IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FROM CALIBRATION TO HALLUCINATION PREDICTION: INVESTIGATING TOKEN-LEVEL UNCERTAINTY DYNAMICS IN LARGE LANGUAGE MODELS

Ishan Kumar Gupta
Student
Department of Mining Engineering
Indian Institute of Technology, Kharagpur

Abstract

The remarkable capabilities of modern Large Language Models (LLMs) have been accompanied by a persistent reliability challenge: the generation of factually incorrect yet highly confident responses. Existing approaches to uncertainty quantification primarily focus on response-level confidence estimates and calibration metrics, implicitly assuming that hallucinations are associated with elevated uncertainty. However, empirical observations frequently contradict this assumption, as incorrect generations often exhibit confidence levels comparable to those assigned to correct outputs.

This study investigates whether the temporal evolution of uncertainty provides a more informative signal than uncertainty magnitude alone. The research proceeds through a progressively refined analysis of model reliability, beginning with calibration assessment, extending to token-level uncertainty characterization, and culminating in uncertainty-driven hallucination prediction. Calibration experiments on **GPT-4o** and **Gemini 3.5** reveal substantial overconfidence despite strong predictive performance, motivating a transition from response-level confidence analysis to token-wise uncertainty modelling.

Using **Qwen2.5-1.5B-Instruct**, token-level predictive entropy trajectories were extracted throughout autoregressive generation. A suite of uncertainty-derived features, including **mean entropy**, **entropy variance**, **entropy spike frequency**, **entropy jump magnitude**, and **entropy drift**, was subsequently evaluated on a human-annotated dataset of factual and hallucinated generations. Contrary to conventional expectations, aggregate uncertainty statistics exhibited limited discriminative power. In contrast, entropy drift emerged as a robust predictive signal, achieving a cross-validated **ROC-AUC of approximately 0.80.**

The findings suggest that hallucinations are not primarily characterized by elevated uncertainty levels, but rather by distinctive temporal uncertainty dynamics. More broadly, the results indicate that uncertainty evolution constitutes a substantially richer source of information than static uncertainty measures and may provide a promising foundation for future hallucination detection and reliability estimation systems.

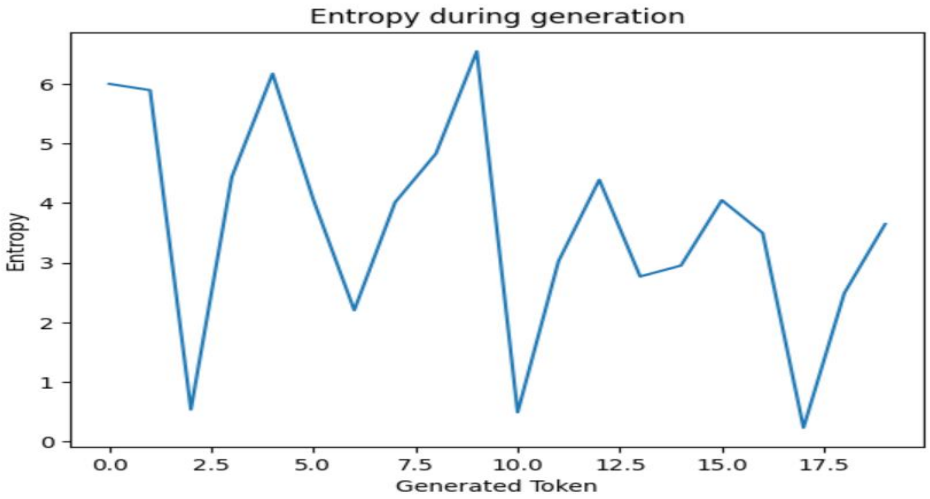

Figure 1. Shannon entropies (H) of all the generated tokens for a given prompt in GPT-2o

1. Introduction

The rapid advancement of Large Language Models has transformed the landscape of natural language processing, enabling systems to perform increasingly sophisticated reasoning, knowledge retrieval, and generative tasks. Despite these advances, hallucinations remain a fundamental limitation. Models frequently produce responses that are syntactically coherent, semantically plausible, and delivered with apparent confidence, yet factually incorrect.

The challenge is particularly significant because confidence and correctness are often poorly aligned. A model may exhibit high confidence in a false statement while expressing uncertainty regarding a correct one. Consequently, understanding the relationship between uncertainty and factual reliability has become an important research direction within contemporary LLM interpretability and evaluation literature.

Most existing analyses treat uncertainty as a static quantity. Metrics such as **confidence scores**, **entropy averages**, and **calibration errors** summarize uncertainty at the response level. While informative, these aggregate statistics neglect the inherently sequential nature of language generation. Each generated token is produced under a changing probability distribution, implying that uncertainty itself evolves throughout the generation process.

This observation motivates the central hypothesis of the present work:

The temporal dynamics of uncertainty may provide more information about hallucination behaviour than uncertainty magnitude alone.

To investigate this hypothesis, the study was structured around three complementary objectives:

1. Evaluate whether modern LLMs are well calibrated.
2. Characterize uncertainty at the token level during generation.
3. Determine whether uncertainty dynamics can predict hallucinations.

Together, these analyses establish a progression from model reliability assessment to uncertainty-based hallucination prediction.

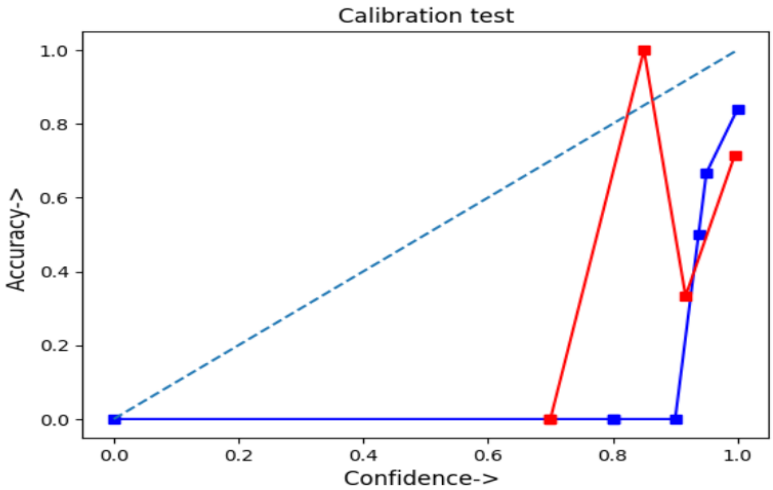

Figure 2. Reliability diagrams comparing predicted confidence and empirical accuracy across GPT-4o and Gemini 3.5.

2. Motivation and Research Questions

The project originated from a simple but consequential observation: incorrect model outputs frequently appear just as confident as correct ones.

This observation raises several research questions:

RQ1: Are modern LLMs well calibrated?

RQ2: Do hallucinations exhibit elevated uncertainty relative to correct generations?

RQ3: Can token-level uncertainty trajectories reveal patterns that are invisible to aggregate uncertainty statistics?

RQ4: Which uncertainty-derived features, if any, possess predictive utility for hallucination detection?

Rather than assuming that hallucinations are associated with high uncertainty, the study explicitly tests whether uncertainty dynamics themselves constitute the defining signal.

3. Calibration Analysis

Before investigating token-level uncertainty, it is necessary to establish whether model-reported confidence is itself a reliable indicator of factual correctness. Confidence calibration provides a natural starting point because it measures the degree to which predicted confidence aligns with empirical accuracy.

Two state-of-the-art proprietary language models, GPT-4o and Gemini 3.5, were evaluated using a suite of calibration metrics, including Expected Calibration Error (ECE), average confidence, accuracy, overconfidence gap, and HCWR. Reliability diagrams were additionally employed to visualize deviations from ideal calibration.

The reliability diagrams reveal systematic departures from ideal calibration. Both models frequently assign confidence scores that exceed their observed accuracy, indicating a tendency toward overconfidence. Although both systems demonstrate strong predictive performance, confidence estimates are consistently inflated relative to empirical correctness.

The calibration metrics further reinforce this observation. While accuracy remains relatively high, the observed confidence values substantially exceed realized performance. This discrepancy highlights a fundamental limitation of response-level confidence estimation: confidence alone is insufficient for reliably identifying factual correctness.

The implications of this finding are significant. If incorrect generations frequently receive confidence scores comparable to correct generations, then response-level confidence may conceal important uncertainty signals. Consequently, a more granular analysis becomes necessary. Rather than treating uncertainty as a single scalar quantity, the remainder of this study investigates uncertainty at the token level and examines how it evolves during generation.

4. Token-Level Uncertainty Characterization

Large Language Models generate text autoregressively, producing each token conditioned on all previously generated tokens. At every generation step, the model constructs a probability distribution over the vocabulary, from which predictive uncertainty can be quantified using Shannon entropy.

Formally, token-level entropy is defined as:

$$H(p) = - \Sigma\, p(x) \log p(x)$$

where $p(x)$ denotes the predicted probability assigned to token x.

Entropy provides a natural measure of uncertainty because it captures the dispersion of probability mass across the vocabulary. Highly concentrated distributions correspond to low entropy and high confidence, whereas diffuse distributions correspond to high entropy and increased uncertainty.

Unlike response-level confidence metrics, token-wise entropy enables uncertainty to be examined as a dynamic process unfolding throughout generation. For each generated response, entropy values were extracted at every generation step, producing a complete uncertainty trajectory.

The following uncertainty descriptors were subsequently derived:

- Mean Entropy
- Entropy Variance
- Entropy Spike Rate
- Entropy Jump Magnitude
- Entropy Drift

Collectively, these features were designed to capture complementary aspects of uncertainty behaviour, including magnitude, volatility, local instability, and long-range temporal evolution.

5. Mean Entropy Trajectories

A natural first hypothesis is that hallucinations should exhibit elevated uncertainty throughout generation. To evaluate this hypothesis, entropy trajectories were normalized to a fixed length and averaged separately for correct and incorrect generations.

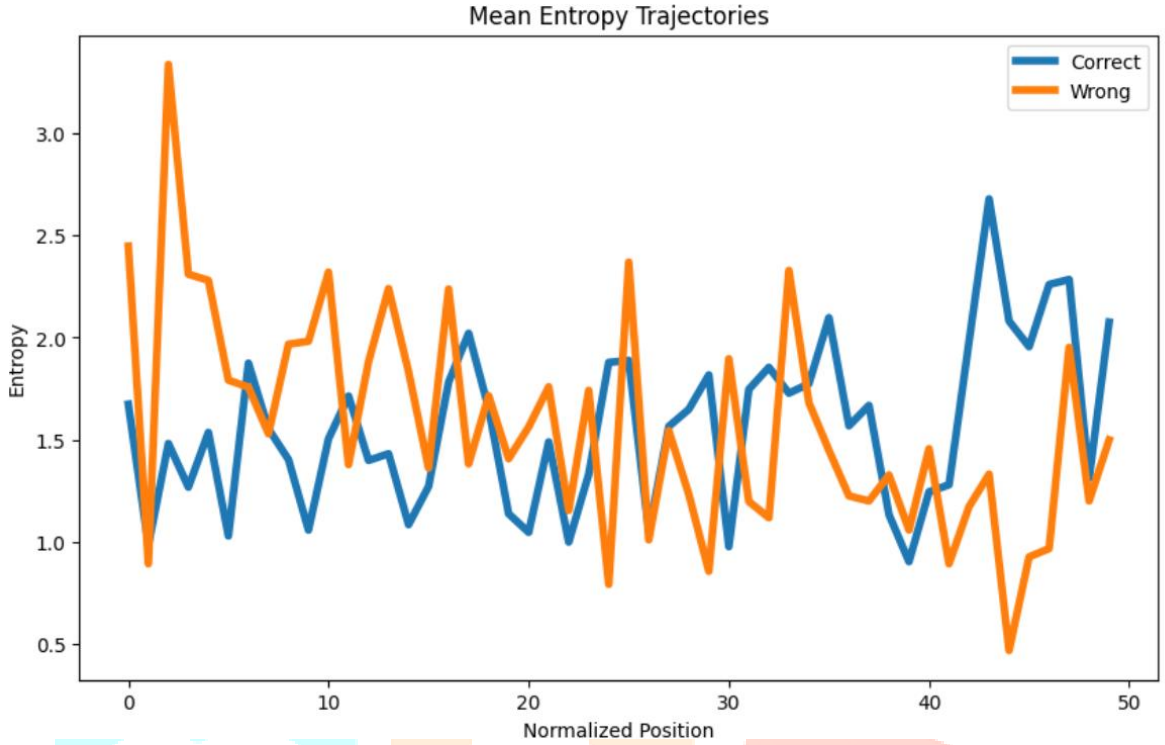

Figure 3. Average normalized entropy trajectories for correct and incorrect generations.

Contrary to expectation, the resulting trajectories exhibit substantial overlap across nearly the entire generation horizon. Although minor local deviations are observable, no systematic separation emerges between the two classes.

This finding suggests that hallucinations are not consistently associated with higher uncertainty levels. In practical terms, incorrect generations frequently display uncertainty magnitudes comparable to those observed in correct responses.

The absence of trajectory-level separation provides an important negative result. It indicates that average uncertainty evolution alone does not provide a reliable basis for hallucination prediction.

6. Entropy Volatility and Local Instability

While average uncertainty levels may be similar, hallucinations could potentially exhibit greater volatility or instability.

To investigate this possibility, entropy variance was computed for every generation.

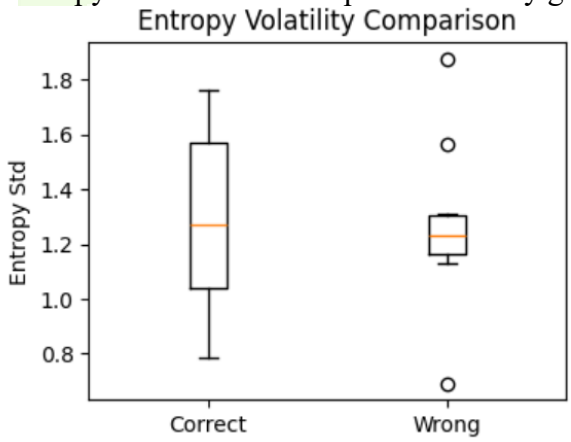

Figure 4. Distribution of entropy variance across correct and incorrect generations.

The distributions exhibit extensive overlap, with only negligible differences between classes. Variance therefore provides limited discriminative utility.

A complementary analysis examined entropy spike frequency. Spikes were defined as entropy values exceeding one standard deviation above the trajectory mean.

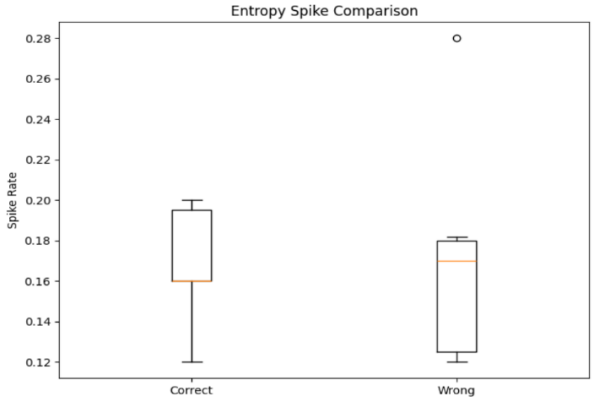

Figure 5. Entropy spike-rate distributions.

The average spike rates were virtually identical across correct and incorrect generations. This observation indicates that localized bursts of uncertainty occur at comparable frequencies regardless of factual correctness.

Taken together, these results suggest that neither uncertainty magnitude nor uncertainty volatility constitutes a defining signature of hallucination behaviour.

7. Entropy Jumps and Local Uncertainty Transitions

Although aggregate volatility fails to distinguish classes, hallucinations may nevertheless exhibit abrupt token-to-token uncertainty transitions.

To test this hypothesis, entropy jumps were computed as the absolute difference between consecutive entropy values.

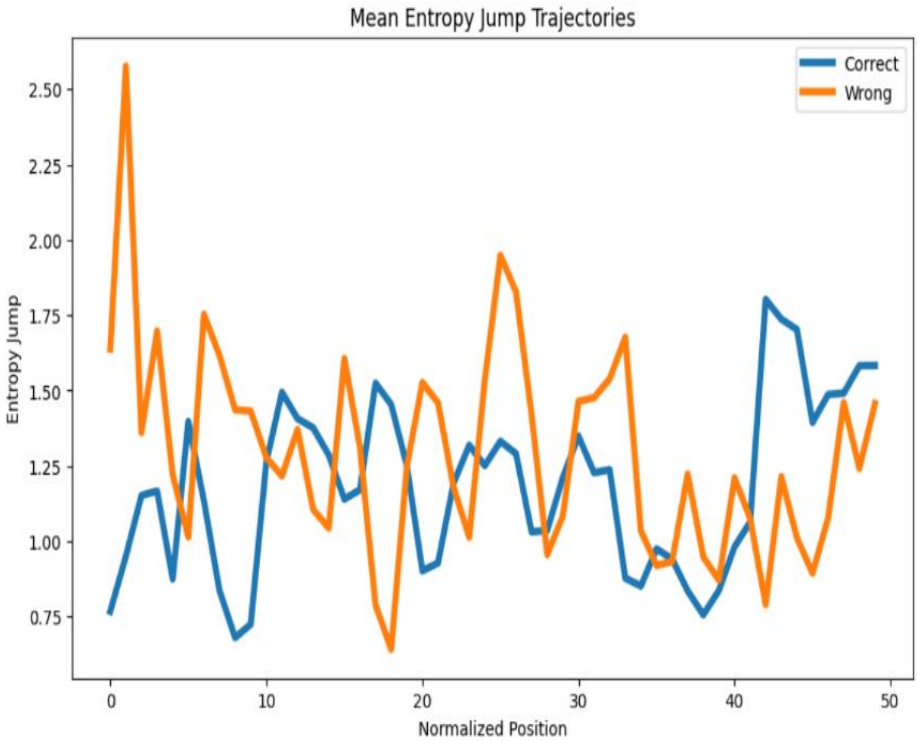

Figure 6. Entropy-jump trajectories for correct and incorrect generations.

While incorrect generations display slightly larger average jump magnitudes, the distributions remain heavily intermixed. Correct responses frequently exhibit jump patterns comparable to those observed in hallucinated outputs.

The substantial overlap suggests that local uncertainty transitions alone do not provide a sufficiently robust signal for hallucination discrimination.

At this stage of the investigation, every uncertainty feature examined had failed to produce meaningful separation between correct and incorrect generations.

However, one important possibility remained unexplored.

All preceding analyses focused on uncertainty magnitude or local fluctuations. None explicitly captured the direction of uncertainty evolution throughout generation.

This observation motivated the introduction of entropy drift.

8. Entropy Drift: The Emergence of a Predictive Signal

The preceding analyses produced a consistent conclusion: conventional uncertainty statistics exhibit limited ability to distinguish factual generations from hallucinations. Mean entropy trajectories displayed substantial overlap, entropy variance provided little discriminative information, spike frequencies were nearly identical across classes, and entropy jump magnitudes failed to reveal meaningful separation.

Collectively, these findings challenge the intuitive assumption that hallucinations are simply associated with higher uncertainty.

However, an important limitation remained. All previously evaluated metrics focused primarily on uncertainty magnitude or local fluctuations. None explicitly captured the long-range temporal evolution of uncertainty throughout the generation process.

This observation motivated the introduction of **entropy drift**, a feature designed to quantify the directionality of uncertainty dynamics.

Entropy drift was defined as:

Entropy Drift = Mean Late-Stage Entropy − Mean Early-Stage Entropy

where each generation trajectory was partitioned into an early segment and a late segment. Positive drift indicates increasing uncertainty over time, whereas negative drift indicates decreasing uncertainty as generation progresses.

Unlike previously evaluated features, entropy drift captures how uncertainty evolves throughout an entire generation rather than measuring uncertainty at isolated points.

8.1 Distributional Analysis

The first evaluation examined the distribution of entropy drift values across correct and incorrect generations.

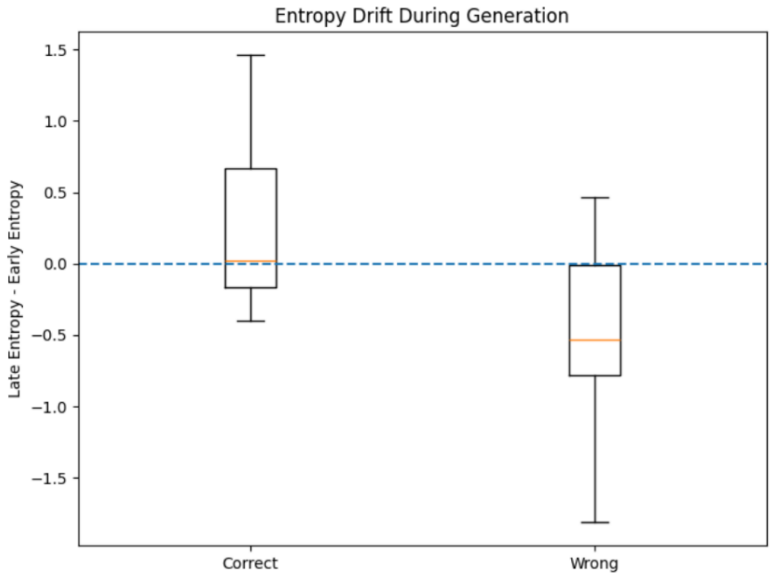

Figure 7. Distribution of entropy drift values for factual and hallucinated generations.

A clear separation emerges between the two classes. Correct generations predominantly exhibit positive entropy drift, whereas hallucinated generations display negative entropy drift. The median values of the two distributions fall on opposite sides of zero, indicating fundamentally different uncertainty behaviours.

Importantly, this pattern was not observed for any previously evaluated uncertainty metric.

Whereas earlier features exhibited extensive overlap and near-random discrimination capability, entropy drift produces a systematic distributional shift that is visible even before formal predictive evaluation.

This result suggests that factual and hallucinated generations may operate under distinct uncertainty regimes.

8.2 Distributional Robustness

To verify that the observed separation was not driven by a small number of extreme observations, the distribution of entropy drift values was examined in greater detail.

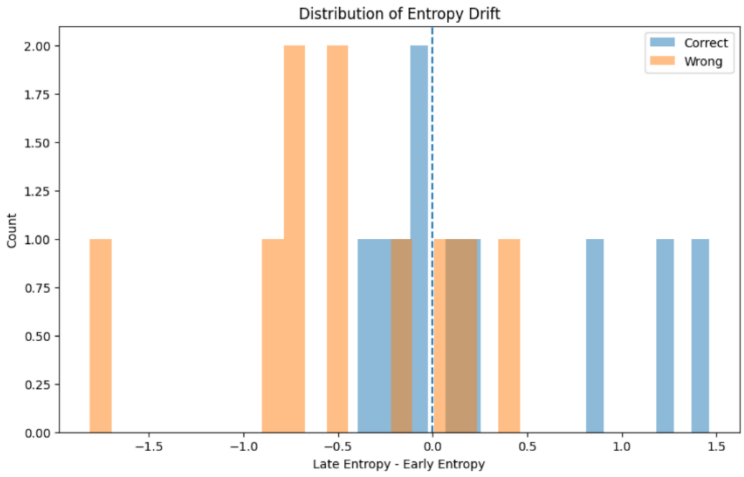

Figure 8. Histogram of entropy drift values for factual and hallucinated generations.

The histogram confirms that the separation extends across a substantial portion of the dataset rather than being attributable to isolated outliers.

Although overlap remains present—as expected in a realistic prediction problem—the overall distributional structure differs markedly between the two classes. Correct generations cluster around positive drift values, whereas hallucinated generations are concentrated within the negative region.

This finding is particularly significant because it demonstrates that uncertainty evolution exhibits systematic behavioural differences even when uncertainty magnitude itself remains comparable.

The result therefore provides the first strong evidence that hallucination prediction may depend more on temporal uncertainty dynamics than on uncertainty intensity.

8.3 Early-Stage versus Late-Stage Uncertainty

To better understand the mechanism underlying entropy drift, average entropy values were compared between the early and late portions of generation.

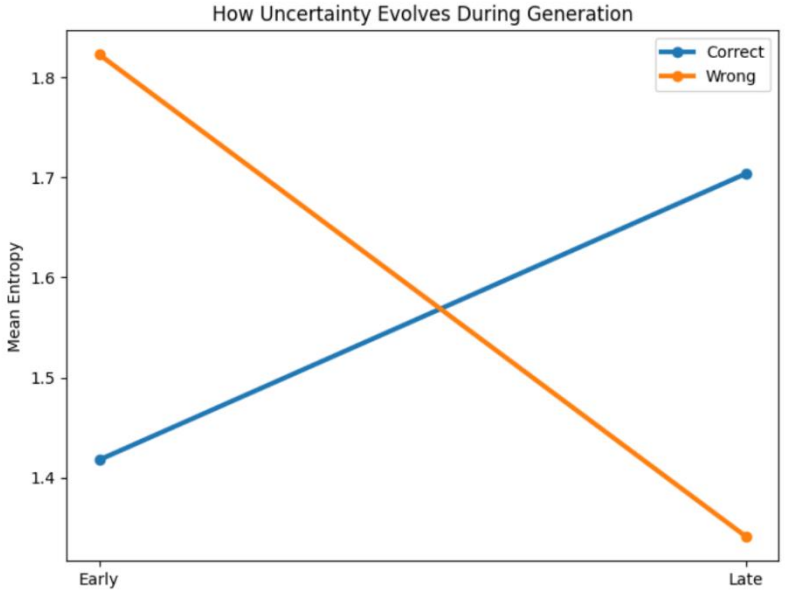

Figure 9. Comparison of early-stage and late-stage entropy for correct and incorrect generations.

A striking crossing pattern emerges.

Correct generations display increasing uncertainty throughout generation, resulting in late-stage entropy values that exceed early-stage entropy. In contrast, hallucinated generations exhibit the opposite behaviour, with uncertainty progressively decreasing over time.

This crossing pattern represents the most important empirical observation of the study.

Notably, the distinction does not arise because hallucinations are more uncertain. Rather, the distinction arises because uncertainty evolves differently.

Correct responses appear to maintain or accumulate uncertainty as generation proceeds, whereas hallucinated responses become increasingly confident.

From an interpretability perspective, this finding suggests that hallucinations may be associated with premature confidence consolidation. Once an incorrect trajectory is established, the model appears to commit to that trajectory with increasing certainty, despite the factual inaccuracy of the resulting output.

Although additional investigation is required to establish causal mechanisms, the observed pattern provides a plausible explanation for why hallucinations often appear highly confident despite being incorrect.

8.4 Implications for Hallucination Prediction

The emergence of entropy drift fundamentally alters the interpretation of uncertainty in language models.

Traditional uncertainty analyses implicitly assume that hallucinations should exhibit elevated uncertainty. The results of this study provide little support for that assumption. Static uncertainty measures consistently failed to distinguish correct and incorrect generations.

Instead, the findings suggest that the critical signal lies in the trajectory of uncertainty rather than its magnitude.

In other words:

How uncertainty changes appears substantially more informative than how uncertain the model is at any individual point in time.

This observation has important implications for future reliability research. It suggests that hallucination detection systems may benefit from modelling uncertainty as a temporal process rather than reducing it to aggregate summary statistics.

More broadly, the results indicate that uncertainty dynamics constitute a richer and potentially more informative representation of model behaviour than conventional confidence-based metrics.

Entropy drift therefore emerges not merely as another uncertainty feature, but as evidence that uncertainty evolution itself may represent a fundamentally important dimension of LLM reliability.

8.5 Summary of Findings

Among all evaluated uncertainty-based features, entropy drift demonstrated the strongest discriminative capability.

The analysis revealed:

- Positive entropy drift for most factual generations.
- Negative entropy drift for most hallucinated generations.
- A clear distributional shift between classes.
- A distinctive early-to-late uncertainty crossing pattern.
- Evidence that uncertainty dynamics outperform uncertainty magnitude.

Most importantly, entropy drift represents the first uncertainty-derived feature in this study to exhibit substantial predictive potential.

The next section evaluates this potential quantitatively through cross-validated predictive modeling and ROC-AUC analysis.

9. Predictive Evaluation

The preceding analyses established that entropy drift exhibits a clear distributional distinction between factual and hallucinated generations. However, visual separation alone does not guarantee predictive utility. Consequently, a formal evaluation was conducted to determine whether uncertainty-derived features can reliably discriminate between correct and incorrect generations.

To assess predictive performance, individual classifiers were trained using each uncertainty feature independently. This approach was intentionally chosen to isolate the contribution of each feature and prevent interactions between variables from obscuring their individual predictive value.

Performance was evaluated using Receiver Operating Characteristic – Area Under the Curve (ROC-AUC), a threshold-independent metric widely used in binary classification tasks. ROC-AUC measures the probability that a classifier ranks a randomly selected positive example higher than a randomly selected negative example, making it particularly suitable for evaluating uncertainty-based predictors.

To mitigate the effects of favourable train-test partitions and to obtain a more robust estimate of generalization performance, stratified k-fold cross-validation was employed.

9.1 Initial Predictive Assessment

An initial train-test evaluation was conducted using entropy drift as the sole predictive feature.

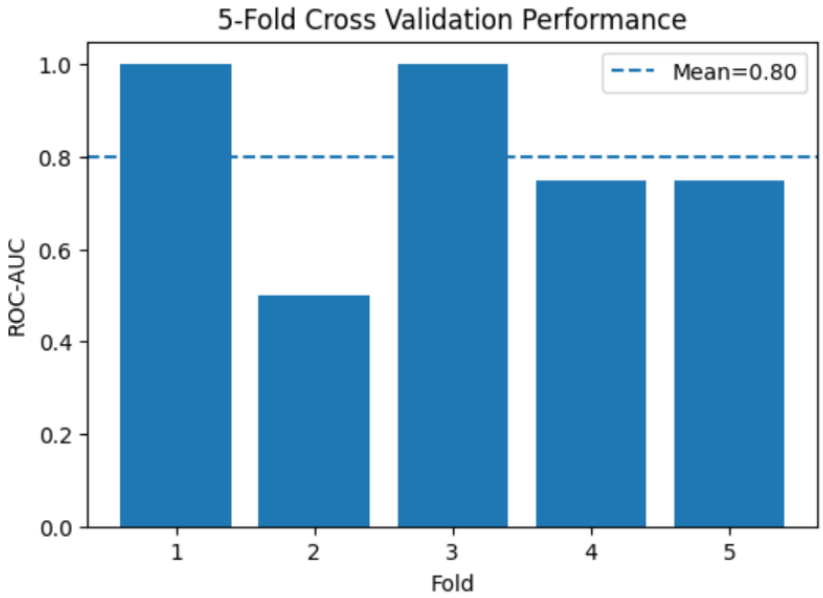

Figure 10. ROC curve for entropy drift under an initial train-test split.

The resulting ROC-AUC approached 1.0, indicating near-perfect separation on the evaluation split. While encouraging, such performance warrants caution because small datasets can occasionally produce optimistic estimates under favourable partitioning.

To determine whether the observed performance reflected a genuine signal rather than a dataset artifact, a more rigorous cross-validation procedure was subsequently performed.

9.2 Cross-Validated Feature Comparison

Five-fold stratified cross-validation was performed for each uncertainty-derived feature.

Table 1. Cross-validated ROC-AUC performance of uncertainty-derived features.

Parameters	Value
mean_entropy	0.41
entropy_variance	0.5
spike_rate	0.56
mean_jump	0.45
entropy_drift	0.79

Feature	Mean ROC-AUC
Mean Entropy	0.25
Entropy Variance	0.25
Spike Rate	0.275
Mean Jump Magnitude	0.35
Entropy Drift	0.80

The results reveal a substantial disparity in predictive performance across uncertainty metrics.

Mean entropy, entropy variance, and spike rate all perform near or below chance level. These findings are consistent with the earlier qualitative analyses, which demonstrated extensive overlap between factual and hallucinated generations across these metrics.

Entropy jump magnitude exhibits a modest improvement but remains insufficient for reliable discrimination.

In contrast, entropy drift substantially outperforms every alternative feature. The observed cross-validated ROC-AUC of approximately 0.80 indicates meaningful predictive capability despite the inherent difficulty of the task.

Importantly, this performance was achieved using a single feature derived exclusively from uncertainty dynamics.

9.3 Interpretation of Predictive Results

The predictive evaluation provides quantitative confirmation of the central empirical finding of this study.
The failure of static uncertainty metrics suggests that hallucinations cannot be reliably identified through uncertainty magnitude alone. Incorrect generations are frequently associated with uncertainty values comparable to those observed in correct responses.
Entropy drift, however, captures an entirely different aspect of model behaviour.
Rather than measuring how uncertain a model is, entropy drift measures how uncertainty evolves throughout generation.
The superior performance of entropy drift therefore provides evidence that temporal uncertainty dynamics contain substantially richer information than aggregate uncertainty statistics.
This distinction is conceptually important.
Traditional uncertainty estimation assumes that hallucinations should be characterized by elevated uncertainty. The present results suggest a different interpretation: hallucinations may instead be characterized by a distinctive trajectory of confidence accumulation during generation.
From this perspective, uncertainty evolution emerges as a potentially fundamental dimension of model reliability.

10. Discussion

The results of this study reveal a striking contrast between static uncertainty measures and uncertainty dynamics.
Across multiple analyses, conventional uncertainty metrics consistently failed to distinguish factual generations from hallucinations. Mean entropy trajectories, entropy variance distributions, spike frequencies, and entropy jumps all exhibited extensive overlap between classes. These findings indicate that uncertainty magnitude alone provides limited information regarding factual correctness.
Entropy drift represents a notable exception.
The emergence of entropy drift as the dominant predictive signal suggests that hallucination behaviour is fundamentally dynamic rather than static. Correct and incorrect generations are not separated by how uncertain they are, but by how their uncertainty changes over time.
This observation has broader implications for uncertainty quantification research. Much of the existing literature focuses on aggregate uncertainty statistics, confidence calibration, or single-point estimates of uncertainty. The findings presented here suggest that such approaches may overlook important temporal structure embedded within the generation process.
More generally, the study demonstrates the value of treating uncertainty as a trajectory rather than a scalar quantity.
The results therefore support a shift toward dynamic uncertainty modelling, trajectory-based reliability assessment, and temporally-aware hallucination detection methodologies.

11. Limitations

Several limitations should be acknowledged.
First, token-level analyses were conducted using a single open-weight language model. Although the resulting uncertainty patterns are informative, additional validation across model architectures and scales is necessary to assess generality.
Second, the dataset size remains relatively modest. While cross-validation provides evidence that the observed signal is genuine, larger datasets would enable more precise estimation of predictive performance.
Finally, the present study establishes predictive associations rather than causal mechanisms.

12. Future Work

Several promising research directions emerge from the findings.
Future studies may investigate:

- Cross-model validation across proprietary and open-weight architectures.
- Larger benchmark datasets with diverse hallucination categories.
- Sampling-based decoding and stochastic generation regimes.
- Richer uncertainty trajectory representations beyond entropy drift.
- Mechanistic investigations into the causes of negative entropy drift.
- Integration of uncertainty dynamics into reliability-aware decoding strategies.

Such extensions would contribute toward a more comprehensive understanding of the relationship between uncertainty evolution and factual reliability.

13. Conclusion

This study investigated whether token-level uncertainty dynamics can predict hallucinations in Large Language Models.

The research began with a calibration analysis of GPT-4o and Gemini 3.5, revealing substantial overconfidence despite strong predictive performance. These findings motivated a transition from response-level confidence estimation to token-level uncertainty modelling.

Subsequent analyses examined multiple uncertainty-derived features, including mean entropy, entropy variance, spike frequency, entropy jump magnitude, and entropy drift. Contrary to conventional expectations, static uncertainty measures exhibited little predictive utility and failed to reliably distinguish factual generations from hallucinations.

Entropy drift emerged as the sole feature demonstrating strong discriminative capability. Correct generations consistently exhibited positive entropy drift, whereas hallucinated generations displayed negative entropy drift. Cross-validated evaluation confirmed the predictive significance of this phenomenon, with entropy drift achieving a ROC-AUC of approximately 0.80 and substantially outperforming all alternative uncertainty metrics.

The central finding of this work is therefore not that hallucinations are more uncertain, but that they evolve differently.

More specifically, the results indicate that uncertainty dynamics contain substantially richer information than uncertainty magnitude alone. This observation suggests that future reliability estimation and hallucination detection systems may benefit from modelling the temporal evolution of uncertainty rather than relying exclusively on aggregate confidence measures.

In summary, the study provides empirical evidence that uncertainty trajectories constitute a meaningful and previously underexplored signal for hallucination prediction. By demonstrating the predictive value of entropy drift, this work contributes toward a deeper understanding of uncertainty in autoregressive language models and highlights uncertainty dynamics as a promising direction for future research in LLM reliability, interpretability, and trustworthy AI.