Temporal Instability Phases Precede and Predict Reasoning Error in Generative Pre-Trained Transformers

Venkata S. Pendyala

***Abstract*— Large language model failures are often treated as isolated turn-level events. Here we show that this view is incomplete. Analyzing multi-turn GPT conversations, we identify persistent latent risk states, inferred from a frozen prospective instability signal, that are temporally non-random, persist across multiple consecutive turns, and forecast elevated future failure probability over subsequent horizons. Higher latent states are associated with systematically higher future failure risk, a monotone ordering that replicates across four disjoint held-out GPT datasets, with state-transition χ^2statistics ranging from 91.64 to 486.15 and high-versus-low risk ratios reaching 2.76. Within-conversation analyses further show that instability rises before failure events, arguing against a purely cross-sectional explanation. Because the latent states are inferred from observable behavioral features without access to ground-truth failure labels at inference time, the resulting signal is prospective and potentially usable during deployment. These findings suggest that reasoning failure in large language models is better understood not as isolated noise, but as entry into temporally persistent high-risk regimes. This reframes model unreliability as a dynamical systems problem and has direct implications for real-time monitoring, safety evaluation and training-time intervention.**

***Index Terms*— LLMs, GPT, reasoning errors, hallucinations, error prediction, safety systems, instability, reasoning**

I. INTRODUCTION

Generative pre-trained transformers have become central to how people access information, construct arguments and reason through consequential decisions. They are used in medicine, law, education and scientific research, and for many users represent the primary interface with modern artificial intelligence. Understanding the conditions under which their reasoning can be trusted has therefore become a problem with broad scientific and societal stakes.

Existing approaches address this problem by asking whether a particular response is wrong. Methods based on hallucination detection, factuality checking, semantic entropy and self-consistency analysis have each improved the identification of erroneous or uncertain outputs. However, these methods evaluate reliability primarily at the level of the individual response. They do not test whether errors are preceded by temporally persistent latent risk states in multi-turn conversation, nor whether such states can be inferred prospectively from observable behavioral signals and used to predict future failure over subsequent turns. Whether GPT reasoning fails abruptly or instead emerges from a detectable pre-failure phase therefore remains unresolved.

This distinction matters in conversation, where GPTs are most often deployed. Each response is conditioned on the full preceding interaction, and each answer shapes the reasoning context of those that follow. Reliability may therefore be temporally organized: a model may pass through stretches of stable reasoning and then enter a regime in which coherence weakens, justification thins and the probability of subsequent failure rises. If so, the scientifically relevant object is not only the erroneous response at the end of the process, but the latent behavioral state that precedes and enables it. Identifying such states prospectively, without access to ground-truth failure labels at inference time, would provide a principled foundation for real-time monitoring and intervention in deployed systems.

Here we test this possibility directly in multi-turn GPT conversations. We construct a prospective instability signal from a frozen seven-criterion reasoning-quality classifier trained on a disjoint dataset, infer discrete latent risk states using unsupervised methods fit exclusively on the training partition, and ask whether these states persist across turns and predict future failure over subsequent conversational horizons. We find that they do. Across a primary corpus of naturalistic GPT interactions and four independent held-out GPT evaluation sets, the inferred latent states exhibit strongly non-random temporal structure, higher-risk states are associated with monotonically higher future failure probability and within-conversation analyses confirm that instability rises before failure relative to matched reference windows from the same trajectory. These results suggest that reasoning failure in GPT-family models is often not abrupt but the endpoint of a temporally organized, prospectively detectable pre-failure phase — a finding with direct implications for how reliability is monitored and governed in deployed conversational AI systems.

II. RESULTS

A. Instability forecasts future failure

Before testing the latent-state hypothesis directly, we first asked whether the prospective instability signal itself contained measurable information about subsequent reasoning failure. The instability score $I_t$was computed from the frozen seven-criterion reasoning-quality classifier without access to downstream failure labels at inference time, and therefore

constituted a genuinely prospective signal rather than a retrospective summary.

On the primary GPT test split, this signal showed strong predictive performance at the pre-specified primary horizon $H = 3$, with similarly strong performance at nearby horizons.

Metric	Primary horizon $H = 3$	Peak observed
ROC–AUC	0.806	—
PR–AUC	0.912	0.921 ($H = 6,7$)
Average precision	0.907	—

These values were well above chance. Permutation-based negative controls further indicated that the signal was not an artifact of random label structure.

Horizon	Observed PR–AUC	Null mean PR–AUC	p-value
$H = 3$	0.912	0.518	0.006
$H = 6$	0.921	0.568	0.010

Thus, even before discretization into latent states, the instability signal already carried substantial prospective information about future reasoning failure. All subsequent latent-state analyses build on this forecasting result.

Latent states are temporally persistent

We next asked whether the inferred latent-state sequence $\{Z_t\}$ exhibited non-random temporal organization. Under the null hypothesis of independent state assignment, observed transition counts should not differ materially from those implied by the product of row and column marginals. This null was decisively rejected in every GPT dataset evaluated.

Dataset	χ^2 statistic	p-value	d.f.
Primary split	486.15	6.63×10^{-104}	4
Disjoint held-out subset	200.35	3.17×10^{-42}	4
Holdout A	91.64	5.89×10^{-19}	4
Holdout B	126.75	1.93×10^{-26}	4

The scale of these deviations argues against the interpretation that the latent states are arbitrary bins imposed on noisy turn-level behavior. Instead, the state process shows strong short-range dependence, consistent with temporally coherent instability phases in GPT conversations.

B. Instability phases span multiple turns

A temporally meaningful phase should persist for more than a single turn. We therefore quantified contiguous run lengths within each latent state. Under an independent three-state process, the expected mean run length is 1.5 turns. Observed run lengths exceeded this benchmark, particularly for the intermediate- and high-risk states.

Analysis / dataset	R_0 low	R_1 mid	R_2 high
Primary split (persistent-state)	2.27	3.59	3.76
Primary split (mixture-state)	1.14	4.06	2.72
Disjoint held-out subset	1.03	2.86	2.39

Across constructions and datasets, the intermediate- and high-risk states occupied multiple consecutive turns at rates incompatible with random assignment. GPT failure is therefore not merely preceded by isolated spikes in local instability, but by sustained entry into multi-turn risk regimes.

C. Higher-risk states imply higher future failure probability

The central empirical claim of this study is that future failure risk is ordered by latent state. For each horizon $h$and state z, we estimated

$$\mathrm{risk}(z, h) = \Pr\left(Y_{t+h} = 1 \mid Z_t = z\right),$$

together with the principal contrast

$$RR(H) = \frac{\mathrm{risk}(2, H)}{\mathrm{risk}(0, H)}.$$

On the primary split, the ordering was strong under both the persistent-state and mixture-state formulations.

H	$\mathrm{risk}(0, H)$	$\mathrm{risk}(1, H)$	$\mathrm{risk}(2, H)$	$RR(H)$	p-value
1	—	—	—	2.76	—
3 (persistent)	0.382	0.684	0.892	2.33	2.46×10^{-9}
3 (mixture)	0.437	0.782	0.794	1.82	4.71×10^{-10}
5	0.490	0.806	0.809	1.65	2.00×10^{-8}
7	0.497	0.813	0.816	1.64	1.68×10^{-8}

At the primary horizon $H = 3$, the corresponding Fisher odds ratio was 13.35 under the persistent-state formulation and 4.97 under the mixture-state formulation. The effect was therefore not only statistically significant, but also substantively large: occupancy of the highest-risk state implied a sharply increased probability of failure over subsequent turns.

D. Replication in Held-out Datasets

The strongest test of the hypothesis is whether the ordered risk structure persists under fully held-out evaluation. We therefore froze the pipeline learned on the primary training split and applied it unchanged to a disjoint held-out subset and to two further independent GPT evaluation sets.

On the first fully disjoint held-out subset, the ordering replicated clearly, particularly at moderate and longer horizons.

H	risk$(0,H)$	risk$(1,H)$	risk$(2,H)$	$RR(H)$	p-value
5	0.523	0.655	0.656	1.25	0.0026
7	0.543	0.682	0.697	1.28	3.93×10^{-4}
10	0.568	0.703	0.736	1.30	8.74×10^{-5}
13	—	—	—	1.33	8.21×10^{-6}
20	—	—	—	1.36	5.34×10^{-7}

Soft-state analyses on this same holdout were directionally identical, with strict soft ordering especially clear for $H \geq 7$ and high-versus-low soft-state risk ratios in the range of approximately 1.28–1.34. Holdout A reproduced the same pattern, with stronger separation at short and moderate horizons.

H	risk$(0,H)$	risk$(1,H)$	risk$(2,H)$	$RR(H)$	p-value
1	0.280	0.455	0.476	1.70	0.00188
5	0.460	0.699	0.713	1.55	5.0×10^{-5}
7	—	—	—	1.54	3.7×10^{-5}
20	0.520	0.756	0.774	1.49	2.4×10^{-5}

In Holdout A, hard-state ordering held for 8 of 9 tested horizons, whereas soft-state ordering held for all 9, with high-versus-low soft-state risk ratios ranging from approximately 1.44 to 1.70.

Holdout B was noisier at the shortest horizons, but the predicted medium- and long-horizon separation re-emerged clearly.

Horizon	$RR(h)$	**p-value**	**Ordering**
$H = 1$	—	—	Not significant
$H = 10$	1.19	0.0438	Ordering holds
$H = 13$	1.22	0.0131	Ordering holds
$H = 15$	1.25	0.00337	Ordering holds
$H = 20$	1.27	0.00185	Ordering holds

Taken together, these results show that the monotone risk-state structure is not confined to a single favorable partition, but reappears across multiple fully disjoint GPT evaluation sets, with the clearest and most stable separation at moderate and longer horizons.

E. Instability rises within conversations before failure

We next asked whether the observed phenomenon was genuinely temporal, rather than a by-product of between-conversation differences in baseline quality. To test this, we centered instability within each conversation and compared matched pre-failure windows with earlier reference windows drawn from the same conversational trajectory.

Across 312 valid pre-failure windows, mean instability in the pre-failure interval exceeded that of the matched reference interval by a substantial margin.

D_{obs}	**Null mean**	**One-tailed p-value**	**Valid windows**
0.0718	0.00036	0.0067	312

This result rules out a purely cross-sectional account. Instability does not merely characterize some conversations as globally worse than others; it rises within conversations before failure occurs.

F. Upward state transitions carry additional hazard information

Beyond static state occupancy, directional movement toward riskier regimes also carried predictive information. For turns satisfying $Z_t > Z_{t-1}$, future failure probability was significantly elevated relative to turns without such upward movement.

H	Pr $(Y_{t+H} = 1 \mid$ no up$)$	Pr $(Y_{t+H} = 1 \mid$ up$)$	**RR**	**Fisher OR**	**p-value**
3	0.698	0.889	1.273	3.45	0.0059
5	0.727	0.911	1.253	—	0.0047
7	0.733	0.933	1.273	—	0.0021

Transition direction sharpened the same pattern. At $H = 3$, future failure rates were 0.889 after upward transitions, 0.776 after no state change and 0.733 after downward transitions. At $H = 7$, the corresponding values were 0.933, 0.802 and 0.756. Thus, movement into riskier regimes carried additional hazard information beyond static state occupancy alone.

G. Directional Stability Across Repeated Splits

To test whether the findings depended on a single favorable partition, we repeated the full pipeline across ten random conversation-level splits. Although effect sizes varied with conversational composition, the main pattern remained directionally stable. Mean high-versus-low risk ratios at longer horizons remained above 1, reaching approximately 1.36 for hard-state analyses and 1.33 for soft-state analyses at the longest horizons.

This repeated-split stability indicates that the empirical effect is not an artifact of one particular partition, but a persistent property of the data-generating process.

To account for turn non-independence within conversations, key associations were re-estimated using GEE with exchangeable working correlation and conversation-level clustering. The burst coefficient remained positive and significant at the primary horizon (β=0.18, SE=0.07, z=2.51, p=0.012), and at $H = 1$ (β=0.18, SE=0.07, z=2.76, p=0.006), consistent with the main analyses. GEE estimates were unavailable at horizons 10, 15 and 17 due to convergence failure, likely reflecting reduced event density at longer horizons.

H. Empirical Summary

Taken together, the results support a temporally organized account of GPT reasoning failure. The prospective instability signal forecast future failure before discretization, and the inferred latent states showed temporal persistence, multi-turn dwell structure and an ordered relationship with future failure risk. This ordering replicated across multiple fully disjoint GPT evaluation sets, instability rose within conversations before failure, and upward transitions into higher-risk states carried additional hazard information. Collectively, these findings support the empirical monotone risk-state property

$$\Pr(Y_{t+h} = 1 \mid Z_t = 2) \geq \Pr(Y_{t+h} = 1 \mid Z_t = 1) \geq \Pr(Y_{t+h} = 1 \mid Z_t = 0),$$

with the clearest support at moderate and longer horizons.

III. Discussion

The temporal organization of reasoning failure observed here admits a plausible mechanistic interpretation. When a GPT-family model enters a high-risk conversational regime, successive responses are generated under a partially degraded inferential configuration: premises are weakly grounded, justification is thin, and apparent coherence is maintained by narrative momentum rather than active self-correction. Because autoregressive generation conditions each new output on prior model-produced text, such configurations propagate across turns. The latent state is therefore best understood not as a literal internal variable but as an observable statistical signature of this dynamical condition, a footprint of elevated future failure risk that is detectable from behavioral features alone.

This account explains a feature of the results that would otherwise be puzzling: why effects are strongest at moderate and longer horizons rather than at the immediately subsequent turn. Entry into a high-risk state need not produce immediate breakdown. Instead it opens a temporally extended vulnerability window in which instability persists beneath a surface of coherence, with failure emerging only when the conversation demands cross-turn integration or factual grounding that the degraded regime cannot sustain. The monotone risk-state ordering is not an arbitrary statistical regularity but the expected signature of a process in which internal instability is carried forward long enough to increase the hazard of eventual collapse.

These findings reframe LLM reliability as a dynamical systems problem with direct consequences for deployment monitoring and intervention design. Because the instability index and latent-state assignments are computed online from observable response features without access to outcome labels, they constitute a candidate framework for prospective safety monitoring, one that could flag high-risk regimes before a confident failure is delivered. The temporal persistence of inferred states implies that the natural unit of intervention is the multi-turn behavioral regime rather than the individual erroneous response. Accordingly, an effective intervention should be evaluated not only through reduced aggregate error incidence but through shortened dwell times in high-risk states, reduced future-failure risk separation across states and weakened pre-failure instability buildup within conversations, quantities that are directly measurable and largely architecture-agnostic.

Several limitations apply. The analyses cover GPT-family conversations from a single public corpus, and whether equivalent temporal structure emerges in other architectures or deployment contexts remains to be established. The instability signal is an operational proxy derived from behavioral surface features rather than a direct measurement of internal model state. The study is observational: temporal co-occurrence of instability and subsequent failure is documented, but the underlying causal mechanisms are not identified. Latent-state assignments depend on specific modelling choices, though the qualitative pattern replicated across four held-out datasets and ten independent partition seeds.

II. Methods

The aim of this study was to test whether GPT conversations exhibit latent instability phases that precede and predict reasoning failure. Specifically, we evaluated an empirical monotone risk-state property: for a discrete latent state $Z_t \in \{0,1,2\}$assigned at turn t, and a binary indicator Y_{t+h}denoting whether failure occurs within a subsequent horizon h, we tested whether

$$\Pr(Y_{t+h} = 1 \mid Z_t = 2) \geq \Pr(Y_{t+h} = 1 \mid Z_t = 1) \geq \Pr(Y_{t+h} = 1 \mid Z_t = 0).$$

Under this formulation, higher latent state corresponds to greater prospective failure risk. Testing this claim required four components: a turn-level operational definition of failure, a prospective scalar instability signal computable without ground-truth labels at inference time, a procedure for converting that signal into discrete latent states, and statistical tests that directly evaluate persistence, dwell, ordered future risk and within-conversation temporal buildup.

This study focuses exclusively on GPT-family models. As discussed in the main text, the framework may extend to other

autoregressive architectures if the same underlying mechanism is present.

A. Evaluation datasets

Multi-turn GPT conversations were drawn from a publicly available Kaggle corpus of approximately 89,000 GPT-3.5 and GPT-4 conversations. From this source, we constructed one primary evaluation dataset and two held-out datasets.

Split	Conv Range	Convs.	Turns	Failures	Fail
Primary	10k–20k	553	3,242	1,480	45.7
Holdout A	19k–25k pool	84	616	303	49.2
Holdout B	19k–25k pool	83	629	256	40.7

Within the primary dataset, conversations were partitioned at the conversation level into training, validation and test subsets in a 60/20/20 ratio so that no turns from the same conversation appeared in multiple splits. Holdout A and Holdout B were evaluated using the same frozen pipeline trained on the primary training split.

B. Operational definition of failure

Failure in open-ended dialogue cannot be reduced to a single error type. We therefore adopted a multi-category labeling framework in which each assistant turn was evaluated for the presence of a substantive, expert-identifiable defect. Minor stylistic weakness, harmless ambiguity and brevity were not treated as failures.

Each assistant turn was labeled by Gemini-2.5-Flash acting as a deterministic zero-shot labeling oracle before being reviewed in a second round by human judges. The model was shown the full conversation transcript with the target turn explicitly marked, allowing contextual assessment of contradiction, relevance and coherence. The rubric consisted of ten mutually exclusive categories:

- **none**: broadly acceptable response; minor issues do not qualify
- **hallucination**: fabricated or unsupported factual claims, entities or numbers
- **logical_inconsistency**: conclusions do not follow from premises
- **self_contradiction**: explicit contradiction within the same response
- **non_answer**: failure to substantively address the user's request
- **unsupported_overclaim**: claims stronger than warranted by the available evidence
- **incoherent_reasoning**: semantically broken or disorganized reasoning
- **harmful_or_misleading_guidance**: dangerous or materially misleading advice
- **severe_incompleteness**: omission so substantial that the response becomes unusable
- **other**: clear failure not captured by the above categories

The primary binary outcome was defined as

$$Y_t = \begin{cases} 1, & \text{if failure type}_t \neq \text{none} \\ 0, & \text{otherwise.} \end{cases}$$

A confidence label (high, medium or low) was also collected but was not used in downstream analyses. The labeling prompt enforced a high threshold for failure designation: a failure was to be assigned only when a reasonable expert reviewer would clearly judge the response to be meaningfully defective. Tie-breaking instructions were included to improve consistency when multiple categories appeared plausible. Temperature was fixed at zero to ensure deterministic outputs. To ensure that the Gemini-2.5-Flash labeling was consistent and fair, turns were reviewed at random by humans. Across 10 independent samples of 25 randomly selected turns, human reviewers blindly judged a mean of 24 turns per sample (96%) to be correctly labeled, with the remaining cases considered defensible under the rubric.

C. Reasoning-quality classifier

To construct a prospective instability signal without directly using the downstream binary failure label, we first trained a multi-label reasoning-quality classifier that scored each assistant turn on seven interpretable dimensions of response quality. The resulting criterion-level probabilities were then transformed into temporal instability features. Because these quantities were produced by a frozen model trained on a disjoint dataset, they provided a prospective signal independent of the downstream binary failure annotation.

The classifier was trained on 1,158 assistant turns drawn from 99 GPT conversations. Of these, 924 were original turns and 234 were synthetic augmentations introduced to improve balance and robustness. The final training set contained 558 failure turns with aggregate reasoning score <7 and 600 passing turns with aggregate reasoning score $= 7$, producing an approximately balanced training distribution.

Each turn was labeled under a seven-criterion binary rubric. These labels were used only to train the classifier bundle and were not reused as downstream test labels in the temporal analyses. The seven criteria were:

1. relevance to prompt
2. directly addresses question

3. step-by-step or structured reasoning
4. uses justification or explanation
5. internally consistent
6. acknowledges uncertainty or limits when needed
7. sufficiently complete for prompt

These criteria were chosen to cover the major response-quality dimensions that can be assessed without external ground truth. Task fidelity was captured by relevance and directness; reasoning quality by structure and justification; logical coherence by internal consistency; epistemic calibration by appropriate acknowledgement of uncertainty; and coverage by completeness. Fluency and grammaticality were excluded because frontier-class GPT models fail on these dimensions rarely in naturalistic dialogue and thus contribute little discriminative value. External factual accuracy was excluded because it cannot be assessed prospectively without external knowledge, which would violate the design goal of a deployment-usable signal.

For turn t, the aggregate reasoning score was defined as

$$S_t = \sum_{k=1}^{7} c_{t,k}\,, c_{t,k} \in \{0,1\},$$

so that $S_t \in [0,7]$. Across the 1,158-turn training set, positive-label means were 0.770 for relevance, 0.699 for directness, 0.692 for structured reasoning, 0.743 for justification, 0.848 for internal consistency, 0.804 for uncertainty acknowledgement and 0.595 for completeness. Each training example was represented by a concatenated prompt–response text field,

$$\text{text}_t = [\text{PROMPT}] + \text{prompt}_t + [\text{RESPONSE}] + \text{response}_t.$$

Text features were encoded using TF–IDF with sublinear term-frequency scaling, lowercase normalization, Unicode accent stripping, and unigram–bigram tokenization $(1, 2)$, with a vocabulary cap of 12,000 features.

Step markers included *first, second, therefore, thus, because, if, then, finally,* and *step*. Hedge markers included *maybe, perhaps, possibly, might, could, likely,* and *probably*. Contradiction markers included *however, but, although, yet,* and *nevertheless*. Numeric features were standardized using training-set means and standard deviations only. The final feature representation was

$$X_t = [\text{TF} - \text{IDF}(\text{text}_t) \mid \text{scale}(x_t)].$$

D. Model training

A one-vs-rest logistic regression classifier was trained to predict each of the seven binary criteria. Each component model used L2 regularization with $C = 2.0$, balanced class weights, the liblinear solver, and a maximum of 4,000 iterations. The seven models were combined in a `OneVsRestClassifier` framework. In addition, a Ridge regression model with $\alpha = 3.0$ was trained to predict the aggregate score S_t directly; this auxiliary regressor was used only in robustness analyses and did not define the primary instability signal.

Classifier performance was assessed by five-fold cross-validation using micro-averaged F1, macro-averaged F1, exact-match accuracy and per-label F1. Performance was sufficient to justify the model's use as a frozen feature extractor for downstream temporal analyses. After training, the classifier bundle was frozen. For each evaluation turn t, we extracted a seven-dimensional vector of criterion-level failure probabilities,

$$p_t = [p_{t,1}, p_{t,2}, \ldots, p_{t,7}],$$

where

$$p_{t,k} = \Pr(\text{criterion } k \text{ fails} \mid \text{text}_t, x_t).$$

We then defined the expected number of failed reasoning criteria as

$$f_t = \sum_{k=1}^{7} p_{t,k}.$$

This scalar f_t served as the primary input to instability-feature construction.

E. Instability Feature Construction

A single-turn estimate $f_t$is an imperfect proxy for latent reasoning quality. Our central hypothesis was that failure does not arise as isolated turn-level noise, but through entry into persistent high-risk regimes. We therefore constructed temporal features that summarized the recent trajectory of $f_t$within each conversation.

Let $W_S = 3$ and $W_L = 10$ denote short and long rolling windows. To capture recent upward movement in expected failure burden, we defined

$$\delta_t = \text{mean}_{s\in[t-W_S,t]} f_s - \text{mean}_{s\in[t-W_L,t]} f_s.$$

Positive values indicate that recent turns exhibit greater expected failure burden than the longer-run local baseline. We next defined a high-risk event indicator

$$e_t = \begin{cases} 1, & f_t \geq \tau \\ 0, & f_t < \tau, \end{cases}$$

where τ was selected from the training set as the validation-optimal quantile of f_t from the grid: $\{0.55, 0.60, 0.65, 0.70, 0.75, 0.80\}$.

Burst was defined as the longest contiguous run of high-risk events within the short window, normalized by window size:

$$\text{burst}_t = \frac{\text{LCR}(\{e_s : s \in [t - W_S, t]\})}{W_S},$$

where LCR denotes longest contiguous run. This feature captures short-term persistence of elevated risk rather than isolated spikes.

To detect localized degradation in any one reasoning dimension, we defined

$$\text{maxshift}_t = \max_{k \in \{1,\ldots,7\}} \left(\text{mean}_{s \in [t-W_S,t]} p_{s,k} - \text{mean}_{s \in [t-W_L,t]} p_{s,k}\right).$$

All three features were standardized using training-set moments only:

$$\delta_t^{(z)} = \frac{\delta_t - \mu_\delta}{\sigma_\delta}, \text{burst}_t^{(z)} = \frac{\text{burst}_t - \mu_{\text{burst}}}{\sigma_{\text{burst}}}, \text{maxshift}_t^{(z)} = \frac{\text{maxshift}_t - \mu_{\text{maxshift}}}{\sigma_{\text{maxshift}}}.$$

These standardized features were compressed into a single instability index by principal component analysis fit on the training set:

$$I_t = \text{PC1}(\delta_t^{(z)}, \text{burst}_t^{(z)}, \text{maxshift}_t^{(z)}).$$

The sign of I_t was oriented so that larger values corresponded to higher instability. Specifically, if the Spearman correlation between I_t and future failure at horizon $H = 3$ on the training set was negative, the first principal component was multiplied by -1.

F. Latent-State Discovery

The theoretical framework concerns discrete latent risk states rather than a purely continuous instability scale. We therefore partitioned the instability index into three ordered states—low, intermediate and high risk—as the smallest representation capable of expressing qualitatively distinct regimes while remaining statistically interpretable.

K-means clustering with $k = 3$and 20 random restarts was fit to training-set values of I_t. Let the ordered centroids satisfy

$$\mu_0 < \mu_1 < \mu_2.$$

State assignments were then defined by nearest-centroid classification:

- $Z_t = 0$: low-risk state
- $Z_t = 1$: intermediate-risk state
- $Z_t = 2$: high-risk state

Centroids were frozen after training, and validation, test and held-out datasets were assigned by nearest-centroid classification without refitting. Soft state-membership weights based on normalized inverse-squared distances were also computed for supplementary analyses.

The central empirical claim tested in this study was that, for each tested horizon h,

$$\Pr(Y_{t+h} = 1 \mid Z_t = 2) \geq \Pr(Y_{t+h} = 1 \mid Z_t = 1) \geq \Pr(Y_{t+h} = 1 \mid Z_t = 0),$$

with the ordering expected to be most stable at moderate and longer horizons. In addition, the state sequence was expected to exhibit non-random transition dynamics and multi-turn persistence.

G. Statistical Tests

Four hypotheses were evaluated in total. **H1: state persistence.** The observed state-transition matrix differs from that expected under independent random assignment. Let N_{ij}denote the number of observed transitions from state $i$to state j. Expected counts under independence are

$$E_{ij} = \frac{\left(\sum_k N_{ik}\right)\left(\sum_k N_{kj}\right)}{\sum_{i,j} N_{ij}}.$$

The test statistic is

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - E_{ij})^2}{E_{ij}},$$

which is asymptotically χ^2-distributed with $(3-1)^2 = 4$ degrees of freedom.

H2: multi-turn dwell. States persist across consecutive turns, producing run lengths above random expectation. For each state z, contiguous runs within conversations were

identified and summarized by lengths $r_{c,z,j}$. Mean run length was computed as

$$R_z = \text{mean}_{c,j}\, r_{c,z,j}.$$

Under a random three-state process with equal state probability, the expected mean run length is 1.5 turns.

H3: ordered future risk. Higher latent states imply higher future failure probability. For each horizon $h$and state z, future failure risk was defined as

$$\text{risk}(z, h) = \Pr\,(Y_{t+h} = 1 \mid Z_t = z).$$

The principal contrast was the high-versus-low risk ratio

$$RR(h) = \frac{\text{risk}(2, h)}{\text{risk}(0, h)}.$$

Significance was assessed using Fisher's exact test. Spearman rank correlation between $Z_t$and $Y_{t+h}$was also reported as a non-parametric sensitivity analysis. For soft-state analyses, weighted risk was computed as

$$\text{soft\_risk}(z, h) = \frac{\sum_t w_{t,z}\, Y_{t+h}}{\sum_t w_{t,z}},$$

where $w_{t,z}$denotes the soft membership weight.

H4: within-conversation temporal buildup. Instability rises in the turns immediately preceding failure relative to matched within-conversation reference windows. For each failure event at turn t_f, we defined a pre-failure window $[t_f - W, t_f - 1]$ and a matched reference window $[t_f - 2W, t_f - W - 1]$, with $W = 5$. Within-conversation centered instability was

$$I_t^{\text{within}} = I_t - \text{mean}_{s \in c} I_s.$$

For each failure event, we computed

$$d_{c,t_f} = \text{mean}_{s \in \text{pre}} I_s^{\text{within}} - \text{mean}_{s \in \text{ref}} I_s^{\text{within}}.$$

The observed summary statistic was

$$D_{\text{obs}} = \text{mean}_{c,t_f} d_{c,t_f}.$$

A one-tailed permutation test with 150 within-conversation random shuffles was used to generate the null distribution. Across the nine-horizon sweep, p-values from slope and Spearman analyses were adjusted using the Benjamini–Hochberg false discovery rate procedure. The primary horizon was pre-specified as $H = 3$; remaining horizons were treated as exploratory.

Because turns within a conversation are not independent, key associations were additionally estimated using generalized estimating equations with a logistic link, exchangeable working correlation structure, and conversation-level clustering. To prevent leakage, the following quantities were estimated exclusively on the training data and then frozen: TF–IDF vocabulary and inverse-document-frequency weights, numeric-feature standardization parameters, reasoning-classifier coefficients, burst threshold τ, PCA components and sign orientation, and K-means centroids. The complete frozen pipeline was then applied unchanged to validation, test, Holdout A and Holdout B.

This protocol ensured that, at turn t, the instability index I_t depended only on information available up to and including turn t, together with parameters estimated from training data only. It therefore functions as a valid prospective indicator rather than a retrospective summary. To assess sensitivity to partition choice, the full pipeline was repeated across 10 random conversation-level splits (seeds 10042–10051), and the distribution of predictive metrics, including PR–AUC and risk ratios, was recorded. We then applied this frozen pipeline to the primary test split and all held-out datasets to evaluate persistence, dwell, ordered future risk, and within-conversation temporal buildup.

References

[1] A. Arteaga, J. Schön, and N. Pielawski, "Hallucination Detection in LLMs: Fast and Memory-Efficient Fine-Tuned Models," *arXiv preprint*, 2024.

[2] A. Ahadian and Y. Guan, "A Survey on Hallucination in Large Language and Foundation Models," *Preprints.org*, 2025.

[3] B. Bang et al., "HalluLens: LLM Hallucination Benchmark," *arXiv preprint*, 2025.

[4] H. Cheng et al., "Integrative Decoding: Improve Factuality via Implicit Self-consistency," *arXiv preprint*, 2024.

[5] 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, p. 103425, 2023.

[6] S. Dhuliawala et al., "Chain-of-Verification Reduces Hallucination in Large Language Models," *arXiv preprint*, 2023.

[7] X. Du, C. Xiao, and Y. Li, “HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection,” *arXiv preprint*, 2024.

[8] 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, 2024.

[9] J. Gawlikowski et al., “A Survey of Uncertainty in Deep Neural Networks,” *arXiv preprint*, 2022.

[10] D. Hendrycks and K. Gimpel, “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks,” *arXiv preprint*, 2016.

[11] L. Huang et al., “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *ACM Transactions on Information Systems*, 2024.

[12] S. Ivgi, O. Yoran, J. Berant, and M. Geva, “From Loops to Oops: Fallback Behaviors of Language Models Under Uncertainty,” *arXiv preprint*, 2024.

[13] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, “How Can We Know What Language Models Know?,” *arXiv preprint*, 2020.

[14] S. Kang et al., “Uncertainty Quantification for Hallucination Detection in Large Language Models: Foundations, Methodology, and Future Directions,” *arXiv preprint*, 2025.

[15] J. Ko, S. Baek, and S. Hwang, “Real-time Verification and Refinement of Language Model Text Generation,” *arXiv preprint*, 2025.

[16] J. Kossen et al., “Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs,” *arXiv preprint*, 2024.

[17] A. Kulkarni et al., “Evaluating Evaluation Metrics — The Mirage of Hallucination Detection,” *arXiv preprint*, 2025.

[18] M. Lage and S. Ostermann, “OpenFActScore: Open-Source Atomic Evaluation of Factuality in Text Generation,” *arXiv preprint*, 2025.

[19] Y. Li et al., “Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources,” *arXiv preprint*, 2023.

[20] Y. Li et al., “The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models,” *arXiv preprint*, 2024.

[21] Y. Li et al., “Loki's Dance of Illusions: A Comprehensive Survey of Hallucination in Large Language Models,” *arXiv preprint*, 2025.

[22] Y. Liu et al., “The Scales of Justitia: A Comprehensive Survey on Safety Evaluation of LLMs,” *arXiv preprint*, 2025.

[23] Y. Lu et al., “Toward Human-Like Evaluation for Natural Language Generation with Error Analysis,” *IEEE Access*, 2023.

[24] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,” *EMNLP*, 2023.

[25] A. Marinescu et al., “FactReasoner: A Probabilistic Approach to Long-Form Factuality Assessment for Large Language Models,” *arXiv preprint*, 2025.

[26] T. Nie et al., “FactTest: Factuality Testing in Large Language Models with Finite-Sample and Distribution-Free Guarantees,” *arXiv preprint*, 2024.

[27] Y. Ovadia et al., “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift,” *arXiv preprint*, 2019.

[28] B. Paudel et al., “HalluciNot: Hallucination Detection Through Context and Common Knowledge Verification,” *arXiv preprint*, 2025.

[29] R. Phillips et al., “Geometric Uncertainty for Detecting and Correcting Hallucinations in LLMs,” *arXiv preprint*, 2025.

[30] X. Piskala et al., “Mind the Goal: Data-Efficient Goal-Oriented Evaluation of Conversational Agents and Chatbots using Teacher Models,” *arXiv preprint*, 2025.

[31] Y. Qi, J. He, and X. Yuan, “Can We Catch the Elephant? The Evolvement of Hallucination Evaluation on Natural Language Generation: A Survey,” *arXiv preprint*, 2024.

[32] 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*, 2024.

[33] Z. Wu et al., “Improve Decoding Factuality by Token-wise Cross Layer Entropy of Large Language Models,” *arXiv preprint*, 2025.

[34] Z. Xu, S. Jain, and M. Kankanhalli, “Hallucination is Inevitable: An Innate Limitation of Large Language Models,” *arXiv preprint*, 2024.

[35] Z. Yang, H. Yoo, and J. Lee, “MAQA: Evaluating Uncertainty Quantification in LLMs Regarding Data Uncertainty,” *arXiv preprint*, 2024.

[36] Y. Wang et al., “Survey on Factuality in Large Language Models,” *ACM Computing Surveys*, 2025.

[37] X. Zhang et al., “SAC3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency,” *EMNLP*, 2023.

[38] X. Zhang et al., “Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus,” *EMNLP*, 2023.