GUARANTEES ON DYNAMICAL SYSTEM DISTINGUISHABILITY FOR LLM TOKEN GENERATION

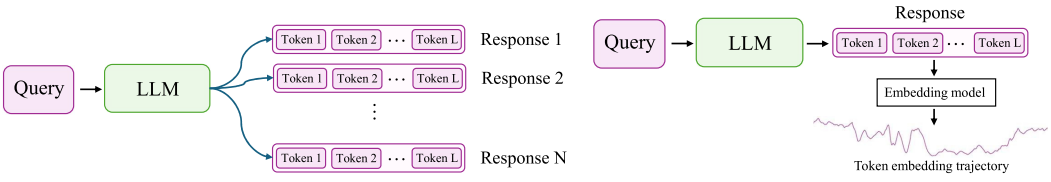
Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has shown that classifying large language models (LLMs)' responses can be distinguished by modeling token embeddings as trajectories of a black-box dynamical system (DS) and comparing prediction residuals of two DSs. Despite the empirical success of this dynamical approach, a theoretical understanding of *why* it works, *how well* it scales as a function of the token sequence, and *when* it transfers across embedding models remains lacking. We address these questions by formalizing the classification task as a binary hypothesis test between two stochastic linear DSs. We show that the total variation distance between the stationary marginal distributions of the two DSs can be arbitrarily small even when the dynamics differ substantially, which provides a fundamental accuracy floor for any classifier that ignores token dynamics. We then show that the misclassification probability of DS-based classification decays exponentially in the sequence length L, with the decay governed by a *dynamical discriminability* quantity δ^2 that captures the spectral distance between the two DSs. We also characterize cross-embedding generalization by introducing an *approximate intertwining* condition between embedding models and establishing a lower bound on the transferable discriminability in terms of the intertwining map's smallest singular value. These results explain the empirical performance of DS-based classification and motivate further investigation into using DS theory to analyze AI systems, in contrast to the more common approach of using AI to model dynamical systems [1].

1 INTRODUCTION

The problem of distinguishing between two dynamical systems from their output trajectories arises in numerous applications, including change-point detection in control systems (Xin et al., 2025), fault diagnosis in engineering (Sheikhi et al., 2026), and, more recently, hallucination detection in Large Language Models (LLMs) (Wilson & Akrout, 2026). In the LLM context, recent work has demonstrated that factual and hallucinated text can be distinguished by modeling the evolution of token embeddings as trajectories of a black-box dynamical system and fitting separate Koopman operator approximations for each regime (Wilson & Akrout, 2026).



(a) Multiple LLM responses (b) token dynamics of one LLM response

Figure 1: The two approaches to analyze the properties of LLM responses: (a) multiple responses where tokens are assumed i.i.d. within responses, and (b) one single response where the token dynamics is considered.

Existing approaches for classifying LLM responses (e.g., distinguishing safe from unsafe, or correct from hallucinated) require multiple responses to the same query as shown in Fig. 1(a) in order to

[1] Opus 4.8 was used in the preparation of this manuscript for text editing and grammar checking.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

make a reliable decision Lewis et al. (2020); Manakul et al. (2023); Zhang et al. (2023); Dhuliawala et al. (2024); Meng et al. (2022). This is because they treat the tokens within each generated sentence as independent and identically distributed (i.i.d.), and therefore ignore the embedding evolution of the tokens across the response. As a result, these methods must aggregate statistics over several independent generations to compensate for the loss of token dynamics, which increases both latency and API costs. By accounting for the rich token dynamics of a single LLM response in an embedding space as shown in Fig. 1(b).

Despite empirical demonstrations that token embedding dynamics enable LLM response classification (Wilson & Akrout, 2026), the theory behind this remains unexplored. This raises fundamental questions that we address in this work:

i) *Why do dynamics matter?* If the marginal distributions of token embeddings differ between two response categories, why is a dynamical model necessary?

ii) *How well does the dynamical system method work?* What is the classification error as a function of sequence length of the response, and what dynamical system quantities govern the classification performance?

iii) *Are embedding dynamics generalizable?* To what extent do token embeddings transfer across embedding models, and what conditions govern their cross-model generalization? From a DS perspective, embedding transferability is a striking and unexpected finding. This is because, at a fundimental level, data driven methods for Koopman operator approximation depend on the specific observables used in the model fitting. Even minor permutations to the ordering of the observable coordinates can ruin the predictive ability; switching to entirely different embedding model represents an even more drastic transformation. The fact that discriminative information survives such a change requires a theoretical explanation and has practical implications for reducing the cost of fitting models across different embedding backends.

In this paper, we provide rigorous answers to all three questions. Our contributions are:

- **Result 1 (Limitation of marginal density analysis):** We show that two stable DSs can have completely different dynamics yet produce stationary distributions that are nearly indistinguishable. This means that any classifier relying only on marginal statistics has a fundamental accuracy floor and cannot do better than random guessing when the distributions are too similar. This explains why static features like SVD modes often overlap for different embedding trajectory classes.

- **Result 2 (Dynamical separability):** We prove that a classifier based on prediction residuals can achieve misclassification error that decays exponentially with sequence length, with the rate decay determined by how dynamically distinct the two DSs are and the noise level. This breaks the accuracy ceiling of classifers with marginal density analysis obtained in Result 1. We also show that reliable classification with error probabilty ϵ requires a number of tokens $L^* = \mathcal{O}\big(\log \frac{1}{\epsilon}\big)$, which explains why richer embedding models already achieve high accuracy with shorter sequences.

- **Result 3 (Cross-embedding transfer):** We prove that dynamical systems fitted on one embedding can be transferred to another, provided the two embedding spaces share similar geometric structure described by an affine transformation. The lower-bound on the ability to discriminate between classes is dependent on the singular value of the transformation matrix and the norm of the operator difference of the two DSs. This explains why cross-embedding classification performs above chance even without retraining.

Notation: We use bold lowercase for vectors ($\mathbf{y}, \mathbf{z}$), bold uppercase for matrices ($\mathbf{A}, \mathbf{P}$), and calligraphic letters for manifolds and sets ($\mathcal{M}_c, \mathcal{M}_h$). The operator $\|\cdot\|$ denotes the Euclidean 2-norm for vectors and the spectral norm for matrices; $\|\cdot\|_F$ denotes the Frobenius norm. For a positive definite matrix $\mathbf{M}$, we write $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{\mathbf{x}^\top \mathbf{M}\, \mathbf{x}}$ for the Mahalanobis norm of $\mathbf{x}$. We write $\mathrm{tr}(\cdot)$ for the trace, $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ for the largest and smallest eigenvalues of a symmetric matrix, $\sigma_{\min}(\cdot)$ for the smallest singular value, and $\rho(\cdot)$ for the spectral radius. We use $\mathcal{O}(\cdot)$ to denote asymptotic upper bounds and $\Theta(\cdot)$ for asymptotically tight bounds in both directions, both up to constant factors.

2 PRIOR WORK

Distinguishability of dynamical systems. The problem of distinguishing between dynamical systems from observed data has a rich history in DS theory and signal processing. Classical system identification theory (Ljung, 1999) addresses the problem of estimating system parameters from input-output data, while the related problem of *hypothesis testing* between candidate models has been studied in the context of fault detection and isolation (Sheikhi et al. (2025)). In the linear setting, the question of whether two systems can be distinguished from output trajectories is closely tied to the notion of *observability* (Kailath (1980)): two systems with distinct dynamics but identical output distributions are observationally indistinguishable. Recent work has provided finite-sample guarantees for learning linear dynamical systems from a single trajectory (Simchowitz et al. (2018); Sarkar & Rakhlin (2019)), establishing that the sample complexity depends on the spectral radius and the signal-to-noise ratio of the system. Online change-point detection between linear dynamical regimes has been studied in (Xin et al. (2025)), who provide finite-sample guarantees for detecting switches between system parameters. Our work differs from these in that we characterize the error as a function of the dynamical discriminability between two dynamical systems.

Dynamical systems for neural networks. The perspective of treating neural network inference as a dynamical system has gained traction in recent years. Neural Ordinary Differential Equations (Chen et al. (2018)) established a formal connection between residual networks and continuous-time dynamical systems, while Koopman operator theory has been applied to analyze and predict the evolution of nonlinear systems from data (Budišić et al. (2012); Mezić (2013); Williams et al. (2015)). In the LLM context,, token embedding sequences of LLM responses were treated as trajectories of a black-box dynamical system and fitted to Koopman-based linear models for factual and hallucinated regimes Wilson & Akrout (2026), enabling single-response classification. However, the theoretical foundations of this approach remain unclear. The present work fills this gap.

3 PROBLEM SETUP AND OUTLINE

3.1 DYNAMICAL SYSTEMS FOR CLASSIFICATION

We consider a discrete-time system generating a sequence of observables $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L \in \mathbb{R}^d$ through an embedding map applied to the output of a generative model (e.g., an LLM). The observables are obtained by projecting the raw embeddings onto the top d singular value decomposition (SVD) modes:

$$\mathbf{y}_k = \mathbf{\Phi}^\top H(\mathbf{x}_k) \in \mathbb{R}^d, \tag{1}$$

where $\mathbf{x}_k$ is the internal state of the generative model, H is the composition of the token selection and embedding maps, and $\mathbf{\Phi} \in \mathbb{R}^{M\times d}$ is the SVD projection matrix with M being the raw embedding dimension. Following the Extended Dynamic Mode Decomposition (EDMD) framework (Williams et al., 2015), we lift the observables to a higher-dimensional space:

$$\mathbf{z}_k = \begin{bmatrix} \mathbf{y}_k \\ f_{\text{lift}}(\mathbf{y}_k) \end{bmatrix} \in \mathbb{R}^{d+\gamma}, \tag{2}$$

where $f_{\text{lift}} : \mathbb{R}^d \to \mathbb{R}^\gamma$ is a nonlinear lifting function (e.g., polynomial combinations of the observable components). A set of snapshot pairs $s_k = (\mathbf{z}_k, \mathbf{z}_{k+1})$ is collected and arranged into matrices $\boldsymbol{X} = [\mathbf{z}_1 \quad \ldots \quad \mathbf{z}_q]$ and $\boldsymbol{X}^+ = [\mathbf{z}_2 \quad \ldots \quad \mathbf{z}_{q+1}]$ where q is the number of snapshot pairs used. An approximation of the Koopman operator $\mathbf{z}_{k+1} = \mathbf{A}\mathbf{z}_k$ can be obtained according to $\mathbf{A} = \mathbf{X}^+\mathbf{X}^\dagger$, where † denotes the pseudoinverse. A prediction for the evolution of the token embeddings can be obtained according to $\tilde{\mathbf{y}}_{k+1} = [\boldsymbol{I} \quad \mathbf{0}]\, \boldsymbol{A}\boldsymbol{z}_k$ where $\boldsymbol{I} \in \mathbb{R}^{M\times M}$ denotes the identity matrix, $\mathbf{0} \in \mathbb{R}^{M\times\gamma}$ is a matrix of zeros.

In this work, we obtain distinct approximations of the Koopman operator, denoted $\mathbf{A}_1$ and $\mathbf{A}_2$, from data belonging to two different classes. In the LLM context, and without loss of generality, these correspond to the approximated operators $\mathbf{A}_h$ and $\mathbf{A}_c$ which we estimate from hallucinated and correct responses, respectively. Given an arbitrary LLM response, with sequences of token embeddings $\mathbf{z}_1, \ldots, \mathbf{z}_L$, the prediction errors under each model are:

$$\epsilon_{c,k} \equiv \|\mathbf{y}_{k+1} - \mathbf{P}\mathbf{A}_c\mathbf{z}_k\|\,, \tag{3}$$

$$\epsilon_{h,k} \equiv \|\mathbf{y}_{k+1} - \mathbf{P}\mathbf{A}_h\mathbf{z}_k\|\,, \tag{4}$$

where $\mathbf{P} = [\mathbf{I}_d \;\; \mathbf{0}_{d\times\gamma}]$ extracts the observable components. The cumulative squared errors are:

$$S_c = \sum_{k=1}^{L-1} \epsilon_{c,k}^2, \qquad S_h = \sum_{k=1}^{L-1} \epsilon_{h,k}^2, \tag{5}$$

and the differential residual score is:

$$\Delta\mathcal{E} = \sqrt{S_h} - \sqrt{S_c}. \tag{6}$$

The decision rule classifies the trajectory as hallucinated if $\Delta\mathcal{E} < \eta$ for a classification threshold η.

3.2 Classification as a hypothesis testing problem

We model the classification task as a binary hypothesis test. Under each hypothesis, the lifted observables evolve according to a stochastic linear dynamical system:

$$\mathcal{H}_c: \quad \mathbf{z}_{k+1} = \mathbf{A}_c\mathbf{z}_k + \boldsymbol{\xi}_k^{(c)}, \qquad \boldsymbol{\xi}_k^{(c)} \sim \mathcal{N}(\mathbf{m}_c, \boldsymbol{\Sigma}_c), \tag{7}$$

$$\mathcal{H}_h: \quad \mathbf{z}_{k+1} = \mathbf{A}_h\mathbf{z}_k + \boldsymbol{\xi}_k^{(h)}, \qquad \boldsymbol{\xi}_k^{(h)} \sim \mathcal{N}(\mathbf{m}_h, \boldsymbol{\Sigma}_h), \tag{8}$$

where $\mathbf{A}_c, \mathbf{A}_h \in \mathbb{R}^{(d+\gamma)\times(d+\gamma)}$ are the finite-dimensional approximations of the Koopman operator for the correct (factual) and hallucinated regimes, respectively. The noise terms $\boldsymbol{\xi}_k^{(c)}$ and $\boldsymbol{\xi}_k^{(h)}$ capture both the approximation error of the finite-dimensional Koopman truncation and any genuine stochasticity in the generation process. We allow each regime to have its own noise mean ($\mathbf{m}_c \in \mathbb{R}^{d+\gamma}$ or $\mathbf{m}_h \in \mathbb{R}^{d+\gamma}$) and its own noise covariance ($\boldsymbol{\Sigma}_c$ or $\boldsymbol{\Sigma}_h$), reflecting the fact that the fidelity of the linear approximation may differ between factual and hallucinated dynamics.

Assumption 1 (Stability)**.** *Both systems are stable:* $\rho(\mathbf{A}_c) < 1$ *and* $\rho(\mathbf{A}_h) < 1$.

Assumption 2 (Gaussian noise)**.** *The noise sequences* $\{\boldsymbol{\xi}_k^{(c)}\}$ *and* $\{\boldsymbol{\xi}_k^{(h)}\}$ *are i.i.d. Gaussian with means* $\mathbf{m}_c, \mathbf{m}_h \in \mathbb{R}^{d+\gamma}$ *and positive definite covariance matrices* $\boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_h \in \mathbb{R}^{(d+\gamma)\times(d+\gamma)}$*, respectively.*

Under Assumptions 1 and 2, each system possesses a unique stationary distribution. The *stationary means* are:

$$\boldsymbol{\mu}_c = (\mathbf{I} - \mathbf{A}_c)^{-1}\mathbf{m}_c, \qquad \boldsymbol{\mu}_h = (\mathbf{I} - \mathbf{A}_h)^{-1}\mathbf{m}_h, \tag{9}$$

where the invertibility of $\mathbf{I} - \mathbf{A}_c$ and $\mathbf{I} - \mathbf{A}_h$ is guaranteed by $\rho(\mathbf{A}_c) < 1$ and $\rho(\mathbf{A}_h) < 1$. The *stationary covariances* $\mathbf{G}_c$ and $\mathbf{G}_h$ are the unique positive definite solutions to the respective discrete Lyapunov equations:

$$\mathbf{G}_c = \mathbf{A}_c\mathbf{G}_c\mathbf{A}_c^\top + \boldsymbol{\Sigma}_c, \;\text{ and }\; \mathbf{G}_h = \mathbf{A}_h\mathbf{G}_h\mathbf{A}_h^\top + \boldsymbol{\Sigma}_h. \tag{10}$$

The covariances in (10) are equivalently given by the convergent series $\mathbf{G}_c = \sum_{j=0}^{\infty} \mathbf{A}_c^j\boldsymbol{\Sigma}_c(\mathbf{A}_c^j)^\top$ and $\mathbf{G}_h = \sum_{j=0}^{\infty} \mathbf{A}_h^j\boldsymbol{\Sigma}_h(\mathbf{A}_h^j)^\top$. Note that the existence, convergence, and uniqueness of these solutions is a classical result in linear systems theory (Hamilton (1994); Kailath (1980)).

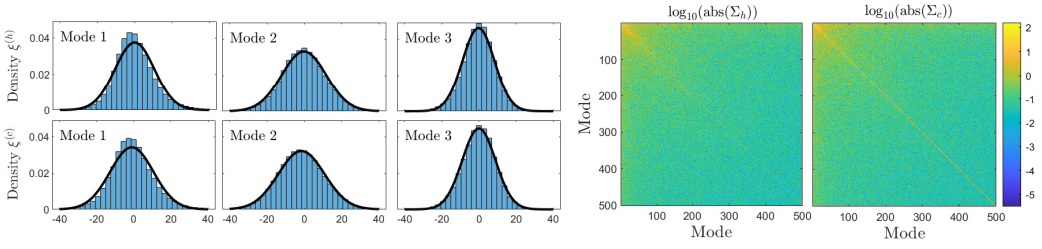


Figure 2: Extended DMD is applied to the HaluEval dataset Li et al. (2023) dataset. 80,000 samples of $\boldsymbol{\xi}_k^{(h)}$ and $\boldsymbol{\xi}_k^{(c)}$ are taken; the distributions are well-approximated by Gaussian curves (left) with mean and variance chosen appropriately. An estimate of the covariances is also provided (right).

When applied to the HaluEval Li et al. (2023) dataset, extended DMD yields predictive models that are well-approximated by equation 7 and equation 8. After fitting matrices $\mathbf{A}_h$ and $\mathbf{A}_c$ with 20% of

the dataset, 80,000 samples of $\boldsymbol{\xi}_k^{(h)}$ and $\boldsymbol{\xi}_k^{(c)}$ are taken over hallucinated responses, with distributions of the responses shown for the first 3 SVD modes in 2. Gaussian curves with identical mean and variance are superimposed highlighting that the distribution of errors is well approximated by a normal distribution. The noise covariance is also computed and included in Fig. 2.

3.3 Outline

Fig. 3 summarizes the logical dependencies between the key DS quantities already defined and our three main results.

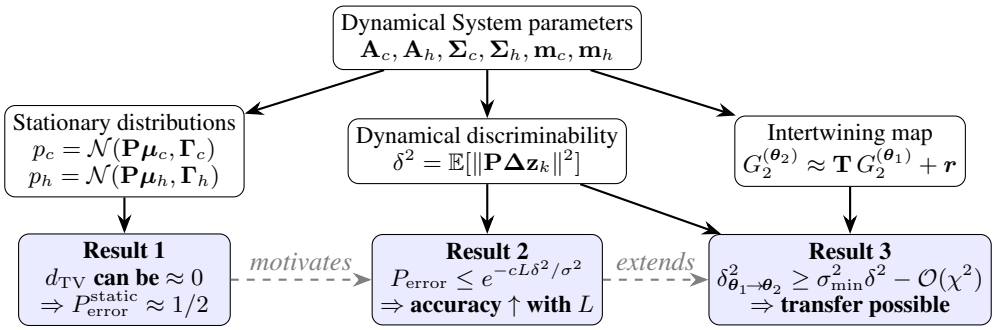


Figure 3: Logical dependencies between the key DS variables and the obtained results.

4 The Dynamical Discriminability

Given a high-dimensional embedding token sequence $\{\boldsymbol{z}_k\}_{\ell=1}^{L}$ of length L, we seek to determine whether it originates from dynamical system ($\mathcal{H}_c$ or $\mathcal{H}_h$) and characterize how the distinguishability between the two depends on L and the parameters of the dynamical systems in (7) and (8). Toward this goal, we define the *operator difference*:

$$\mathbf{\Delta} \equiv \mathbf{A}_c - \mathbf{A}_h. \tag{11}$$

Definition 1 (Dynamical discriminability)**.** *The* dynamical discriminability *of the embedding under $\mathcal{H}_c$ is:*

$$\delta^2 \equiv \mathbb{E}\Big[\|\mathbf{P}\mathbf{\Delta}\mathbf{z}_k\|^2\Big] = \mathrm{tr}\big(\mathbf{P}\mathbf{\Delta}\mathbf{G}_c\mathbf{\Delta}^\top\mathbf{P}^\top\big) + \|\mathbf{P}\mathbf{\Delta}\boldsymbol{\mu}_c\|^2\,, \tag{12}$$

where $\mathbf{G}_c$ is the stationary covariance and $\boldsymbol{\mu}_c$ is the stationary mean under $\mathcal{H}_c$. The mean-bias interaction *is:*

$$\beta_c \equiv 2\,\mathbf{m}_c^\top\mathbf{P}^\top\mathbf{P}\mathbf{\Delta}\boldsymbol{\mu}_c. \tag{13}$$

The discriminability $\delta^2 \geq 0$ measures the expected squared model-mismatch signal per time step and represents the systematic error that arises when the wrong model $\mathbf{A}_h$ is used to predict a trajectory generated by $\mathbf{A}_c$. The mean-bias interaction β_c (which can be positive, negative, or zero) captures the correlation between the noise mean and the model-mismatch signal evaluated at the stationary operating point. When the noise has zero mean ($\mathbf{m}_c = \mathbf{0}$), we have $\boldsymbol{\mu}_c = \mathbf{0}$ and $\beta_c = 0$, and the discriminability reduces to $\delta^2 = \mathrm{tr}(\mathbf{P}\mathbf{\Delta}\mathbf{G}_c\mathbf{\Delta}^\top\mathbf{P}^\top)$.

To build intuition, consider what happens when we use the *wrong* model to predict the next observation. Suppose the true dynamics follow $\mathcal{H}_c$, so that $\mathbf{z}_{k+1} = \mathbf{A}_c\mathbf{z}_k + \boldsymbol{\xi}_k^{(c)}$. If we instead predict using $\mathbf{A}_h$, the prediction for the observable component is $\widehat{\mathbf{y}}_{k+1} = \mathbf{P}\mathbf{A}_h\mathbf{z}_k$, while the true observable evolves as $\mathbf{y}_{k+1} = \mathbf{P}\mathbf{A}_c\mathbf{z}_k + \mathbf{P}\boldsymbol{\xi}_k^{(c)}$. The prediction error therefore decomposes as:

$$\mathbf{y}_{k+1} - \widehat{\mathbf{y}}_{k+1} = \mathbf{P}(\mathbf{A}_c - \mathbf{A}_h)\mathbf{z}_k + \mathbf{P}\boldsymbol{\xi}_k^{(c)} = \underbrace{\mathbf{P}\mathbf{\Delta}\mathbf{z}_k}_{\text{model-mismatch signal}} + \underbrace{\mathbf{P}\boldsymbol{\xi}_k^{(c)}}_{\text{noise}}. \tag{14}$$

The model-mismatch signal $\mathbf{P}\mathbf{\Delta}\mathbf{z}_k$ has two components: a random component driven by the stationary covariance $\mathbf{G}_c$, and a deterministic offset $\mathbf{P}\mathbf{\Delta}\boldsymbol{\mu}_c$ arising from the stationary mean. The expected squared magnitude of this signal is:

$$\begin{aligned}\delta^2 = \mathbb{E}\Big[\|\mathbf{P}\mathbf{\Delta}\mathbf{z}_k\|^2\Big] &\overset{(a)}{=} \mathbb{E}\big[\mathbf{z}_k^\top\mathbf{\Delta}^\top\mathbf{P}^\top\mathbf{P}\mathbf{\Delta}\mathbf{z}_k\big] = \mathrm{tr}\big(\mathbf{\Delta}^\top\mathbf{P}^\top\mathbf{P}\mathbf{\Delta}\,\mathbb{E}[\mathbf{z}_k\mathbf{z}_k^\top]\big) \\ &\overset{(b)}{=} \mathrm{tr}\big(\mathbf{P}\mathbf{\Delta}(\mathbf{G}_c + \boldsymbol{\mu}_c\boldsymbol{\mu}_c^\top)\mathbf{\Delta}^\top\mathbf{P}^\top\big)\,,\end{aligned} \tag{15}$$

where (a) follows from the identity $\mathbb{E}[\mathbf{z}^\top \mathbf{M}\, \mathbf{z}] = \mathrm{tr}(\mathbf{M}\, \mathbb{E}[\mathbf{z}\, \mathbf{z}^\top])$ and (b) uses the decomposition of the second moment $\mathbb{E}[\mathbf{z}_k \mathbf{z}_k^\top] = \mathbf{G}_c + \boldsymbol{\mu}_c \boldsymbol{\mu}_c^\top$ into covariance and mean outer product.

The dynamical discriminability δ^2 captures two distinct sources of discriminative information. The first one is the operator difference $\boldsymbol{\Delta} = \mathbf{A}_c - \mathbf{A}_h$ which encodes how differently the two dynamical regimes evolve the state. If $\mathbf{A}_c = \mathbf{A}_h$, then $\boldsymbol{\Delta} = \mathbf{0}$ and $\delta^2 = 0$, meaning the two systems are indistinguishable from their dynamics regardless of the embedding. The second aspect is the state distribution, characterized jointly by the covariance $\mathbf{G}_c$ and the mean $\boldsymbol{\mu}_c$, which determines which regions of the state space the trajectory typically visits. Even if $\boldsymbol{\Delta}$ is large in some directions, those directions only contribute to δ^2 if the stationary distribution places significant probability mass along them. However, if $\mathbf{G}_c$ concentrates the state along directions where $\boldsymbol{\Delta}$ acts strongly, or if the stationary mean $\boldsymbol{\mu}_c$ has a large component in a direction where $\boldsymbol{\Delta}$ acts, then δ^2 will be large.

A natural question arises: how reliably can a classifier distinguish whether $\mathbf{y}_k$ is drawn under $\mathcal{H}_c$ or $\mathcal{H}_h$? In the next section, we prove theoretical bounds on this DS discrimination task.

5 Sub-optimality of disregarding dynamics

Here we establish that the marginal stationary distributions of the observables $\mathbf{y}_k$ under the two hypotheses can have bounded divergence even when the dynamics are very different. This provides a theoretical classification error floor when the dynamic behind the generation of $\mathbf{y}_k$ is disregarded.

Result 1 (marginal density analysis limitation)**.** *Under Assumptions 1 and 2, let $p_c = \mathcal{N}(\mathbf{P}\boldsymbol{\mu}_c, \boldsymbol{\Gamma}_c)$ and $p_h = \mathcal{N}(\mathbf{P}\boldsymbol{\mu}_h, \boldsymbol{\Gamma}_h)$ be the stationary distributions of the observables $\mathbf{y}_k$ under $\mathcal{H}_c$ and $\mathcal{H}_h$, respectively, where $\boldsymbol{\Gamma}_c = \mathbf{P}\mathbf{G}_c\mathbf{P}^\top$ and $\boldsymbol{\Gamma}_h = \mathbf{P}\mathbf{G}_h\mathbf{P}^\top$ are the observable-space stationary covariances. Defining $\|\cdot\|$ as the Mahalanobis distance, the total variation (TV) distance $d_{\mathrm{TV}}(p_c, p_h)$ between p_c and p_h satisfies:*

$$d_{\mathrm{TV}}(p_c, p_h) \leq \frac{1}{2} \left\| \boldsymbol{\Gamma}_c^{-1/2} (\boldsymbol{\Gamma}_c - \boldsymbol{\Gamma}_h) \boldsymbol{\Gamma}_c^{-1/2} \right\|_F + \frac{1}{2} \left\| \mathbf{P}(\boldsymbol{\mu}_c - \boldsymbol{\mu}_h) \right\|_{\boldsymbol{\Gamma}_h^{-1}}, \tag{16}$$

and the Bayes error of any classifier using only the marginal distribution satisfies:

$$P_{\mathrm{error}} \geq \frac{1}{2} \left(1 - d_{\mathrm{TV}}(p_c, p_h)\right). \tag{17}$$

Moreover, there exist pairs $(\mathbf{A}_c, \mathbf{A}_h)$ and noise parameters $(\mathbf{m}_c, \boldsymbol{\Sigma}_c, \mathbf{m}_h, \boldsymbol{\Sigma}_h)$ with $\|\boldsymbol{\Delta}\|_F$ bounded away from zero such that both $\|\boldsymbol{\Gamma}_c - \boldsymbol{\Gamma}_h\|_F$ and $\|\mathbf{P}(\boldsymbol{\mu}_c - \boldsymbol{\mu}_h)\|$ are arbitrarily small.

Proof. See Appendix A. □

This result justifies the situation encountered in practice where the SVD mode distributions (marginals) of factual and hallucinated embeddings are nearly indistinguishable, but the transition dynamics differ. This was observed empirically in (Wilson & Akrout, 2026) where the first two SVD modes of the token embeddings showed distributional differences that are insufficient for hallucination detection. Note that when the noise means differ ($\mathbf{m}_c \neq \mathbf{m}_h$), the second term in (16) can be nonzero even for scaled rotations with equal noise covariance, providing limited discriminability through the mean shift alone.

Rather than classifying based on marginal distributions alone, we ask whether incorporating the prediction error as described in Section 3.1 leads to a significant improvement in classification performance, which is the focus of the next section.

6 Exponential Separability via Dynamical System Prediction

In this section, we prove that the dynamical prediction approach achieves exponentially decaying classification error.

Result 2 (Dynamical separability)**.** *Under Assumptions 1 and 2, suppose $\delta^2 + \beta_c > 0$ where δ^2 and β_c are defined in (12) and (13). Let $\eta \in \mathbb{R}$ be an arbitrary classification threshold and define the mean discriminability gap under $\mathcal{H}_c$ as*

$$\mu_L \equiv \mathbb{E}[\Delta\mathcal{E} \mid \mathcal{H}_c] > 0. \tag{18}$$

Then, for any threshold $\eta < \mu_L$, the misclassification probability satisfies:

$$\mathbb{P}(\Delta\mathcal{E} < \eta \mid \mathcal{H}_c) \leq 2\exp\left(-\frac{(L-1)(\mu_L-\eta)^2}{C_1\left(\sigma^2_{\max,c}(\delta^2+|\beta_c|)+\sigma^2_{\max,c}R^2\lambda_{\max}(\mathbf{G}_c)+P_mR^2\lambda_{\max}(\mathbf{G}_c)\right)}\right), \tag{19}$$

where $\sigma^2_{\max,c} = \lambda_{\max}(\mathbf{P}\boldsymbol{\Sigma}_c\mathbf{P}^\top)$, $R = \|\mathbf{P}\boldsymbol{\Delta}\|$, $P_m = \|\mathbf{P}\mathbf{m}_c\|^2$ and the constant $C_1 > 0$. An similar bound holds under $\mathcal{H}_h$ with $\boldsymbol{\mu}_c, \mathbf{G}_c, \boldsymbol{\Sigma}_c, \mathbf{m}_c$ replaced by $\boldsymbol{\mu}_h, \mathbf{G}_h, \boldsymbol{\Sigma}_h, \mathbf{m}_h$, and with the event $\{\Delta\mathcal{E} \geq \eta\}$.

Proof. See Appendix B. □

Result 2 establishes that the misclassification probability decays exponentially in the sequence length L, but the bound involves several variables (i.e., δ^2, β_c, $\sigma^2_{\max,c}$, R, $\lambda_{\max}(\mathbf{G}_c)$) whose interplay may obscure the essential scaling behavior. To present the result into a single interpretable quantity, we introduce the *separability factor* $S(L)$, defined as the standardized mean of the test statistic $\Delta\mathcal{E}$ under $\mathcal{H}_c$. This quantity, borrowed from signal detection theory, measures how many standard deviations the expected score is separated from zero: when $S(L)$ is large, the distribution of $\Delta\mathcal{E}$ under $\mathcal{H}_c$ is concentrated far from the decision boundary, making misclassification unlikely. The following corollary shows that $S(L)$ grows as $\sqrt{L}$, providing a direct link between sequence length and classification reliability, and yields an explicit formula for the minimum number of tokens needed to achieve a target error rate.

Corollary 1 (Convergence rate and minimum sequence length)**.** *Under the conditions of Result 2, the separability factor $S(L)$ of $\Delta\mathcal{E}$ under $\mathcal{H}_c$ grows as $\sqrt{L}$:*

$$S(L) \equiv \frac{\mathbb{E}[\Delta\mathcal{E} \mid \mathcal{H}_c]}{\sqrt{\mathrm{Var}(\Delta\mathcal{E} \mid \mathcal{H}_c)}} = \Theta\left(\sqrt{L}\cdot\frac{(\delta^2+\beta_c)}{\sigma^2_{\max,c}}\right). \tag{20}$$

Consequently, for any threshold η satisfying $|\eta| \leq \mu_L/2$, the number of tokens needed for classification with error probability $\leq \epsilon$ scales as:

$$L^* = \mathcal{O}\left(\frac{\sigma^4_{\max,c}}{(\delta^2+\beta_c)^2}\log\frac{1}{\epsilon}\right). \tag{21}$$

Proof. See Appendix C. □

This result explains the empirical observation that larger embedding models (e.g., Llama-3) achieve near-perfect classification even with short sequences, while smaller models (e.g., Jina-v5 with 30M parameters) require longer sequences.

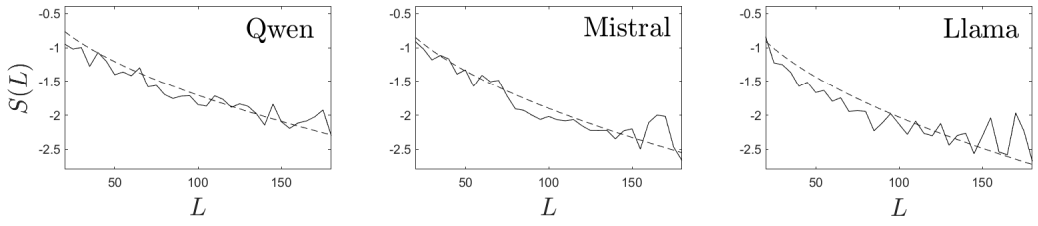


Figure 4: Estimates of the separability factor (solid line) from equation 20 for three different text embedding models for the HaluEval Li et al. (2023) dataset. Dashed lines of the form $K\sqrt{L}$ are fit to the data, where K is a constant used for fitting.

Extended DMD is applied to the HaluEval Li et al. (2023) dataset to obtain predictive models of the form equation 7 and equation 8 using token embeddings from `Qwen`, `Mistral` and `Llama-3`. On the remaining 8000 hallucinated responses not used for model inference, the differential residual score $\Delta\mathcal{E}$ from equation 6 is computed for 8000 hallucinated responses. $S(L)$ is computed for responses of different length, L, and plotted in Fig. 4. As predicted by Corollary 1, the separability factor is proportional to $\sqrt{L}$.

7 CROSS-EMBEDDING GENERALIZATION

An important practical question is whether the dynamical models fitted using one embedding model can transfer to another. It has been shown in (Wilson & Akrout (2026)) that fitting Koopman operators $\mathbf{A}_c^{(\boldsymbol{\theta}_1)}$ and $\mathbf{A}_h^{(\boldsymbol{\theta}_1)}$ on token embeddings from one model parametrized by $\boldsymbol{\theta}_1$ then run prediction on token embeddings from a different model parametrized by $\boldsymbol{\theta}_2$ achieve above-chance accuracy despite operating in an entirely different observation space. We now provide a theoretical explanation for this phenomenon.

Consider two embedding models $G_2^{(\boldsymbol{\theta}_1)} : \mathcal{Q} \to \mathbb{R}^{M_1}$ and $G_2^{(\boldsymbol{\theta}_2)} : \mathcal{Q} \to \mathbb{R}^{M_2}$, where $\mathcal{Q}$ denotes the token space and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ parametrize the respective model architectures and weights. Each embedding model induces its own observation space and, consequently, its own Koopman operator approximations. We write:

$$\mathbf{y}_k^{(\boldsymbol{\theta}_i)} = G_2^{(\boldsymbol{\theta}_i)}(q_k) \in \mathbb{R}^{M_i}, \qquad i \in \{1, 2\}, \tag{22}$$

for the token embeddings under each model, and denote the corresponding fitted Koopman approximations by $\mathbf{A}_c^{(\boldsymbol{\theta}_i)}, \mathbf{A}_h^{(\boldsymbol{\theta}_i)}$ with operator differences $\boldsymbol{\Delta}^{(\boldsymbol{\theta}_i)} = \mathbf{A}_c^{(\boldsymbol{\theta}_i)} - \mathbf{A}_h^{(\boldsymbol{\theta}_i)}$.

The central question is: under what conditions on the pair $(G_2^{(\boldsymbol{\theta}_1)}, G_2^{(\boldsymbol{\theta}_2)})$ can dynamical models fitted in the $\boldsymbol{\theta}_1$-observation space provide discriminative information about trajectories observed in the $\boldsymbol{\theta}_2$-observation space?

Assumption 3 (Approximate intertwining)**.** *There exists a linear map $\mathbf{T} \in \mathbb{R}^{M_2 \times M_1}$ such that the two embedding models are approximately related by:*

$$G_2^{(\boldsymbol{\theta}_2)}(q) = \mathbf{T}\, G_2^{(\boldsymbol{\theta}_1)}(q) + \boldsymbol{r}(q), \qquad \forall\, q \in \mathcal{Q}, \tag{23}$$

where $\boldsymbol{r} : \mathcal{Q} \to \mathbb{R}^{M_2}$ is a residual satisfying $\mathbb{E}[\|\boldsymbol{r}(q_k)\|^2] \le \chi^2$ along the token sequences.

The map $\mathbf{T}$ acts as a *translation* between the two observation spaces: it captures the extent to which the geometric structure of one embedding is a linear transformation of the other. The residual $\boldsymbol{r}$ measures the *intertwining defect*, i.e., the degree to which this linear relationship breaks down. When $\chi^2 = 0$, the two embeddings are exactly linearly related, and any dynamical structure in one space is perfectly preserved in the other. When $\chi^2 > 0$, the transfer is approximate.

Definition 2 (Cross-embedding discriminability)**.** *The* cross-embedding discriminability *when fitting on $G_2^{(\boldsymbol{\theta}_1)}$ and testing on $G_2^{(\boldsymbol{\theta}_2)}$ is:*

$$\delta^2_{\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2} \equiv \mathbb{E}\left[\left\| \mathbf{P}^{(\boldsymbol{\theta}_2)} \boldsymbol{\Delta}^{(\boldsymbol{\theta}_1)} \mathbf{z}_k^{(\boldsymbol{\theta}_2)} \right\|^2 \right], \tag{24}$$

where $\mathbf{z}_k^{(\boldsymbol{\theta}_2)}$ is the lifted observable from the $\boldsymbol{\theta}_2$-embedding and $\mathbf{P}^{(\boldsymbol{\theta}_2)}$ is the corresponding projection onto the observable components.

Result 3 (Cross-embedding transfer bound)**.** *Under Assumptions 1, 2, and 3, the cross-embedding discriminability satisfies:*

$$\delta^2_{\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2} \ge \sigma^2_{\min}(\mathbf{T})\, \delta^2(\boldsymbol{\theta}_1) - C_2 \left(\chi^2 \left\| \boldsymbol{\Delta}^{(\boldsymbol{\theta}_1)} \right\|^2 \lambda_{\max}\big(\mathbf{G}_c^{(\boldsymbol{\theta}_2)}\big) + \chi^2 \delta^2(\boldsymbol{\theta}_1) \right), \tag{25}$$

where $\sigma_{\min}(\mathbf{T})$ is the smallest singular value of $\mathbf{T}$, $C_2 > 0$ is a constant, and $\delta^2(\boldsymbol{\theta}_1) = \operatorname{tr}(\mathbf{P}^{(\boldsymbol{\theta}_1)} \boldsymbol{\Delta}^{(\boldsymbol{\theta}_1)} \mathbf{G}_c^{(\boldsymbol{\theta}_1)} (\boldsymbol{\Delta}^{(\boldsymbol{\theta}_1)})^\top (\mathbf{P}^{(\boldsymbol{\theta}_1)})^\top) + \left\| \mathbf{P}^{(\boldsymbol{\theta}_1)} \boldsymbol{\Delta}^{(\boldsymbol{\theta}_1)} \boldsymbol{\mu}_c^{(\boldsymbol{\theta}_1)} \right\|^2$ is the discriminability under $\boldsymbol{\theta}_1$.

Consequently, the classification accuracy when fitting on $G_2^{(\boldsymbol{\theta}_1)}$ and testing on $G_2^{(\boldsymbol{\theta}_2)}$ satisfies:

$$\begin{aligned} \text{Accuracy}_{\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2} &= \frac{1}{2}\mathbb{P}(\Delta\mathcal{E} \ge 0 \mid \mathcal{H}_c) + \frac{1}{2}\mathbb{P}(\Delta\mathcal{E} < 0 \mid \mathcal{H}_h) \\ &\ge \frac{1}{2} + \frac{1}{2}\left(1 - 2\exp\left(- \frac{(L-1)\left(\delta^2_{\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2}\right)^2}{C_1\left(\sigma^2_{\max,c}\, \delta^2_{\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2} + \sigma^2_{\max,c}\, R^2\, \lambda_{\max}(\mathbf{G}_c^{(\boldsymbol{\theta}_2)})\right)} \right) \right). \end{aligned} \tag{26}$$

Proof. See Appendix D. □

For empirical validation of the bound in (25), we select a model pair (`Qwen3`, `F2LLM`) whose cross-transfer behavior in (Wilson & Akrout, 2026) was intriguing: fitting the DS on `Qwen3` embedding and testing on `F2LLM` embeddings led to a cross-model accuracy of 63.0%, while doing the opposite yielded an accuracy of 50.5%. We then evaluate the intertwining map $\mathbf{T}$ and defect χ^2 between the two embedding spaces. As depicted in Fig. 5, while both mapping directions exhibit well-conditioned singular value spectra for $\mathbf{T}$ across the subspace modes, their reconstruction residual errors differ drastically: the intertwining defect for mapping `F2LLM` $\rightarrow$ `Qwen3` ($\chi^2 \approx 8.8 \times 10^6$) is nearly double that of `Qwen3` $\rightarrow$ `F2LLM` ($\chi^2 \approx 4.8 \times 10^6$). This aligns directly with our theoretical bound in Result 3: when transferring from a lower-density embedding space (e.g., `F2LLM`) to a denser and compact space (e.g., `Qwen3`), the large intertwining defect χ^2 acts as a heavy penalty term that suppresses the transferable discriminability $\delta^2_{\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2}$ toward zero, causing accuracy to drop to 50.5%. However, Transferring from `Qwen3` to `F2LLM` incurs a substantially smaller defect, preserving sufficient dynamical separation to maintain an accuracy of 63.0% without refitting.

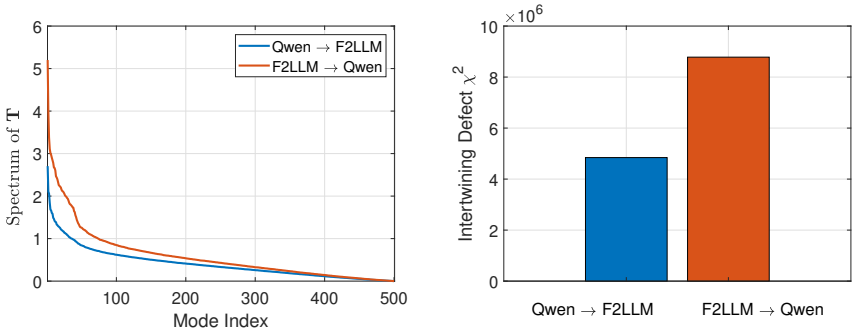


Figure 5: Empirical validation of cross-embedding transfer between `Qwen3` and `F2LLM`. *Left:* Singular values of the linear intertwining map $\mathbf{T}$ across SVD modes. *Right:* The reconstruction residual error, i.e., intertwining defect χ^2 between the two embedding models.

8 Conclusion

We have developed a mathematical framework for understanding the distinguishability guarantees of two dynamical systems to classify LLM responses in the embedding space. First, we establish that classifiers relying solely on the marginal distribution of token embeddings face an irreducible accuracy floor since two dynamical systems can produce nearly identical stationary distributions while having completely different dynamical evolution. Second, we showed that exploiting temporal dynamics through prediction residuals overcomes this limitation, with classification error decaying exponentially in sequence length at a rate governed by the dynamical distinguishability of the two systems. This provides practical guidance for choosing the minimum response length needed for reliable detection. Third, we addressed the surprising phenomenon of cross-embedding transfer, demonstrating that dynamical models fitted on one embedding can transfer to another when the embedding spaces share similar geometric structure. The transfer quality depends on affine geometric compatibility, the fitted discriminability space, and the extent of distortion between the fitted and target embedding spaces. This explains why some embedding pairs transfer well while others do not. Together, these results establish guarantees on the distinguishability between dynamical systems theory for LLM token generations, shifting the focus from static representations of individual tokens to the dynamical patterns that emerge across the embedding trajectory of a LLM response.

Limitations and future directions: Our linear Koopman approximation could be extended to nonlinear models (e.g., kernel EDMD or neural network-based Koopman methods) to capture richer dynamical structure. The Gaussian noise assumption which is realistic after testing it on real-world datasets could be relaxed to heavy tail distributions to broaden the applicability of the obtained bounds. Future work can also devise a joint optimization over the embedding projection and the Koopman operators to seeking the projection that maximizes δ^2 for systematic embedding selection. Such an algorithm will connect to the classical distinction between variance-maximizing and discriminability-maximizing projections.

REFERENCES

Marko Budišić, Ryan Mohr, and Igor Mezić. Applied koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4), 2012.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

James D Hamilton. *Time series analysis*, volume 2. Princeton Princeton University Press, 1994.

Thomas Kailath. *Linear systems*, volume 156. Prentice-Hall Englewood Cliffs, NJ, 1980.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2nd edition, 1999.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Igor Mezić. Analysis of fluid flows via spectral properties of the koopman operator. *Annual review of fluid mechanics*, 45(1):357–378, 2013.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pp. 5610–5618. PMLR, 2019.

Mohammad Amin Sheikhi, Gabriel de Albuquerque Gleizer, Peyman Mohajerin Esfahani, and Tamás Keviczky. Data-driven fault isolation in linear time-invariant systems: a subspace classification approach. *IEEE Control Systems Letters*, 2025.

Mohammad Amin Sheikhi, Gabriel de Albuquerque Gleizer, Tamás Keviczky, and Peyman Mohajerin Esfahani. Fault diagnosis in dynamical systems: Geometric interpretation and tractable algorithms. *Annual Review of Control, Robotics, and Autonomous Systems*, 9, 2026.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pp. 439–473. PMLR, 2018.

M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

Dan Wilson and Mohamed Akrout. Low-cost black-box detection of llm hallucinations via dynamical system prediction. *arXiv preprint arXiv:2605.05134*, 2026.

Lei Xin, George T-C Chiu, and Shreyas Sundaram. Online change points detection for linear dynamical systems with finite sample guarantees. *Automatica*, 171:111894, 2025.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. SAC3: reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15445–15458, 2023.