entropy) on the Figures dataset using Llama 3.1 8B.

Spearman	Kendall	Pearson
Self-entropy vs R^2		
-0.85	-0.68	-0.84
Self-entropy vs Spearman score		
-0.96	-0.88	-0.93

The uncertainty scores discussed above, such as variance and entropy, are measured at the semantic (concept) level. In this study, we further evaluate whether the observed correlation remains robust at the token level by using self-entropy as an alternative uncertainty measure. Specifically, for an answer such as "1799", which is tokenized into two tokens (e.g., "179" and "9"), we define the self-entropy as the average logit entropy across all constituent tokens.

The correlation between self-entropy and probe performance is reported in Table 8. Consistent with our previous findings, we still observe a clear correlation on the Historical Figures dataset when using LLaMA 3.1 8B, demonstrating that the relationship is robust to the choice of uncertainty granularity.

C.4 Additional sensitivity experiments

In this study, we present additional sensitivity experiments in term of temperatures (Figure 8), the sliding window strategy (Figure 9) for training the probe, and prompt templates (Table 7) used for generating responses with more datasets and models. Our experiments show that:

- **Temperature**. The main findings remain stable for temperatures between 0.1 and 1.0; only extreme values (very low or very high) noticeably distort the correlation.
- **Sliding-window parameters**. Windows that are too narrow, or strides that are too wide, alter both the number of segments and the training samples per segment, which can in turn affect the observed trend. Within typical parameter ranges, however, the conclusions persist.
- **Prompt templates**. Across a variety of commonly used templates, our results are consistently robust.

C.5 Correlation Analysis via Feature Ablation

To complement the analyses in Section 4.5, we run an additional set of experiments reported in Figure 10. Because AttnLRP is tailored to smaller Llama variants and does not yet support Llama-3.1 (70 B), we instead employ Llama-2 (14 B) to test whether both the probe and the generation process rely on a shared subset of concept representation of $\phi(x)$. Generally, retaining the top 40% importance features to LLM response is sufficient to achieve a good fit for the probe models. In particular, in the remove-and-retrain experiments, keeping only the top 20% of the most informative features already yields a near-optimal probe fit, confirming that the salient information is shared by both the probe and LLM generation process.

C.6 Verifying the hypothesis

Figure 11 reports the remove and test results on different uncertainty subsets using Artworks dataset. To test the generalization of this conclusion, we use variance to measure the uncertainty here. We gradually preserve the top important features to LLM response and observe the resulting performance of the probe. From the bottom figure, we can see that preserving the top 5% of important features approximately achieves the similar Spearman rank correlation for low-uncertainty subset while 30% is needed for high-certainty subset. This experiment provides additional evidence to support our hypothesis. Note that the three subsets with varying levels

Table 5: Target layer, representation dimension and sliding window strategy (window size and stride) for different LLMs in probe model training.

Model	Target layer	#Layers	Dimension	Window size	Stride
Llama 3.1 (8B)	25	32	4096	6000	1000
Llama 3.1 (70B)	60	80	8192	12000	1000
Qwen 2.5 (14B)	36	48	5120	8000	1000
Qwen 2.5 (72B)	60	80	8192	12000	1000
Mistral-Small (24B)	30	40	5120	8000	1000
Gemma 2 (27B)	35	48	4608	8000	1000

Table 6: Pearson coefficient between probe performance and LLM response uncertainty on Llama 3.1 (8B, 70B), Qwen 2.5 (14B, 72B), Mistral-Small (24B), and Gemma-2 (27B) across three temporal datasets.

Dataset	LLama 8B	LLama 70B	Qwen 14B	Qwen 72B	Mistral 24B	Gemma 27B
Pearson coefficient: ***variance*** *vs* ***probe performance*** *(R^2).*						
Figures	-0.46	-0.53	-0.20	-0.37	-0.48	-0.73
Artworks	-0.36	-0.54	-0.20	-0.52	-0.50	-0.30
News	-0.07	-0.24	-0.3	-0.15	-0.54	-0.40
Pearson coefficient: ***variance*** *vs* ***probe performance (Spearman rank score)***						
Figures	-0.40	-0.56	-0.36	-0.26	-0.61	-0.42
Artworks	-0.40	-0.57	-0.15	-0.50	-0.50	-0.25
News	-0.06	-0.32	-0.30	-0.26	-0.61	-0.42
Pearson coefficient: ***entropy*** *vs* ***probe performance*** *(R^2).*						
Figures	-0.93	-0.96	-0.97	-0.96	-0.95	-0.96
Artworks	-0.98	-0.98	-0.98	-0.89	-0.94	-0.91
News	-0.99	-0.99	-0.99	-0.93	-0.97	-0.94
Pearson coefficient: ***entropy*** *vs* ***probe performance (Spearman rank score)***						
Figures	-0.96	-0.97	-0.81	-0.93	-0.97	-0.97
Artworks	-0.98	-0.99	-0.97	-0.89	-0.93	-0.96
News	-0.99	-0.99	-0.99	-0.95	-0.97	-0.92

Table 7: The correlation coefficients of different prompt templates on Llama 3.1 (8B) using the Figures dataset. τ, and ρ denote Kendall and Spearman rank correlation coefficients, respectively. The response uncertainty is measured by variance.

Prompt	Variance vs R^2		Variance vs ρ	
	τ	ρ	τ	ρ
System prompt: Play as a historian and return the death year of given person. User prompt: when did {} die?	-0.85	-0.96	-0.49	-0.66
System prompt: Return the death year of given person. User prompt: when did {} die?	-0.75	-0.85	-0.43	-0.66
Return the death year of {}:	-0.89	-0.98	-0.67	-0.82
What is the death year of {}? only return the year.	-0.61	-0.78	-0.62	-0.76

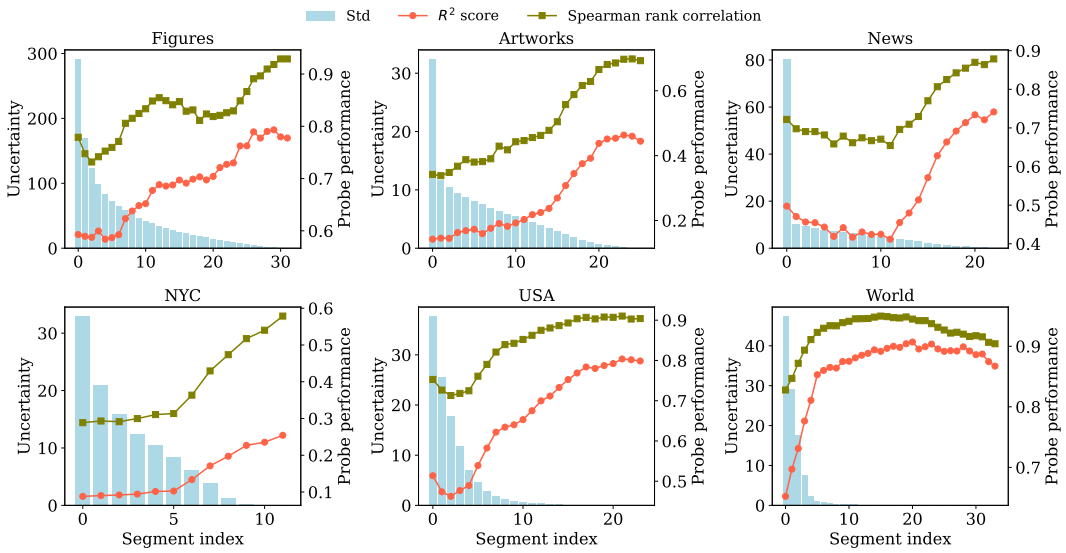


Figure 6: Segment-wise report of uncertainty score (variance) and probe performance on the Llama 3.1 (8B).

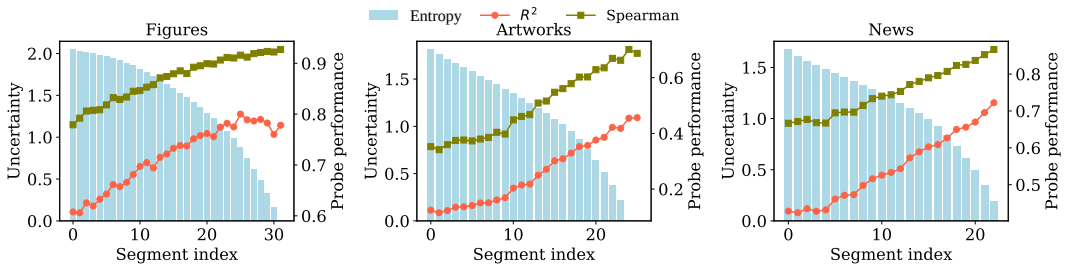


Figure 7: Segment-wise report of uncertainty score (entropy) and probe performance on the Llama 3.1 (8B).

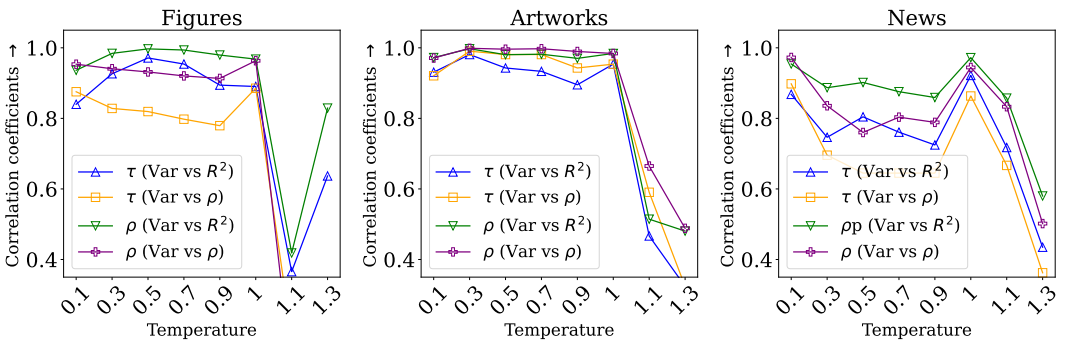


Figure 8: The correlation analysis with different temperatures during response generation for Mistral-Small (24B). 'Var', τ, and ρ denote variance (uncertainty estimator), Kendall and Spearman rank correlation coefficients. Here we report absolute values for better visualization performance.

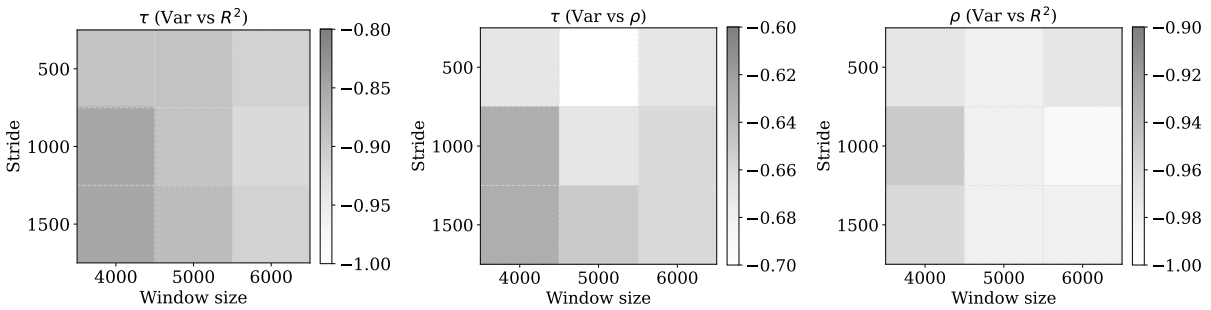


Figure 9: Correlation coefficient under different settings in the sliding window strategy, conducted on Historical Figures dataset using Llama 3.1 (8B). 'Var', τ, and ρ denote variance (uncertainty estimator), Kendall and Spearman rank correlation coefficients.

of uncertainty are selected based on response variance in above experiments. The same conclusion holds when using entropy as the uncertainty estimator to select three subsets with varying levels of uncertainty, as shown in Figure 12.

C.7 Impact of Entity Frequency

To address concerns that the strong negative correlation between LLM response uncertainty and probing performance is merely a byproduct of static data artifacts—specifically, entity frequency within the training corpus—we conducted a controlled analysis on Llama 3.1 8B using OpenWebText (Gokaslan et al., 2019), a publicly released 30GB corpus that is used to reproduce the training of GPT-2, as a proxy for entity popularity. Headlines and NYC datasets are excluded because only a small number of entities can be matched with OpenWebText.

As shown in Table 9, we observe **no clear or consistent correlation between entity frequency and LLM response uncertainty**. The Pearson and Spearman coefficients are generally small and exhibit inconsistent directions across domains, indicating that entity frequency alone is insufficient to account for variations in model uncertainty.

Table 9: Correlation (ρ) between entity frequency ($Freq$) and LLM response uncertainty ($Entropy$). No clear or consistent correlation (Pearson or Spearman) is observed across domains.

Domain	Pearson	Spearman
Historical	0.0383	-0.3486
Art	-0.0008	0.1429
US	-0.2398	-0.2658
World	-0.0373	-0.1691

To rigorously isolate the effect of frequency, we stratified the entities into Low-Frequency (bottom two-thirds) and High-Frequency (top one-third) subsets and re-evaluated the correlations. As shown in Table 10 and 11, the strong negative correlation between probe accuracy and response uncertainty persists significantly across both subsets. This robustly demonstrates that **negative correlation is the inherit characteristic of the LLM's reasoning process, rather than a trivial artifact of statistical shortcut.**

Table 10: The Spearman score under the low-frequency subset $\rho(Uncertainty, Probe|\text{Low Freq})$. The low-frequency subset consists of the bottom two-thirds of entities by frequency, with the average count shown in parentheses after each dataset name.

Dataset	Probe R^2 vs $Entropy$	Probe sp vs $Entropy$
Historical (3.38)	-0.77	-0.66
Art (53.10)	-1.00	-1.00
US (13.54)	-0.95	-0.97
World (4.93)	-0.61	-0.35

Table 11: The Spearman score under the low-frequency subset $\rho(Uncertainty, Probe|\text{High Freq})$. The high-frequency subset consists of the top two-thirds of entities by frequency, with the average count shown in parentheses after each dataset name.

Dataset	ProbeR^2 vs $Entropy$	Probe sp vs $Entropy$
Historical (1421.68)	-0.88	-0.70
Art (75442.00)	-1.00	-1.00
US (9510.12)	-0.96	-0.95
World (813.18)	-0.45	-0.28

C.8 Predictive Utility of Probes for Uncertainty Estimation

To directly address the practical utility, we evaluate whether a trained probe can actively predict LLM generation uncertainty without relying on the computational overhead of multiple sampling. For simplicity, we formulate this uncertainty estimation as a binary classification task.

We use the median uncertainty score within each dataset as a threshold to binarize the samples into high-uncertainty and low-uncertainty classes. The data is partitioned into training and testing sets with an 8:2 ratio. We train the probe to predict the uncertainty class relying solely on internal entity representations, reporting its classification performance on the held-out test set. As shown in Table 12, the internal representations exhibit a strong potential for indicating uncertainty. Across all datasets, the probe achieves robust predictive performance, with AUROC scores consistently ranging from approximately 0.76 to 0.87.

To further validate this capability, we adopt a more rigorous experimental setting by discarding the ambiguous middle cases. We retain only the top 30% of samples (highest uncertainty) and the bottom 30% (lowest uncertainty) and retrain the probe following the identical training and testing split strategy.

As demonstrated in Table 13, when trained on these distinct polarities, the probe's predictive performance becomes substantially stronger. The AU-

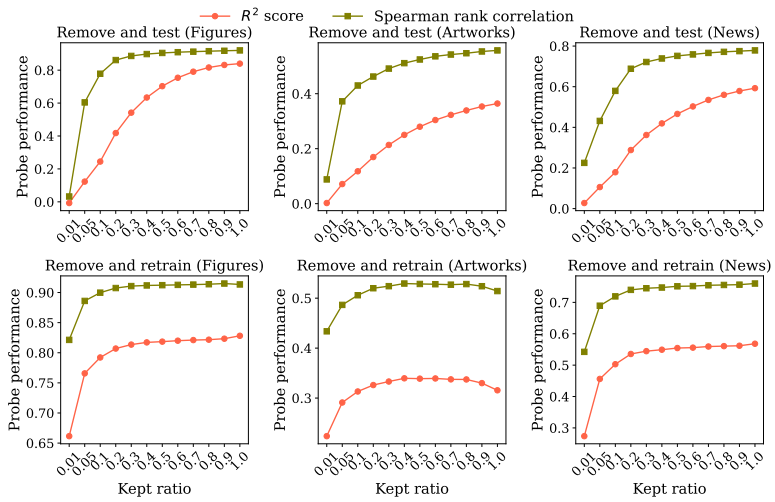


Figure 10: Correlation experiments from lens of feature importance on Llama 2 (14B) model. x-axis represents the ratio of important features preserved, from 0.01 to 1 and y-axis denotes the resulting probe performance. In the top panels, the probe weights remain fixed; in the bottom panels, the probe is retrained on the masked data.

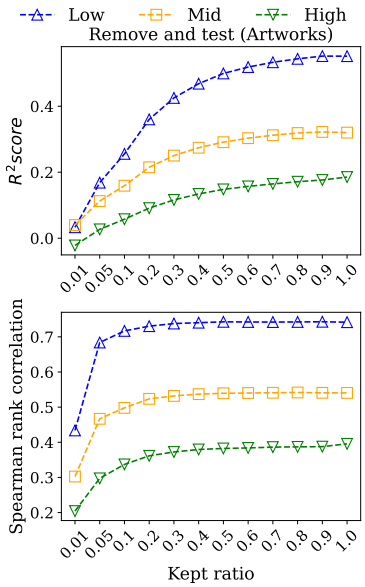


Figure 11: Correlation experiments from lens of feature importance on Llama 3.1 (8B). Three probes are trained on subsets stratified by varying uncertainty levels (measured by variance) in Artworks. We observe that on the low uncertainty subset, only a small number of salient features are needed to reach near-optimal probe performance.

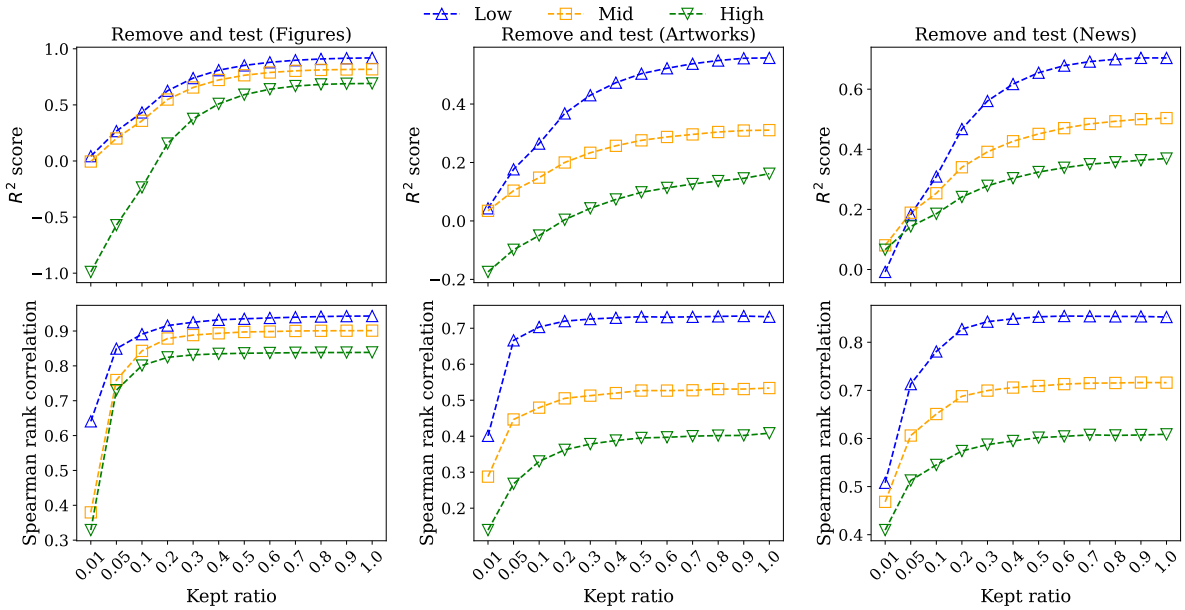


Figure 12: Correlation experiments from lens of feature importance on Llama 3.1 (8B). Three probes are trained on subsets stratified by varying uncertainty levels (measured by entropy) in Artworks. We also observe that on the low uncertainty subset, only a small number of features are needed to reach near-optimal probe performance.

Table 12: Binary classification performance of the probe using the median uncertainty threshold.

Dataset	AUROC	Accuracy	F1-Score
Historical	0.8053	0.7729	0.7728
Headline	0.7663	0.6988	0.6982
Art	0.8058	0.7439	0.7438
NYC	0.8734	0.8151	0.8157
US	0.8001	0.7389	0.7380
World	0.8612	0.7827	0.7828

Table 13: Binary classification performance of the probe trained exclusively on the top 30% (high) vs. bottom 30% (low) uncertainty samples.

Dataset	AUROC	Accuracy	F1-Score
Historical	0.8942	0.8256	0.8247
Headline	0.8169	0.7457	0.7444
Art	0.8424	0.7983	0.7963
NYC	0.9525	0.8956	0.8953
US	0.8313	0.7800	0.7764
World	0.9184	0.8573	0.8568

ROC scores increase across all domains, with the highest performance achieved on the NYC dataset (0.9525). These results indicates that lightweight probes can distinguish between high and low uncertainty states using only single-forward-pass internal representations

C.9 Nonlinear Probes and Uncertainty Metrics Evaluation

To validate our findings under nonlinear conditions, we replaced the linear probe with 2-layer and 3-layer nonlinear Multilayer Perceptrons (MLPs) and calculated the corresponding Spearman coefficients and R^2 values on the LLaMA 3.1 8B datasets.

- 2-layer MLP architecture: $input \times 16 \times 1$.
- 3-layer MLP architecture: $input \times 16 \times 2 \times 1$.

The detailed experimental results are presented in Table 14. **Our observations indicate consistent negative correlation across these nonlinear probes, demonstrating that the relationship between probe performance and response uncertainty (measured using variance) remains robust**. The specific network architectures utilized in this experiment are as follows:

C.10 Spherical distance as Uncertainty Estimator on Space Datasets

Below, we report results using spherical distance (SD). As shown in the table, **the observed correlations remain consistent**.

Table 15: Spearman coefficients between probe performance and response uncertainty (measured by variance of spherical distance).

Dataset	Probe R^2 vs SD	Probe Spearman vs SD
NYC	-0.81	-0.89
US	-0.98	-0.95
World	-0.70	-0.98

Table 14: Evaluation results of nonlinear MLP probes on LLaMA 3.1 8B datasets (uncertainty measured using variance)

Dataset	**2-MLP (R^2 vs Var)**	**2-MLP (Spearman vs Var)**	**3-MLP (R^2 vs Var)**	**3-MLP (Spearman vs Var)**
Historical	-0.87	-0.76	-0.91	-0.80
Art	-0.87	-0.99	-0.92	-0.99
Headline	-0.72	-0.73	-0.69	-0.70