Temporal Instability Phases Precede Reasoning Failure in Generative Models

Venkata Pendyala

***Abstract*— Current AI systems are bent on hallucination and error detection, with prediction of failure lagging far behind. This paper introduces the concept of instability as a temporal phase in large language model behavior: a measurable regime that precedes overt failure events such as noncompliance and logical breakdowns. Rather than detecting errors at the moment they occur, instability is operationalized through a structured five-dimensional diagnostic framework probing risk awareness, factual grounding, adversarial robustness, stakeholder sensitivity, and revision readiness. Sequential analysis of model responses shows that instability exhibits statistically significant temporal persistence and clustering, distinguishing it from random error noise. Permutation testing confirms that observed state persistence exceeds shuffled baselines, supporting the claim that instability constitutes a genuine behavioral phase. Entry into this phase is associated with up to 58% elevated near-term probability of downstream failure, providing measurable lead time before visible breakdown. This reframes reliability monitoring from reactive detection to proactive phase identification and suggests that monitoring latent reasoning health can enable early-warning systems for generative models.**

***Index Terms*—Failure Forecasting, Hallucination Prediction, Large Language Models, Latent Instability, LLM Behavior, Temporal Phase.**

I. Introduction

The large language model is one of the best known and most used types of artificial intelligence, yet it is plagued by the problem of reasoning failures. Advances in quality and capability have continuously been made, but these models tend to generate outputs that are logically inconsistent, violate explicit task instructions, or fail to adhere strictly to given constraints. Undermining reliability, especially in high-risk or high-stress contexts such as spacecraft systems, legal situations, or medical assistance, this problem wastes substantial resources and time.

These reasoning failures are more than temporary or isolated glitches: they reflect major limitations in how current architectures learn and generalize from data. Large language models optimize next-token prediction over massive text corpora, a training objective that does not explicitly enforce logical coherence or instruction compliance. As a result, models may generate responses that plausibly follow surface patterns in language but fail deeper reasoning tests, violate user intent, or misinterpret constraints embedded in the prompt. Work in rigorous studies of LLMs has proven via a variant of Cantor's diagonalization argument that it is theoretically impossible in a formal world to completely rid generative models of hallucinations or other logical errors, as we can always construct some prompt for which the model will output a nonsensical answer. Therefore, the inherent nature of failures and their economic, developmental, and research problems prompt the creation of early warning systems, or frameworks that forecast reasoning issues rather than detecting them. In this way, a potentially significant lead time allows for further analysis and prevention.

Significant research has demonstrated that reasoning quality deteriorates in specific structural contexts, such as longer interactions or tasks requiring compositional reasoning across multiple steps. In multi-turn settings, error patterns often shift from isolated mistakes to sequences of degraded reasoning, suggesting the existence of phases in which the model's internal representations become unstable relative to earlier behavior. This is consistent with empirical observations that models can transition from coherent, instruction-aligned behavior toward sequences of noncompliant, incoherent, or logically invalid outputs as interaction history grows and error accumulation compounds.

Characterizing these phases and understanding their relationship to different types of failures is critical for building dependable AI systems. If elevated probabilities of logical inconsistency or instruction violation are preceded by measurable changes in the model's diagnostic profile, then it may be possible to forecast reasoning failure risk before it manifests in a harmful or incorrect output. Such predictive insight has implications for safe deployment, real-time monitoring, and adaptive interventions during model interaction: unlike prior work that detects failures after they occur, we aim to forecast them. This paper proposes a five-"rung" diagnostic method to identify unstable phases of LLM behavior and provides statistical evidence supporting both the existence of these phases and their implications for failure prediction.

II. Problem Formulation

We consider a generative model producing a sequence of responses $y_1, y_2, y_3, \ldots, y_T$ over discrete interaction turns (prompt-response pairs).

For each response y_t, we evaluate a set of *five* binary reasoning criteria:

$$R_t = (R_t^1, R_t^2, \ldots, R_t^5)$$

Where R_t^k can take on the values of *1* or *0,* respectively indicating a satisfied or failed criterion. The number of failed diagnostics at turn *t* are:

$$f_t = \sum_{k=1}^{5} 1 - R_t^{(k)}$$

A *failure signature* of a response y_t indicates whether there are failed diagnostics present:

$$e_t = \mathbf{1}[f_t > 0]$$

At each turn *t,* we compute a scalar instability score S_t from recent patterns of rung failures. *Instability* itself, therefore, is the condition of being in an elevated phase of this score, denoted by the discrete phase label Z_t.

To accurately consider a pattern or process rather than short-term trend, we calculate the instability score only if the number of turns *N* is greater than 20. We also consider two windows for the score calculation: $w_s = 5$, representing the shorter window of turns, and $w_l = 20$, representing the larger window of turns. Given that instability is a temporally defined degradation phase in which recent multi-criterion rung failures become denser, more bursty, or more skewed towards at least one worsening criterion, we may quantify an instability score S_t through a *mean shift*, a *burstiness score*, and a *max per-rung degradation.*

By the mean shift, we compute:

$$\Delta_t = \mu_s - \mu_l$$

Where:

$$\mu_s = \frac{1}{w_s} \sum_{i=t-w_s+1}^{t} f_i$$

$$\mu_l = \frac{1}{w_l} \sum_{i=t-w_l+1}^{t} f_i$$

For a burstiness score, we may compute:

$$B_t = \frac{r_{max}}{w_s}$$

where r_{max} denotes the length of the longest consecutive run of turns for which $e_i = 1$. The maximum per-rung degradation is computed by:

$$M_t = \max_{k\epsilon\{1,\dots,5\}}^{(p_{s,k} - p_{l,k})}$$

Where $p_{s,k}$ and $p_{l,k}$ are defined as:

$$p_{s,k} = \frac{1}{w_s} \sum_{i=t-w_s+1}^{t} R_i^k$$

The instability score is simply the sum of all these components: $S_t = \Delta_t + B_t + M_t$. The discrete phase variable $Z_t\epsilon\{\text{stable}, \text{warning}, \text{unstable}\}$ uses fixed thresholds to assign a label from the S_t. The aim of this paper is to study whether reasoning failures in such generative model systems are preceded by measurable reasoning degradation patterns (instability phases). Formally, we test whether elevated values of S_t significantly increase the conditional probability of downstream failure events within horizon *H*.

III. Evaluation Methodology

Equipped with definitions, we must determine *two* things to achieve the aims. Firstly, it must be proven that instability as a phase designated by Z_t exhibits *temporal structure* beyond independent fluctuations. Next, if Z_t is truly a temporal regime, we must analyze whether it matters in reliably predicting reasoning and logical failures (or preceding them).

A. Phase Persistence Analysis

If instability was truly noise rather than signal, phase labels and instability scores would lack persistence beyond random chance levels. Thus, we evaluate dependence in both a discrete phase sequence and the continuous instability score.

Let $Z_t\epsilon\{\text{stable}, \text{warning}, \text{unstable}\}$ denote the discrete instability phase at turn *t.* If phase assignments are temporally independent, the ordering of labels should not influence statistical properties such as adjacent agreement (whether two neighboring steps have the same state, or $\mathbf{1}(Z_t = Z_{t+1})$ or run structure (how long consecutive blocks of the same label occur, and how often they switch). If instability were random noise, runs would tend to be short and the system would frequently switch. The probability that two adjacent labels match would be whatever chance produces based on the overall proportions of unstable, warning, and stable.

Mathematically, we compute two statistics regarding runs and adjacent agreement: *lag-1 persistence* and *average run length.* Empirical lag-1 persistence is:

$$P_{lag1} = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{1}(Z_t = Z_{t+1}).$$

It measures the proportion of consecutive turns for which the phase remains unchanged. Under H_0 of temporal independence, adjacent matches occur at rates determined by the proportions of labels.

Average run length, however, is defined as a maximal sequence of consecutive identical phase labels. r_i denotes the length of the *i*-th run, and *K* denotes the total number of runs. Therefore, the average run length is simply:

$$ARL = \frac{1}{K} \sum_{i=1}^{K} r_i.$$

Short run lengths would be expected under temporal independence, while extended runs would indicate regime persistence.

Given these two main tools, we construct a permutation null model that, given the observed sequence Z_1, Z_2, Z_3 and so on, generates *B* random permutations of the phase labels. Each new permutation recomputes lag-1 persistence and average run length, preserving the total number of occurrences of each phase and the marginal phase distribution, while eliminating any potential temporal structure. Empirical *p-values* are computed as the proportion of permutations whose persistence statistics equal or exceed those observed initially. Significant deviation from the permutation distribution is strong evidence that the phase sequence exhibits temporal dependence

inconsistent with independent noise (scrambling loses the signal).

As discrete phase labels are obtained by threshold calculations with the instability score S_t, we test persistence at the level of the continuous score to ensure that the potential phase behavior is not induced suddenly by the discretization. Lag-ℓ autocorrelation is defined as:

$$\rho(\ell) = \frac{\sum_{t=1}^{T-\ell}(S_t-\bar{S})(S_{t+\ell}-\bar{S})}{\sum_{t=1}^{T}(S_t-\bar{S})^2},$$

We focus again on lag-1 autocorrelation $\rho(1)$, which measures the extent to which elevated instability at time *t* predicts elevated instability at the next time (i.e. $t+1$). We also observe the higher order lags to analyze any potential persistence beyond one time step. Under the null hypothesis of temporal independence, we again notice that the expected autocorrelation (and likely higher order correlation) is precisely 0.

To assess statistical significance, we generate *B* random permutations of the instability score sequence S_t in a similar fashion to the permutation test for the discrete phenomena. Permutation is applied to S_t, rather than Z_t.This preserves the distribution of instability magnitudes and the variance of the score, but destroys any type of temporal ordering, suggesting that observed autocorrelation values significantly exceeding the permutation distribution would indicate memory structures rather than random fluctuations.

B. Phase Consequence: Lead Time Failure Risk

If it is established that instability forms a temporally coherent regime or phase, we must next test whether it has any predictive consequences on failures. We let

$$H_t \in \{0,1\}$$

denote whether the response at time t contains hallucinated content. To evaluate predictive consequences of instability, we define a strictly prospective failure event over a fixed horizon $H \in \mathbb{N}$:

$$Y_{t,H} = \mathbf{1}(\exists\, i \in \{t+1, \dots, t+H\} \text{ such that } H_i = 1).$$

Equivalently,

$$Y_{t,H} = \max_{1\le k\le H} H_{t+k}.$$

Thus, $Y_{t,H} = 1$if at least one logical or reasoning failure occurs within the next *H* turns. This metric excludes the current turn to ensure a future focused view. Essentially, the aim is to identify whether the instability phase at time t increases the probability of failure occurring within the next *H* responses.

We define the conditional probabilities:

$$p_{\text{elev}}(H) = P\big(Y_{t,H} = 1 \mid Z_t \in \{\text{Warning,Unstable}\}\big),$$
$$p_{\text{stable}}(H) = P\big(Y_{t,H} = 1 \mid Z_t = \text{Stable}\big).$$

The relative risk, abbreviated here on out as RR, (ratio between occurring a failure in horizon given unstable or warning versus stable) is

$$RR(H) = \frac{P\big(Y_{t,H} = 1 \mid Z_t \in \{\text{Warning,Unstable}\}\big)}{P\big(Y_{t,H} = 1 \mid Z_t = \text{Stable}\big)}.$$

A value $RR(H) > 1$ indicates multiplicative increase of in-window failure probability under elevated instability, either unstable or warning. Importantly, this framework evaluates conditional probability rather than determinism: instability is assessed as a probabilistic leading indicator, not a guaranteed precursor.

For each horizon H, we construct a 2×2 contingency table:

	$Y_{t,H} = 1$	$Y_{t,H} = 0$
$Z_t \in$ {Warning,Unstable}	a	b
$Z_t =$ Stable	c	d

with empirical risk estimates

$$\hat{p}_{\text{elev}}(H) = \frac{a}{a+b}, \qquad \hat{p}_{\text{stable}}(H) = \frac{c}{c+d}.$$

We test the one-tailed hypothesis

H_0: $P(Y_{t,H} = 1 \mid Z_t \in \{\text{Warning,Unstable}\}) = P(Y_{t,H} = 1 \mid Z_t = \text{Stable})$

H_A: $P(Y_{t,H} = 1 \mid Z_t \in \{\text{Warning,Unstable}\}) > P(Y_{t,H} = 1 \mid Z_t = \text{Stable})$

using Fisher's exact test.

IV. Experimental Results and Discussion

After cleaning malformed entries, the dataset consisted of 1,994 usable turns drawn from 237 conversations, with 588 total failure events and 158 hallucinations. Instability scores were computed using combinations of short and long windows. The short window was $w_+ \in \{3,5,7,10\}$ and the long window $w_- \in \{10,15,20\}$, with $w_- > w_+$.

Because instability requires at least w_- prior turns, larger long windows reduce the number of eligible observations. For each configuration, we evaluate:

1. Temporal persistence via lag-1 autocorrelation and permutation testing.
2. Predictive consequence via relative risk (RR) of failure within the next five turns.
3. Hallucination-specific risk as a secondary outcome.

A. Temporal Independence

Across all window configurations, the instability score S_t exhibits strong lag-1 autocorrelation, ranging from *0.60 to 0.85*, depending on window size. Under 300 random permutations of the score sequence (destroying temporal order while preserving marginal distribution), the mean permuted autocorrelation was approximately zero, with empirical

permutation p-values reaching extremely low values (≈ 0.0033) in every configuration.

Given:

1. $w_+ = 5, w_- = 15$: lag-1 correlation = **0.746**, permutation p ≈ **0.0033**
2. $w_+ = 7, w_- = 20$: lag-1 correlation = **0.795**, permutation p ≈ **0.0033**
3. $w_+ = 10, w_- = 20$: lag-1 correlation = **0.847**, permutation p ≈ **0.0033**

These results strongly reject the null hypothesis of temporal independence. Instability does not behave as random fluctuation; rather, elevated instability tends to persist across consecutive turns. This supports the interpretation of instability as a temporally coherent regime rather than a transient anomaly.

B. Implications of Instability on Future Failure

We next evaluate whether elevated instability increases the probability of failure within a five-turn horizon:

$$Y_{t,5} = 1(\exists i \in \{t+1, \ldots, t+5\} \text{ such that failure}_i = 1).$$

High instability is defined as the top 25% of S_t values, and low instability as the bottom 25%, excluding the middle 50% for contrast. For any failure across the next five turns, nearly all window settings show elevated instability significantly increases near-term failure risk. The results are summarized:

(w_+)	(w_-)	Eligible turns	Convos	RR@5	95% Katz CI
3	10	915	57	1.115	(0.959, 1.297)
5	10	915	57	1.165	(1.009, 1.346)
7	10	915	57	1.106	(0.948, 1.291)
3	15	687	36	1.263	(1.056, 1.510)
5	15	687	36	1.504	(1.258, 1.798)
7	15	687	36	1.399	(1.156, 1.694)
10	15	687	36	1.298	(1.099, 1.533)
3	20	542	22	1.344	(1.121, 1.612)
5	20	542	22	1.431	(1.196, 1.711)
7	20	542	22	1.579	(1.297, 1.924)
10	20	542	22	1.418	(1.195, 1.683)

For instance, at $w_+ = 7, w_- = 20$, the relative risk of any failure within five turns is approximately 1.58**,** indicating a *58% increase in short-term failure probability* under elevated instability. The associated p-value (≈ 1×10^{-6}) strongly rejects the null hypothesis of equal risk. We consistently note that even the lower end of the confidence intervals is above 1 in nearly all cases, suggesting anywhere from slight to moderate-high influence of instability on reasoning.

These findings demonstrate that instability functions as a statistically significant probabilistic leading indicator of near-term reasoning failure. In contrast, instability did not significantly predict hallucinations specifically in this dataset. Relative risk values for hallucination within five turns were consistently near or below 1 and statistically non-significant across all window configurations. This suggests that, within this corpus, instability more strongly forecasts general reasoning failures (e.g., logical inconsistency, non-compliance, structural breakdown) rather than hallucination events as labeled here.

References

[1] K. Jazwinska and A. Chandrasekar, "AI Search Has A Citation Problem," *Columbia Journalism Review*, Mar. 06, 2025. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php

[2] S. Hirsch, "The Cost of Hallucinations: A Calculator for Global Business," *Home Business Magazine*, Feb. 14, 2026. https://homebusinessmag.com/businesses/ai/cost-hallucinations-calculator-global-business/ (accessed Feb. 18, 2026).

[3] Y. Sun, D. Sheng, Z. Zhou, and Y. Wu, "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content," *Humanities and Social Sciences Communications*, vol. 11, no. 1, Sep. 2024, doi: https://doi.org/10.1057/s41599-024-03811-x.

[4] M. Chen, W. Yu, and K. Liu, "A meta-analysis of third-person perception related to distorted information: Synthesizing the effect, antecedents, and consequences," *Information Processing and Management*, vol. 60, no. 5, pp. 103425–103425, Sep. 2023, doi: https://doi.org/10.1016/j.ipm.2023.103425.

[5] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is Inevitable: An Innate Limitation of Large Language Models." Available: https://arxiv.org/pdf/2401.11817

[6] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *arXiv (Cornell University)*, Nov. 2023, doi: https://doi.org/10.48550/arxiv.2311.05232.

[7] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,"

arXiv:2303.08896 [cs], Mar. 2023, Available: https://arxiv.org/abs/2303.08896

[8] J. Kossen, J. Han, M. Razzak, L. Schut, S. Malik, and Y. Gal, "Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs," *arXiv.org*, 2024. https://arxiv.org/abs/2406.15927

[9] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, Jun. 2024, doi: https://doi.org/10.1038/s41586-024-07421-0.

[10] S. Dhuliawala *et al.*, "Chain-of-Verification Reduces Hallucination in Large Language Models," *arXiv.org*, 2023. https://arxiv.org/abs/2309.11495

[11] S. Kang, Y. F. Bakman, Y. D. Nur, B. Buyukates, and S. Avestimehr, "Uncertainty Quantification for Hallucination Detection in Large Language Models: Foundations, Methodology, and Future Directions," *arXiv.org*, 2025. https://arxiv.org/abs/2510.12040 (accessed Feb. 22, 2026).

[12] B. Paudel, A. Lyzhov, P. Joshi, and P. Anand, "HalluciNot: Hallucination Detection Through Context and Common Knowledge Verification," *arXiv.org*, 2025. https://arxiv.org/abs/2504.07069

[13] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How Can We Know What Language Models Know?," *arXiv:1911.12543 [cs]*, May 2020, Available: https://arxiv.org/abs/1911.12543

[14] Y. Ovadia *et al.*, "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift," *arXiv.org*, Dec. 17, 2019. https://arxiv.org/abs/1906.02530

[15] J. Gawlikowski *et al.*, "A Survey of Uncertainty in Deep Neural Networks," *arXiv:2107.03342 [cs, stat]*, Jan. 2022, Available: https://arxiv.org/abs/2107.03342

[16] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," *arXiv:1610.02136 [cs]*, Oct. 2018, Available: https://arxiv.org/abs/1610.02136

[17] X. Du, C. Xiao, and Y. Li, "HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection," *arXiv.org*, 2024. https://arxiv.org/abs/2409.17504