Investigating Hallucinations in Time Series Foundation Models through Signal Subspace Analysis

Abstract

Times series foundation models (TSFMs) have emerged as a promising paradigm for time series analysis, showing remarkable generalization performance across different domains. While there has been research on hallucinations in foundation models, hallucinations in TSFMs are underexplored. In this paper, we formally define TSFM hallucinations in the zero-shot forecasting setting by examining whether a generated forecast exhibits different dynamics from those of the context. Our study reveals that TSFM hallucinations are mainly caused by the loss of context information in hidden states during forward propagation. As such, we propose methodologies to identify signal subspaces in TSFMs and magnify the signal subspace information through intervention. Extensive experiments demonstrate that our proposed intervention approach effectively mitigates hallucinations and improves forecast performance. Furthermore, the signal strength measure we compute from signal subspaces has strong predictive power of hallucinations and forecast performance of the model. Our work contributes to deeper understanding of TSFM trustworthiness that could foster future research in this direction.

1 Introduction

Times series analysis is a major research field that facilitates decision making and scientific inference across a broad range of domains, from energy and weather to economy, transport, and system management. As a key task, time series forecasting has motivated the development of distinct approaches including statistical models [6, 32] and deep learning models [28, 40]. Despite competitive performance on specific tasks, these models are typically trained on a single domain, without sufficient capability to generalize to different domains. Inspired by the success of foundation models in fields like natural language processing (NLP) [1, 7], time series foundation models (TSFMs) has recently emerged as a new paradigm towards universal forecasters [4, 8, 11, 16, 24, 30, 39]. By pretraining on large-scale TS data, TSFMs have shown remarkable few-shot and even zero-shot forecasting performance across multiple domains [3, 18], substantially reducing the need for downstream data. The hidden representations of TSFMs are also valuable for downstream tasks by capturing the context TS information.

Yet, the reliability of TSFMs is often hampered by hallucinations, as with other foundation models. Broadly referring to the generation of unsupported statements or nonsensical content, hallucinations primarily stem from a lack of correct knowledge or insufficient inference capability of the model [19]. Among numerous hallucination detection and mitigation approaches proposed, intervention is a powerful mitigation approach requiring no additional training that has demonstrated effectiveness for large language models (LLMs) [25, 31, 43] and vision-language models (LVLMs) [23, 42].

In zero-shot forecasting, since a TSFM is tasked with generating extrapolations based on the extracted information of the context TS such as trends, periodicity, and patterns [18], accurately processing

the context information is essential for generating high-quality forecasts. As such, we study TSFM hallucinations from the perspective of whether a forecast exhibits drastically different dynamics from those of the context, e.g., Figure 1 (a) versus (b). We investigate the underlying mechanisms of TSFM hallucinations through the lens of hidden representations and develop a novel intervention approach to address the identified causes. As far as we know, little has been explored on similar research problems in existing literature. We strive to address these knowledge gaps and contribute to deeper understanding of TSFM trustworthiness that could foster future research in this direction.

We formally define TSFMs hallucination in the zero-shot forecasting setting in §3 and outline the rules for checking hallucinations in practice. In §4.1, we build insights on TSFM hallucinations through experimental analyses, where we find that hallucinations are mainly caused by a lack of context information in hidden states during forward propagation. We then propose a methodology to identify the signal spaces and a measure, SSAS, to quantify the signal strength of hidden states in §4.2. Build upon these results, we propose a novel intervention approach, SSIM, which mitigates hallucinations by magnifying the signal information of hidden states in §4.3. Extensive experiments in §5.2 demonstrate that the forecasting performance of TSFMs suffers from hallucinations and our intervention approach effectively mitigates hallucinations and improves the quality of forecasts, yielding up to 6.62% reduction on hallucination rate, 93.83% gain on R^2, and 13.52% gain on correlation. Moreover, the signal strength measure we propose has strong predictive power of both hallucinations and the forecast performance of TSFMs.

Our main contributions in this work are: (1) We formally define the problem of hallucinations in TSFMs along with a set of procedures to check hallucinations. We are the first to systematically study this problem to our best knowledge. (2) We propose a methodology to identify the signal subspaces in TSFMs and a measure to quantify the signal strength in TSFM hidden states. (3) We propose a simple and efficient intervention approach to mitigate hallucinations by magnifying the signal information in hidden states. (4) We conduct extensive experiments on both synthetic and real-world datasets to demonstrate the effectiveness of our proposed signal information measure and intervention approach.

2 Related Work

Times series foundation models. TSFMs represent a promising paradigm towards generalization across different TS domains and tasks by leveraging the knowledge from large-scale pretrained data [4, 8, 11, 16, 24, 30, 39]. TSFMs not only substantially reduce the need for downstream data but have also shown capabilities of producing accurate forecasts even in zero-shot scenarios, where forecasts are made on inputs from previously unseen domains [3, 18]. While most TSFMs are Transformer based [33] and open sourced, they are diverse in architectural design, tokenization strategies, and pretraining objectives. For instance, Chronos [4] and Chronos-Bolt adopt encoder-decoder architecture, while TimesFM [11] is decoder-only. Chronos-Bolt and TimesFM truncate the normalized TS inputs into patches, while Chronos discretely quantizes the scaled inputs into a fixed vocabulary. Yet, the forecasting performance of TSFMs suffers from hallucinations when they fail to capture enough signal information from the inputs. We study this issue on models from both families.

Hallucinations. Hallucination, defined as the generation of unfaithful or nonsensical content, is a fundamental challenge in Large Foundation Models due to their black-box nature [19]. Recent research has examined models' hidden representations for hallucination detection and mitigation, based on the hypothesis that factual knowledge is encoded in these states [10, 12, 15]. Studies have identified diagnostic signals in hidden states, showing that outlier or inconsistent activation patterns during generation can indicate potential hallucinations [2, 9, 13, 31, 36]. Complementary approaches focus on hidden state manipulation, demonstrating that truthfulness can be elicited through targeted neuron activation interventions, offering promising directions for reducing hallucinations [22, 23, 31, 41, 42]. We are the first to formally define and systematically study hallucinations in time series foundation models. We develop methodologies to both detect and mitigate TSFM hallucinations.

Intervention. Hidden state intervention has emerged as a powerful technique for controlling neural models' behavior, as these internal representations serve as causal factors influencing model outputs. Research by [43], [25], and [21] demonstrates effective control over LLM outputs through activation steering, which identifies linear-interpretable directions in representation space and guides hidden states along these pathways. Some research achieves model output modification by selectively

masking specific neuron activations, preventing corresponding generations from occurring [29, 35]. The intervention approach for TSFMs proposed in [38] alters the outputs but does not address specific challenges of TS forecasting. Differently, we propose a novel intervention approach to specifically address TSFM hallucinations that is context adaptive and selectively intervenes model layers.

3 Definitions and Preliminaries

Formally, we describe the forecast of a time series foundation model and the problem of hallucinations.

Definition 1 (**TSFM forecasts**). A pretrained time series foundation model, denoted as $\mathcal{M}_\theta$, takes a time series $\boldsymbol{x}_{context} = [x_1, \ldots, x_p]$ of context length p as the input and generates a forecast $\hat{\boldsymbol{x}} = \mathcal{M}_\theta(\boldsymbol{x}_{context}) = [\hat{x}_{p+1}, \ldots, \hat{x}_{p+q}]$ of horizon q. For an L-layer time series foundation model, we denote the hidden states at different positions of layer l (the outputs of the layer) as a matrix $\boldsymbol{H}^{(l)} = [\boldsymbol{h}_1^{(l)}, \ldots, \boldsymbol{h}_n^{(l)}] \in \mathbb{R}^{n \times d}$, where d is the dimension of hidden states.

Definition 2 (**TS forecast hallucinations**). Suppose for a time series $\boldsymbol{x}_{full} = [x_1, \ldots, x_T]$, a knowledge set $\mathbb{K}$ can be inferred from a partial time series $\boldsymbol{x}_{context} = [x_i, \ldots, x_j]$, $1 \leq i < j < T$. The knowledge set $\mathbb{K}$ comprises time-dependent knowledge rules r that hold true for $\boldsymbol{x}_{full}$, i.e., $r(x_i, i) = 1$ for $x_i \in \boldsymbol{x}_{full}$, or simply $r(\boldsymbol{x}_{full}) = 1$. In zero-shot time series forecasting, we consider a hallucination to be a forecast that does not conform to the knowledge rules inferred from the context time series and define the set of hallucinations as $Hallu(\boldsymbol{x}_{context}) = \{\hat{\boldsymbol{x}} : \bigwedge_{r \in \mathbb{K}} r(\hat{\boldsymbol{x}}) = 0\}$.

Definition 3 (**Hallucination detection and mitigation**). The goal of hallucination detection is to define a score function f that discriminates hallucinated forecasts of the foundation model, such that for any $\hat{\boldsymbol{x}} = \mathcal{M}_\theta(\boldsymbol{x}_{context}) \in Hallu(\boldsymbol{x}_{context})$, we have $f(\boldsymbol{x}_{context}, \hat{\boldsymbol{x}}, \theta) > \tau$. We mitigate hallucinations through test-time intervention on hidden states so that with the intervention operation $\mathcal{I}$, we obtain non-hallucinated forecasts $\mathcal{M}_{\theta,\mathcal{I}}(\boldsymbol{x}_{context}) \notin Hallu(\boldsymbol{x}_{context})$.

In practice, we sequentially extract a set of knowledge rules $\mathbb{K} = \{r_1, \ldots, r_n\}$ from the context time series to check whether a forecast is hallucinated. Further details are provided in the appendix.

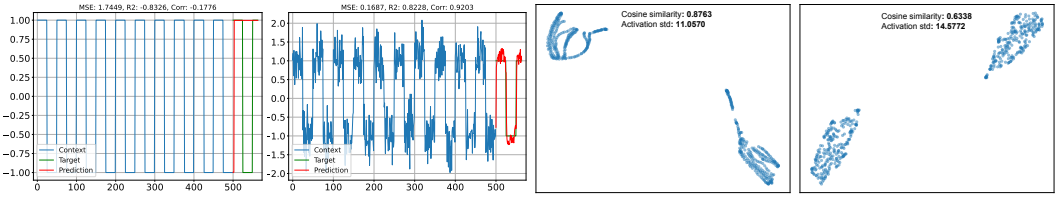
Trend. The trend rule checks whether the trend of the forecast conforms to those of the context. We perform ordinary least-square (OLS) regression on $\hat{\boldsymbol{x}}$ and take the first-degree coefficient c' as the trend if it is significant with the p-value < 0.01. We then perform OLS on rolling windows of $\boldsymbol{x}_{context}$ and take significant trends $[c_1, \ldots, c_n]$. With the relative difference between trends computed as $diff(c, c') = \left|\frac{c'}{c} - 1\right|$, the trend rule is satisfied if the minimum relative difference $\min_i diff(c_i, c') < \epsilon$, or neither the forecast nor the context has significant trends.

Frequency. The frequency rule checks whether the spectral density of the forecast conforms to those of the context. After removing the trend, we compute the spectral densities $[\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n]$ of rolling windows on $\boldsymbol{x}_{context}$ using short-time Fourier transform (STFT) [17] and also the spectral density $\boldsymbol{f}'$ of $\hat{\boldsymbol{x}}$. With the Jaccard distance between spectral densities computed as $\mathcal{D}(\boldsymbol{f}, \boldsymbol{f}') = 1 - \frac{\sum_i \min\{f_i, f_i'\}}{\sum_i \max\{f_i, f_i'\}}$, the frequency rule is satisfied if the minimum distance $\min_j \mathcal{D}(\boldsymbol{f}_j, \boldsymbol{f}') < \epsilon$.

Pattern. The pattern rule checks whether the pattern of the forecast is similar to those of the context. After removing the trend, we compute the relative absolute errors between the forecast and rolling windows $[\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n]$ on $\boldsymbol{x}_{context}$. With the relative absolute error computed as $RAE(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\sum_i |x_i - \hat{x}_i|}{\sum_i |x_i - \bar{x}|}$, the pattern rule is satisfied if the minimum relative error $\min_j RAE(\boldsymbol{w}_j, \hat{\boldsymbol{x}}) < \epsilon$.

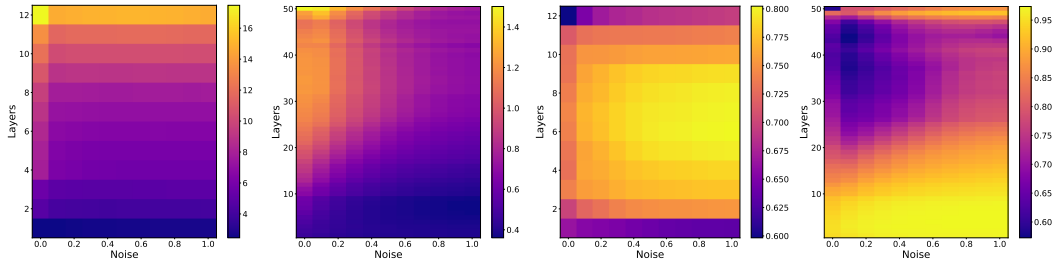
ARMA. The ARMA rule checks whether the ARMA dynamics of the forecast conform to those of the context, which complements the pattern rule since a TS that exhibits strong ARMA dynamics may not have distinct patterns. After removing the trend, we fit a first-order Autoregressive moving average (ARMA) model [6] on $\boldsymbol{x}_{context}$ and take the AR and MA coefficients ϕ and ψ if both are significant with p-values less than 0.01. Let ϕ' and ψ' be the first-order ARMA coefficients on $\hat{\boldsymbol{x}}$, the ARMA rule is satisfied if the relative differences $\left|\frac{\phi'}{\phi} - 1\right| < \epsilon$ and $\left|\frac{\psi'}{\psi} - 1\right| < \epsilon$.

A TSFM forecast that violates the trend, frequency, or both pattern and ARMA rules is considered to be hallucinated, since it is unsupported by the information encoded in the context.



(a) Hallucinated (b) Non-hallucinated (c) Hallucinated (d) Non-hallucinated

Figure 1: (a) (c) Examples of hallucinated and non-hallucinated forecasts from `Chronos`. (b) (d) The UMAP visualizations of hidden states at the last model layer and the statistics of hidden states.



(a) `Chronos` (b) `TimesFM` (c) `Chronos` (d) `TimesFM`

Figure 2: (a) (b) The standard deviations of hidden state activations across positions under varying noise magnitudes. (c) (d) The mean pairwise cosine similarities of hidden states across positions. The x-axis represents the standard deviation of the Gaussian noise added to the context signal.

4 Methodology

To understand the cause of hallucinations in TSFMs, we first build insights from observations, provide intuitive explanations, and then perform experimental analyses to justify our claims. Afterwards, we propose a signal information measure to help detect hallucinations. Finally, we develop a novel test-time intervention approach that mitigates hallucinations by addressing the identified causes.

4.1 Observations and Analyses

We begin with a brief case analysis. Figure 1 (a)(c) presents a hallucination example where the model fails to generate a forecast consistent with the context TS. The UMAP [26] visualization of hidden states at the last layer reveals irregular patterns, with high mean pairwise cosine similarity and low activation variance. We speculate that the forecast failure is caused by the loss of context information in hidden states during forward propagation. In comparison, we find that injecting a small amount of random perturbation to the context TS with Gaussian noise helps address such information loss, as shown in Figure 1 (b)(d). We observe that the hidden states are more evenly distributed in each cluster, with the mean pairwise similarity substantially reduced and activation variance increased.

To understand the effects of context signal and noise on the internal model, we present the results over 10 random perturbations of the context TS with Gaussian noise of varying magnitudes where no hallucination occurs. From Figure 2 (a)(b), we observe that hidden state activations exhibit the greatest variance across positions in the presence of clean signal. As the signal gets mixed with more noise, while the input variance increases, hidden state activations becomes less variant. The decline in activation variance with noise magnitude is more salient at higher layers, suggesting that the model incrementally extracts signals and reduces noises from the input by each layer.

Based on this, we posit that the hidden state space $\mathbb{H}^{(l)}$ at each TSFM layer can be decomposed into signal and noise subspaces $\mathbb{H}^{(l)} = \mathbb{S}^{(l)} \oplus \mathbb{N}^{(l)} \subset \mathbb{R}^d$, handling the signals and noises of the input [14, 20, 27]. In forward model propagation $\boldsymbol{H}^{(1)} \rightarrow \ldots \rightarrow \boldsymbol{H}^{(L)}$, the signal components of a hidden state $\Pi_{\mathbb{S}^{(l)}} \boldsymbol{h}^{(l)}$ are further processed by subsequent layers, while the noise components

$\Pi_{\mathbb{N}^{(l)}} \boldsymbol{h}^{(l)}$ get repressed and eventually removed. Since the signal components are more variant and dissimilar across hidden state positions than the noise components, the hidden states would exhibit greater distinctiveness across positions when the signal strength at a layer is strong (Figure 2 (c)(d)).

Back to the previous case of TSFM hallucinations, the inactivity of signal subspaces of the model leads to highly similar hidden states across positions. In this case, a proper amount of random perturbation injects input variance that helps activate the signal subspaces and facilitates the propagation of context signal information. Nonetheless, it is hard to determine the optimal amount of perturbation, since too much perturbation obscures the input signal and degrades forecast quality. Moreover, a single perturbation is not robust [23], while performing multiple perturbations hampers efficiency. As such, our goal is to magnify the signal information in hidden states through intervention, which would enable us to mitigate hallucinations in a controllable and efficient manner.

4.2 Signal Subspace Identification

Now, we develop a novel methodology to identify the signal subspaces in TSFM layers and provide empirical analysis. We aim to identify a set of hidden state neurons that are most active to context signals by examining the variance of activations across hidden state positions, enlightened by the associations between the activation variance and signal strength we observe in the previous subsection. The activity score of the j-th neuron at layer l given a context input $\boldsymbol{x}$ is computed as:

$$\mathcal{A}^{(l)}(j \mid \boldsymbol{x}) = \sqrt{\frac{1}{n}\sum_{i}(\boldsymbol{H}^{(l)}_{i,j} - \bar{\boldsymbol{h}}^{(l)}_{j})^2} \ . \tag{1}$$

The neuron activity measure we propose is more nuanced compared with the magnitude of neuron activations used in prior works [34, 35], which not only reflects the overall magnitude but also measures the deviation of neuron activations across TS steps.

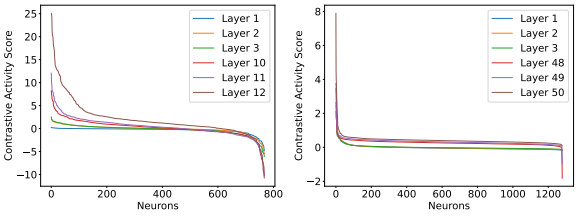
To measure neuron activity in the presence of signals, we collect the activity scores on a synthetic dataset comprising common waveforms that will be described in §5.1. We also vary the magnitude of noises injected to the context signals and initialize them with different random seeds for robustness. With $\mathcal{D}_{signal}$ denoting the set of synthetic TS inputs where no hallucination occurs, we consider neurons with the activity score consistently top ranked across the samples as candidate signal neurons, i.e., $Cand(l) = \bigcap_{\boldsymbol{x}\in\mathcal{D}_{signal}} \{j \mid rank(\mathcal{A}^{(l)}(j \mid \boldsymbol{x})) < \epsilon d\}$. We compute the signal activity score of each neuron using the sample mean $\mathcal{A}^{(l)}_{signal}(j) = \frac{1}{|\mathcal{D}_{signal}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{signal}} \mathcal{A}^{(l)}(j \mid \boldsymbol{x})$.

As hidden state neurons may fulfill multiple roles [5, 37, 42], e.g., processing signals and removing noises concurrently, we want to identify neurons that are primarily responsible for signal processing. To this end, we further collect the activity scores $\mathcal{A}^{(l)}_{noise}$ with Gaussian noises as the input by similar means and then compute the contrastive neuron activity score between signal and noise:

$$\mathcal{A}^{(l)}_{contrastive}(j) = \mathcal{A}^{(l)}_{signal}(j) - \mathcal{A}^{(l)}_{noise}(j) \ . \tag{2}$$

For each model layer $l \in \{1, \ldots, L\}$, we select the candidate neurons with top-ranked contrastive activity scores as the signal neurons, i.e., $Sig(l) = Cand(l) \cap \{j \mid rank(\mathcal{A}^{(l)}_{contrastive}(j)) < \epsilon d\}$. Figure 3 plots the distributions of the contrastive neuron activity scores across different model layers. We observe that at each layer only a small proportion of neurons are exclusively sensitive to signal or noise. Moreover, there is greater contrast at higher layers where neurons get more specialized.

Ranking signal neurons by contrastive activity score, we leverage the top signal neuron's activity score at the final layer as a measure of the strength of signal information the model has processed, i.e., $\mathcal{A}^{(L)}(j \mid \boldsymbol{x})$ with $j = Top_1(Sig(L))$. The final layer is selected as it shows the greatest contrast of neuron activity between signal and noise. We call the proposed measure Signal Subspace Activity Score (SSAS) and will verify its usefulness for



(a) `Chronos` (b) `TimesFM`

Figure 3: Distributions of contrastive activity scores.

TSFM hallucination detection and performance prediction in §5.2. With this, we claim that the model implicitly expresses in the hidden state subspaces how much signal information it is able to capture from the context.

4.3 Signal Subspace Intervention

Built upon the previous results, we propose a Center-Project-Scale (CPS) intervention operation to mitigate hallucinations by magnifying the signal information in hidden states. During forward propagation, for the hidden states $\boldsymbol{H}^{(l)} \in \mathbb{R}^{n \times d}$ at a TSFM layer, CPS works as follows:

1. Centering $\boldsymbol{H}^{(l)}$ by subtracting the mean across positions to obtain $\boldsymbol{H}_c^{(l)} = \boldsymbol{H}^{(l)} - \bar{\boldsymbol{h}}^{(l)}$;
2. Computing the projections on signal subspaces $\Pi_{\mathbb{S}^{(l)}} \boldsymbol{H}_c^{(l)}$ at all positions;
3. Scaling the signal components by a factor λ so that $\tilde{\boldsymbol{H}}_c^{(l)} = \boldsymbol{H}_c^{(l)} + (\lambda - 1)\Pi_{\mathbb{S}^{(l)}} \boldsymbol{H}_c^{(l)}$;
4. Adding back the mean to obtain the intervened hidden states $\tilde{\boldsymbol{H}}^{(l)} = \tilde{\boldsymbol{H}}_c^{(l)} + \bar{\boldsymbol{h}}^{(l)}$.

The intervened hidden states are passed as the inputs to the next layer. We center the hidden states in Step 1 to emphasize the activation differences across positions. Arranging the bases of $\mathbb{S}$ into a orthogonal matrix $\boldsymbol{P} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k] \in \mathbb{R}^{k \times d}$, where $\boldsymbol{e}_i$ is the indicator vector of a signal neuron, the projection in Step 2 can be computed by matrix product. The CPS operation can be formulated simply as $\tilde{\boldsymbol{H}}^{(l)} = \boldsymbol{H}^{(l)} + (\lambda - 1)\boldsymbol{H}_c^{(l)} \boldsymbol{P}^T \boldsymbol{P}$, which can be efficiently computed at each layer in $\mathcal{O}(ndk)$ cost, with $k \ll d$. The cost can be further reduced to $O(nk)$ leveraging the sparsity of $\boldsymbol{P}$.

The CPS operation has desirable properties. First, the mean of hidden state neuron activations is unaltered by the operation, while the standard deviation scales proportionally with λ, which makes the operation easy to control and causes no distribution drift to neuron activations. Moreover, different from previous intervention approaches of adding a static steering vector to hidden representations [21, 23, 38], our approach adaptively alters neuron activations based on their distributions, improving the contrast of hidden states and clustering effects. We mathematically show that in many cases the CPS operation can reduce the cosine similarity between two hidden states (see proofs in the appendix).

We further propose an adaptive scaling approach to help identify the scenarios when it is necessary to apply intervention and determine the scaling magnitude. Since the signal activity scores $\mathcal{A}_{signal}^{(l)}$ measure the neuron activity in the presence of strong signals, we use them as a reference. At each layer, we compute the mean activity scores of the signal neurons $\bar{\mathcal{A}}^{(l)}(\boldsymbol{x}) = \frac{1}{k} \sum_{j \in Sig(l)} \mathcal{A}^{(l)}(j \mid \boldsymbol{x})$. Then we compute the scaling factor as a ratio $\lambda^{(l)} = \bar{\mathcal{A}}_{signal}^{(l)} / \bar{\mathcal{A}}^{(l)}(\boldsymbol{x})$ and apply the intervention when $\lambda^{(l)} > 1$. In this way, we adaptively select the intervened layers with weak signal information and scale the activations of signal neurons to match those of the reference. We call the complete intervention approach Signal Subspace Intervention through Magnification (SSIM).

5 Experiments

In this section, we conduct experiments to address the following questions: (1) How do hallucinations affect the performance of each type of TSFM? (2) How is the effect of our proposed intervention approach on hallucination mitigation? (3) How is the performance of our proposed signal strength measures? (4) How do our designed components affect the intervention performance?

5.1 Experimental Settings

Datasets. We curate a synthetic dataset comprising common waveforms of sine, square, sawtooth, triangle, and pulse waves with varying slopes in $\{-0.01, 0, 0.01\}$. We vary the number of periods in the context in $\{8, 10, 12, 14, 16, 18, 20\}$ and the standard deviation of Gaussian noise added to the context signal in $\{0, 0.1, 0.2, 0.3, 0.4\}$. In addition, we adopt read-world datasets from GIFT-Eval [3] benchmark covering various domains. We take a fixed number of final observations from each time series, dividing them into context and ground truth of fixed lengths. We discard TS instances with over 10% missing values and impute missing values with the segment mean. As defined in §3, we retain TS instances whose ground truth satisfies the knowledge rules extracted from the context such

that the context contains sufficient information for forecasting. Each dataset is randomly split into validation (20%) and test (80%) sets. Further details are available in the appendix.

Baselines. For hallucination mitigation, we compare SSIM with input denoising by smoothing as well as input perturbation and output averaging [23]. For hallucination detection, we compare SSAS with the statistics discussed in §4.1, including the mean pairwise cosine similarity of hidden states and the mean standard deviation of neuron activations.

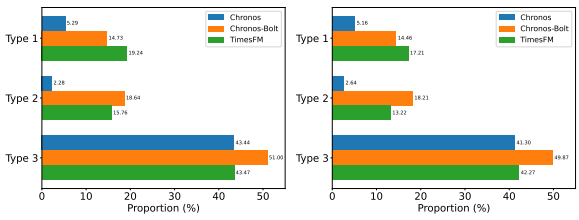
Evaluation metrics. We evaluate forecast quality with R^2 and Pearson correlation which are scale invariant. R^2 measures the goodness of fit to the ground truth; Pearson correlation measures the strength and direction of the linear relationship with the ground truth (invalid values are filled with 0). Whether a forecast is hallucinated is determined according to the knowledge rules defined in §3. We evaluate the effect of hallucination mitigation with hallucination rate reduction and forecast quality improvement. We evaluate the accuracy of hallucination detection with AUROC and performance prediction with Spearman rank correlation.

Implementation details. We evaluate on three mainstream TSFMs: *Chronos* [4], *Chronos-Bolt*, and *TimesFM* [11]. We set the context length to 500 and the forecast horizon to 64 for zero-shot TS forecasting in our main experiments, using the base versions of *Chronos* and *Chronos-Bolt together* with `TimesFM-2.0`. As *Chronos* produces probabilistic forecasts, we set the number of decoding samples to 1 and fix the random seed to ensure reproducibility. We set the frequency configuration of *TimesFM* to 0. For hallucination check, we set the thresholds of the trend, frequency, pattern, and ARMA rules to 0.25, 0.5, 0.5, and 0.25 respectively based on validation. For SSIM, we perform grid search for the proportion of selected top neurons $\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and set it to 0.1 for Chronos and TimesFM and 0.2 for Chronos-Bolt based on validation. For baselines methods, we denoise input TS using the mean of sliding windows of size 5. We perturb input TS by Gaussian noise with a standard deviation of 0.05 times that of the input and repeat for 10 runs. We use the signal strength measures for performance prediction and their negations for hallucination detection.

5.2 Main Experimental Results

TSFM hallucinations (*RQ1*). Table 1 summarizes the performance of original TSFMs. We note that the hallucination rate varies drastically across domains. On *Energy* domain the TS have more periodic patterns, while on *Nature* domain the TS contain more abrupt changes, making it harder to process the context information. The forecasts appear to have stronger correlations with the ground truths on domains where the hallucination rate is lower. Table 2 compares the performance of hallucinated forecasts versus non-hallucinated forecasts by TSFMs. We see that hallucinated forecasts are consistently outperformed by non-hallucinated forecasts. For Chronos-Bolt and TimesFM, the mean R^2 is positive on each domain when no hallucination occurs. Hallucinated forecasts have much weaker correlations with the ground truths than non-hallucinated forecasts for all models, indicating that the captured context signal information is not as strong. The χ^2 test yields $p < 10^5$ against the null hypothesis that the performance of hallucinated and non-hallucinated forecasts is the same. These results show that hallucinations significantly impact the forecasting performance of TSFMs.

Figure 4 (a) compares the distributions of hallucinations in TSFMs, with Type 1 referring to the violation of the trend rule, Type 2 to the frequency rule, and Type 3 to both the pattern and ARMA rules. We observe that Type 3 hallucinations occur most frequently, since these rules require detailed inference on the context. Chronos suffers fewer Type 1 and 2 hallucinations than the other models as it does not apply patching, enabling more accurate capture of the trend and frequency information.



(a) Original (b) SSIM

Figure 4: Distributions of hallucinations.

Hallucination mitigation (*RQ2*). Table 1 compares the forecasting performance with SSIM and the baseline methods. SSIM attains the best performance overall, yields up to 6.62% reduction on

Table 1: Comparison of forecasting performance across domains, with the best results boldfaced.

Model	Domain	Original			Denoising			Perturbation+Averaging			SSIM (ours)		
		$Hal \downarrow$	$R^2 \uparrow$	$Corr \uparrow$	$Hal \downarrow$	$R^2 \uparrow$	$Corr \uparrow$	$Hal \downarrow$	$R^2 \uparrow$	$Corr \uparrow$	$Hal \downarrow$	$R^2 \uparrow$	$Corr \uparrow$
Chronos	Synthetic	0.4524	-0.1625	0.6265	0.4429	-0.6734	0.5714	0.4333	-0.1053	0.6392	**0.4145**	**0.1854**	**0.7150**
	Econ/Fin	0.4115	-3.3554	0.4751	0.5007	-4.5011	0.3413	0.4609	-3.5727	0.4735	**0.4061**	**-3.2037**	**0.5146**
	Energy	0.1389	-0.4839	0.7180	0.2504	-3.0764	0.5315	0.1212	-0.2073	0.7241	**0.1191**	**0.0268**	**0.7707**
	Nature	0.8035	-10.7283	0.0457	0.9514	-8.3558	0.0575	0.8436	-7.2973	0.0552	**0.6715**	**-0.7575**	**0.1082**
	Transport	0.4197	-1.6444	0.5127	0.7565	-1.4804	0.3295	0.4461	-1.4582	0.5315	**0.3938**	**-0.2221**	**0.6081**
	WebOps	**0.5801**	-414.8937	0.2762	0.8833	-139.8035	0.1559	0.6115	-79.3298	0.2822	0.6052	**-21.8389**	**0.3369**
Aggregated Mean		0.4531	-82.3762	0.4458	0.5991	-28.6399	0.3336	0.4759	-17.2700	0.4529	**0.4231**	**-5.0845**	**0.5061**
Chronos-Bolt	Synthetic	0.5381	0.0152	0.5589	0.5810	-0.0625	0.5302	0.5500	0.0099	0.5586	**0.5231**	**0.0238**	**0.5672**
	Econ/Fin	0.4856	-1.3759	0.5727	0.4870	**-1.2344**	0.4243	0.4911	-1.4191	**0.5929**	**0.4787**	-1.2891	0.5811
	Energy	0.0985	0.1499	0.7694	0.1712	-0.0411	0.6291	0.1002	0.1033	0.7671	**0.0843**	**0.1508**	**0.7765**
	Nature	0.9426	-0.0744	0.1400	0.9536	-0.6290	0.1057	0.9404	-0.0876	0.1462	**0.9316**	**-0.0657**	**0.1472**
	Transport	0.6684	0.2039	0.6501	0.8446	-0.2129	0.4072	0.6632	0.2015	0.6488	**0.6522**	**0.2124**	**0.6563**
	WebOps	0.6777	-0.6529	0.3591	0.9024	-1.2424	0.1777	0.6707	**-0.4822**	0.3625	**0.6632**	-0.6680	**0.3646**
Aggregated Mean		0.5308	-0.4260	0.5099	0.6084	-0.6662	0.3848	0.5321	-0.4163	0.5158	**0.5191**	**-0.3766**	**0.5171**
TimesFM	Synthetic	0.1143	0.5661	0.9143	0.1452	0.4685	0.7568	0.1190	0.5688	0.9094	**0.1049**	**0.5699**	**0.9194**
	Econ/Fin	0.3868	-1.8715	0.7793	0.4472	-5.1532	0.3722	0.4005	-1.8277	0.7657	**0.3771**	**-0.3161**	**0.7847**
	Energy	0.1357	0.2745	0.8065	0.1987	-0.0981	0.6150	0.1341	**0.2916**	0.7941	**0.1222**	0.1304	**0.8133**
	Nature	0.9558	-0.1678	**0.1620**	0.9691	-0.2561	0.1302	**0.9492**	-0.1715	0.1451	0.9536	**-0.0902**	0.1559
	Transport	0.5751	**0.4245**	0.7027	0.6321	-0.4210	0.3540	**0.5648**	0.4236	0.7028	0.5733	0.4201	**0.7076**
	WebOps	0.6429	-22.0090	0.4224	0.8676	-20.0077	0.2324	**0.6202**	-7.5448	0.4216	0.6359	**-6.7480**	**0.4291**
Aggregated Mean		0.4441	-4.5461	0.6368	0.5251	-5.3271	0.4119	0.4418	-2.0435	0.6275	**0.4348**	**-1.2529**	**0.6410**

Table 2: Performance comparison of hallucinated and non-hallucinated forecasts by TSFMs.

Metric	Domain	Chronos			Chronos-Bolt			TimesFM		
		Hal	*Non-hal*	*Diff*	*Hal*	*Non-hal*	*Diff*	*Hal*	*Non-hal*	*Diff*
R^2	Synthetic	-1.1206	0.6290	1.7496	-0.4450	0.5512	0.9962	-1.5404	0.8379	2.3783
	Econ/Fin	-6.7964	-0.9492	5.8473	-3.3098	0.4497	3.7595	-5.7181	0.5553	6.2734
	Energy	-2.8794	-0.0974	2.7820	-1.1625	0.2934	1.4559	-0.8971	0.4585	1.3556
	Nature	-12.8573	-2.0209	10.8364	-0.0846	0.0933	0.1779	-0.1923	0.3626	0.5549
	Transport	-3.9208	0.0018	3.9226	-0.0548	0.7253	0.7800	0.2046	0.7223	0.5178
	WebOps	-634.8971	-110.9056	523.9916	-1.1848	0.4657	1.6505	-34.4655	0.4127	34.8783
Aggregated Mean		-161.6823	-16.6599	145.0224	-1.1647	0.4096	1.5743	-10.9572	0.5757	11.5329
$Corr$	Synthetic	0.3478	0.8568	0.5090	0.3596	0.7910	0.4314	0.7436	0.9364	0.1927
	Econ/Fin	0.0853	0.7476	0.6624	0.2558	0.8719	0.6162	0.5927	0.8969	0.3042
	Energy	0.4738	0.7574	0.2836	0.5716	0.7910	0.2194	0.6300	0.8342	0.2043
	Nature	0.0226	0.1401	0.1176	0.0991	0.8118	0.7127	0.1312	0.8289	0.6977
	Transport	0.2776	0.6827	0.4051	0.5407	0.8708	0.3301	0.5744	0.8765	0.3021
	WebOps	0.1036	0.5148	0.4113	0.1760	0.7441	0.5681	0.2135	0.7984	0.5849
Aggregated Mean		0.1459	0.6943	0.5484	0.2441	0.8105	0.5664	0.3429	0.8716	0.5286

hallucination rate, 93.83% gain on R^2, and 13.52% gain on correlation over the original models. While denoising improves R^2 in some cases by reducing the impact of outliers, it leads to higher hallucination rate and lower correlation in general due to the loss of context information. Perturbation averaging improves the forecast quality to some extent, but it does not sufficiently address hallucinations and demands considerably more computation. In comparison, SSIM pre-computes the signal neurons only once for each TSFM and incurs minor additional overheads during test time. The performance margin between SSIM and baselines is statistically significant with $p < 0.01$ by Friedman-Nemenyi test. We also analyze the distributions of hallucinations after SSIM. From Figure 4 (b), we observe that SSIM has the greatest impact on Type 3 hallucinations. By improving the propagation of signal information, the TSFM better captures patterns in the context.

Hallucination detection and performance prediction (*RQ3*). We summarizes the performance of different measures on hallucination detection and forecast performance prediction in Table 3. SSAS has consistently strong predictive power across domains for different TSFMs, with high AUROC for hallucination detection and significantly positive rank correlations with the forecast performance, demonstrating the effectiveness of of our proposed signal strength measure and the critical role of signal neurons in generating reliable forecasts. Simply using the mean neuron activation variance as a measure yields inferior and less consistent results overall, as it is obscured by the activity of irrelevant neurons. While the mean cosine similarity of hidden states shows relatively strong predictive power for Chronos and Chronos-Bolt, it fails to generalize to TimesFM.

Ablation study (*RQ4*). We compare the performance SSIM intervention with the following variants: (1) w/o adaptive scaling: using a constant scaling factor λ for each layer; (2) w/o centering: scaling neuron activations without subtracting the mean across positions [10, 35]. From Figure 5, SSIM

Table 3: The results of hallucination detection and forecast performance prediction for TSFMs, with the best results highlighted in boldface. For each compared method, the first column shows AUROC and the latter two columns show rank correlations. The statistical significance of positive rank correlation is indicated with * for $p < 0.05$ and ** for $p < 0.01$.

Model	Domain	Cosine Similarity			Activation Variance			SSAS (Ours)		
		Hal	R^2	$Corr$	Hal	R^2	$Corr$	Hal	R^2	$Corr$
Chronos	Synthetic	0.7847	0.3834**	0.3262**	0.6786	0.3734**	0.4052**	**0.8316**	**0.4299****	**0.5111****
	Econ/Fin	**0.8495**	**0.6501****	**0.6208****	0.6927	0.5096**	0.5507**	0.7833	0.5034**	0.5258**
	Energy	0.7124	**0.5088****	**0.3528****	0.8093	–0.1116	0.0373	**0.8096**	0.1166**	0.0363
	Nature	0.4978	0.3382**	0.1524**	0.5384	**0.3490****	**0.1886****	**0.5925**	0.3430**	0.1507**
	Transport	0.6601	0.4706**	0.5550**	**0.7466**	**0.5351****	**0.6083****	0.6767	0.4158**	0.5234**
	WebOps	0.5542	**0.2328****	**0.3710****	0.5060	0.1363**	0.2720**	**0.5740**	0.1693**	0.2526**
	Aggregated	0.7903	**0.5866****	**0.5804****	0.7226	0.4197**	0.5277**	**0.8086**	0.5082**	0.5758**
Chronos-Bolt	Synthetic	0.4416	–0.1767	–0.0361	0.4011	–0.3806	–0.3569	**0.5282**	**0.2363****	**0.2142****
	Econ/Fin	0.2247	–0.5253	–0.5530	0.3851	–0.3600	–0.4283	**0.8528**	**0.6289****	**0.5979****
	Energy	0.5360	0.2052**	0.1488**	0.4739	**0.4158****	**0.3741****	**0.7451**	–0.0603	–0.0865
	Nature	0.9343	–0.1155	0.3819**	**0.9395**	–0.1003	**0.3842****	0.6340	**0.0895***	0.1690**
	Transport	**0.7183**	0.2601**	0.2394**	0.6656	0.2730**	0.2064**	0.6416	**0.2850****	**0.2858****
	WebOps	**0.6656**	0.1315**	**0.4091****	0.5224	–0.0236	0.1767**	0.6518	**0.2188****	0.3423**
	Aggregated	0.6279	0.0462*	0.2607**	0.6131	0.0789**	0.1840**	**0.7991**	**0.3860****	**0.5037****
TimesFM	Synthetic	0.4892	–0.2210	–0.1684	0.3821	–0.3302	–0.2738	**0.4902**	**–0.0081**	**–0.0359**
	Econ/Fin	0.2210	–0.6896	–0.6568	0.1930	–0.5539	–0.5192	**0.4625**	**–0.0932**	**–0.0987**
	Energy	0.2898	–0.2597	–0.1570	0.2777	–0.1531	–0.1911	**0.7469**	**0.1407****	**0.1042***
	Nature	0.4042	–0.0565	–0.0041	0.7912	0.0587	0.3480**	**0.8934**	**0.1744****	**0.3870****
	Transport	0.4881	–0.0669	–0.0174	0.6121	0.1892**	0.2710**	**0.6529**	**0.3483****	**0.3622****
	WebOps	0.3756	–0.3497	–0.2117	0.5615	–0.1001	0.1046*	**0.6365**	**0.0637**	**0.2418****
	Aggregated	0.3963	–0.2608	–0.1767	0.5211	–0.0517	0.0634*	**0.6890**	**0.2636****	**0.3552****

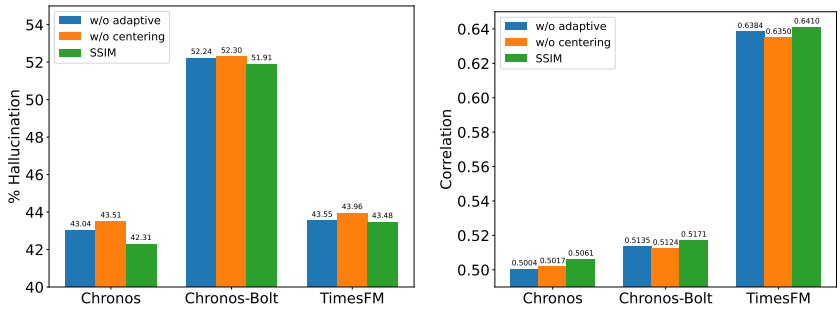


Figure 5: The aggregated mean performance of SSIM and the variants for TSFMs.

consistently outperforms the variants. The performance differences are significant with $p < 0.01$ by paired t-tests, highlighting the effectiveness of our design. The adaptive scaling enables more detailed control of the intervention, providing greater magnification when weak signal information is detected at a layer. The centering operation emphasizes activation differences that facilitate reducing the similarity between hidden states and avoids changing the mean activations of the intervened neurons.

6 Conclusion

TSFMs represent a promising paradigm for time series analysis, yet the issue of hallucinations has been unexplored in existing literature. We have formally defined TSFMs hallucination in the zero-shot forecasting setting and outlined the rules for checking hallucinations in practice. We have found that hallucinations are mainly caused by a lack of context information in hidden states through experimental analyses. We have proposed a methodology to identify the signal spaces and a measure to quantify the signal strength of hidden states. We have further developed an intervention approach that mitigates hallucinations by magnifying the signal information of hidden states. Extensive experiments have demonstrated that our intervention approach effectively mitigates hallucinations and improves the quality of forecasts of TSFMs. The signal strength measure we proposed has shown strong predictive power of both hallucinations and the forecast performance. Our work contributes to deeper understanding of TSFM trustworthiness that could foster future research in this direction.

References

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

[2] Sumukh K Aithal, Pratyush Maini, Zachary Chase Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[3] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

[4] Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

[5] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2, 2023.

[6] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, 2020.

[8] Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024.

[9] Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*, 2024.

[10] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022.

[11] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, pages 10148–10167. PMLR, 2024.

[12] Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*, 2021.

[13] Xuefeng Du, Chaowei Xiao, and Yixuan Li. Haloscope: Harnessing unlabeled LLM generations for hallucination detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[14] Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[15] Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

[16] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, pages 16115–16152. PMLR, 2024.

[17] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

[18] Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

[20] Shuyang Jiang, Yusheng Liao, Ya Zhang, Yanfeng Wang, and Yu Wang. Fine-tuning with reserved majority for noise reduction. In *The Thirteenth International Conference on Learning Representations*, 2025.

[21] Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

[23] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *International Conference on Machine Learning*, pages 32369–32399. PMLR, 2024.

[25] Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and Rene Vidal. PaCE: Parsimonious concept engineering for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[26] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[27] Tam Minh Nguyen, Tan Minh Nguyen, and Richard Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[28] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

[29] Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. Finding and editing multi-modal neurons in pre-trained transformers. *arXiv preprint arXiv:2311.07470*, 2023.

[30] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, et al. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

[31] ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*, 2025.

[32] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[34] Chengxin Wang, Yiran Zhao, Shaofeng Cai, and Gary Tan. Investigating pattern neurons in urban time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025.

[35] Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *The Thirteenth International Conference on Learning Representations*, 2025.

[36] Yiming Wang, Pei Zhang, Baosong Yang, Derek F Wong, and Rui Wang. Latent space chain-of-embedding enables output-free llm self-evaluation. *arXiv preprint arXiv:2410.13640*, 2024.

[37] Zijian Wang, Britney Whyte, and Chang Xu. Locating and extracting relational concepts in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4818–4832, 2024.

[38] Michał Wiliński, Mononito Goswami, Nina Żukowska, Willa Potosnak, and Artur Dubrawski. Exploring representations and interventions in time series foundation models. *arXiv preprint arXiv:2409.12915*, 2024.

[39] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, pages 53140–53164. PMLR, 2024.

[40] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

[41] Yuxin Xiao, Wan Chaoqun, Yonggang Zhang, Wenxiao Wang, Binbin Lin, Xiaofei He, Xu Shen, and Jieping Ye. Enhancing multiple dimensions of trustworthiness in llms via sparse activation control. *Advances in Neural Information Processing Systems*, 37:15730–15764, 2024.

[42] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025.

[43] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.