BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Predictive Cognitive Failure Learning (PCFL): A Pattern Recognition Framework for Early Hallucination Forecasting and Adaptive Reasoning in Large Language Models

Sagar Baliram Ahirrao

Department of Artificial Intelligence and Data Science
Ajeenkya DY Patil School of Engineering, Pune, India
Email : sagar.ahirrao@dypic.in

Dr. Vishal Puranik

Department of Artificial Intelligence and Data Science
Ajeenkya DY Patil School of Engineering, Pune, India
Email : vishal.puranik@dypic.in

Abstract

When errors occur in the form of hallucinations or factual distortions in a Large Language Model (LLM), they are exacerbated by the fluency of the model's output. Error correction techniques are generally post-hoc and result in delays to users. This paper analyses ten different studies including surveys, repeated sampling, in-context validation of dubious tokens, confidence calibration, probing and document retrieval. A gap is identified in the field where risk estimation can be performed during decoding using purely model generated signals. As a solution to this gap, we describe Predictive Cognitive Failure Learning (PCFL), where token probabilities, attention and drift between reasoning steps are used as signals to predict failure using a compact risk estimation module. This risk score is then used by a controller to decide to continue, recheck, rewrite or alert the user to the text. This is a design study and PCFL has not been implemented or assessed. Therefore, we describe how we will develop a prototype and assessment plan for future work.

Keywords : Adaptive reasoning, confidence calibration, hallucination detection, large language models, pattern recognition, trustworthy AI, uncertainty estimation.

I. Introduction

Artificial Intelligence (AI) tools such as GPT, Claude and LLaMA have become commonly used tools for writing, researching and coding. However, these tools can often provide inaccurate results. For example, they may reference non-existent studies, provide incorrect statistics or provide contradictory

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

information to the text they are asked to summarise. This can cause issues for clinicians, lawyers, teachers, and analysts who rely on the results.

These issues can generally be split into two categories [1]–[3]. The first category refers to the text being factually incorrect and the second category refers to the text being inconsistent with the information given to the model or previous sentences. Some reasons for this could be due to the training data being sparse/noisy, models often being trained in a way that can cause inconsistencies when deployed, the models having knowledge limitations and decoding parameters that favour confident language [1].

The defensive tools created so far mostly work in a reactive way: they compare a finished answer with external resources, bring in supporting documents, or look for agreement among several sampled answers. These approaches help, but they come after the damage is done and can slow the system down.

We take a different perspective and ask whether the model's behaviour *while writing* shows detectable signs that something is going wrong.

We make three contributions. First, we categorise ten representative studies by the stage their methods operate at. Second, we explicitly state what capability none of them have. Third, we provide an outline of a conceptual framework and a plan for this to be tested in future experimental work. Section II introduces the problem. Section III outlines the objectives. Section IV discusses current trends. Section V covers a literature review. Section VI outlines the gaps. Section VII provides the framework. Section VIII concludes.

II. Problem Statement and Motivation

There are three issues with current approaches. First, post-hoc checks are only available after text has been generated, and may involve further generations or calls. Second, while calibration work provides information about a model's confidence, this is rarely used to provide a clear signal during decoding. Third, while retrieval provides evidence, if the error is due to the model's reasoning then this approach provides nothing.

There is good reason to believe a better path exists. Internal activations carry information about whether a statement is true [4], and on that basis a model can evaluate its own answers to a useful extent [5]. We therefore ask: *how can the probability of a hallucination be estimated from an LLM's internal behaviour during generation, and how can that estimate be used to alter the model's reasoning before an unreliable answer is shown?*

III. Objectives of the Review

The aim of this review is:

a) to list the kinds and sources of hallucination and the present ways of identifying and controlling it;

b) to examine internal cues, namely token confidence, attention behaviour and reasoning consistency, as possible precursors of failure;

c) to explain PCFL as a conceptual framework that rates risk at the time of generation; and

d) to evaluate whether adapting reasoning on the fly, in response to that rating, is a realistic proposition.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

IV. Relevance to Current Technology Trends

A. Trustworthy AI

Overall, hallucination is one of the more notable outstanding issues in the field [6]. As LLMs are used more in high-stakes regulated domains, showing they can be trusted is now a core requirement.

B. Uncertainty and Calibration

Research on understanding model uncertainty [5] and the encoding of truth related information within the hidden layers [4] demonstrates that uncertainty is encoded within the network. This can be used to benefit methods which aim to predict error.

C. Efficient Real-Time Inference

Chat assistants and enterprise search systems require fast responses. Therefore methods requiring multiple extra generations [7] become expensive. This has led to an increasing interest in tiny monitors that can be run alongside a single generation.

D. Retrieval-Augmented Systems

Retrieval-augmented generation [8] is now popular in industry. It supplies missing knowledge but cannot repair a faulty argument, so it complements, rather than replaces, monitoring of the generation process itself.

V. Literature Survey

A. Taxonomy and Origins of Hallucination

As well as the factual versus faithful distinction [1]–[3], we can make a distinction between knowledge and reasoning errors. Knowledge errors occur when the model has not been taught the fact, and retrieval could be used to address these. Reasoning errors occur when the model has access to the fact, but fails during the sequence of reasoning steps. Monitoring internal signals could help with both types of error, hence this review will focus on reasoning errors.

Reasons for errors have been grouped into three categories [1]. Data errors can result from missing, noisy or outdated data in the training corpus. Training errors can occur due to the difference between training with always having correct previous tokens and during inference having to continue from the model's output. This can cause small errors to build up. Inference errors can occur because the decoding choices will often prefer fluent and likely text. Benchmark experiments [9] have shown that this issue occurs in reasoning tasks and in multiple languages.

B. Overview of Approaches

Existing work belongs to three families: examining finished text, estimating confidence (from probabilities, entropy, self-assessment and hidden-state probes), and grounding answers in retrieved documents. A set of survey and benchmark papers provides shared vocabulary and metrics. The ten studies are summarised in Table I, and Fig. 1 marks the stage at which each family acts.

Overall, the findings indicate four aspects. While the post-editing approach can be applied in all models, it comes at a later stage and incurs a cost [7]. Whilst it has been shown that the required

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

information can be obtained by calibration and interrogation [4], [5], it is rarely used during decoding. The closest approach to being able to do this during generation [10] still relies on external tools and incurs additional costs. Knowledge can be recovered using retrieval [8]. However, it does not allow for correcting faulty reasoning. Finally, whilst the survey and benchmark papers [1]–[3], [6], [9] provide an overview of the landscape, they do not suggest a predictive mechanism.

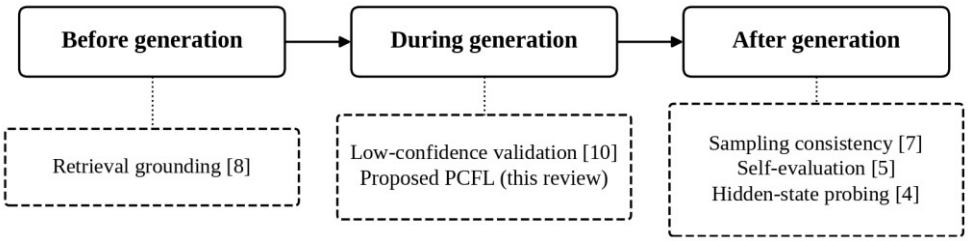


Fig. 1. Position of each reviewed approach in the life of a response. PCFL, in the middle column, is this paper's conceptual proposal.

TABLE I

LITERATURE SURVEY SUMMARY

Ref.	Author / Year	Methodology Used	Key Findings	Limitations
[1]	Huang et al., 2024	Classification scheme and root-cause study	Ties hallucination to data, inference behaviour and knowledge boundaries	Wide scope; no combined real-time predictive remedy
[2]	Zhang et al., 2023	Survey that sorts hallucination into types	Separates factuality from faithfulness errors across tasks	Descriptive; no scheme for detection or prevention
[3]	Ji et al., 2023	Systematic literature review	Offers shared definitions and measures for generation errors	Metric-centred; little on early or predictive detection
[4]	Azaria and Mitchell, 2023	Classifier trained on hidden activations	Hidden states reveal whether a statement is likely true	Requires internal access; mostly tested on factual statements
[5]	Kadavath et al., 2022	Probing of self-assessment and calibration	Models can partly estimate whether their answers are right	Quality differs by task; not linked to forecasting while decoding
[6]	Kaddour et al., 2023	Broad review of open LLM problems	Treats hallucination as one of many reliability concerns	Too general; limited depth on hallucination
[7]	Manakul et al., 2023	Agreement test over sampled outputs (SelfCheckGPT)	Marks unreliable text by comparing samples, no external data	Applied afterwards; repeated sampling is time and compute consuming

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Ref.	Author / Year	Methodology Used	Key Findings	Limitations
[8]	Lewis et al., 2020	Retrieval-augmented generation (RAG)	Fetched documents ground answers and cut factual errors	Bounded by retrieval quality; reasoning faults remain
[9]	Bang et al., 2023	Benchmark experiments across several tasks	Reports how often hallucination occurs in reasoning and multilingual use	Assessment only; no remedy or forecasting idea
[10]	Varshney et al., 2023	Checking of doubtful tokens while text is written	Repairs probable errors early so they do not propagate	Depends on external validators; extra run-time load

C. Comparative Analysis

Table II outlines the primary methods available, based on five criteria. We have also included a row for the proposed PCFL method to show where we would expect it to fit. There appears to be a balance between methods with no inside knowledge (such as sampling consistency) which are easily transportable but take longer and are more expensive to implement. Conversely, methods using inside information are much cheaper to implement but rarely used for forecasting.

D. Evaluation Measures Used in This Area

A set of survey and benchmark papers provides shared vocabulary and metrics for this area. Among them, [3] offers shared definitions and measures for generation errors, and [9] reports how often hallucination occurs in reasoning and multilingual use. The evaluation plan in Section VII-C additionally refers to lead time and latency when PCFL is compared with sampling consistency and token validation, and to ablations that remove one signal group at a time.

VI. Research Gap

There are four issues that we have observed from this review:

a) No forecasting capability. None of the methods were able to flag hallucinations at the point of generation based solely on internal model factors.

b) No confidence indicators. While probabilities, entropy and subjective confidence measures are calculated, they are rarely used as predictive indicators [4], [5].

c) Dependence on external factors. The most advanced method at the point of generation required external tools [10] meaning extra time was needed to perform inference. Sampling methods also had the issue of requiring multiple executions [7].

d) Internal errors. Errors arising from faulty reasoning within the model cannot be corrected by retrieval [8]. Very few works track the consistency between steps of reasoning.

Overall the main issue is that there is no lightweight, pattern-based method which can predict risk during generation based on internal patterns of behaviour that can allow the model to respond.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

VII. Proposed Conceptual Framework

The proposed PCFL (see Fig. 2) comprises a signal extractor module which observes the decoder. It provides three kinds of cues from the signal. A failure pattern memory stores past generation trajectories that resulted in an error. These are processed by a small risk model which combines them with the cues to produce a risk score. This is continually updated during decoding and finally an adaptive controller performs one of four actions, namely continuing, rechecking the current step, rewriting a range of text, or inserting an uncertainty marker.

Let the decoder emit token y_t from $p_t(\cdot) = P(\cdot \mid y_{<t}, x)$. Two example token features are the entropy (1) and the top-two margin (2), where V is the vocabulary and $v^{(1)}$, $v^{(2)}$ are the two most probable tokens at step t:

$$H_t = -\sum_{v \in V} p_t(v) \log p_t(v) \qquad (1)$$

$$m_t = p_t(v^{(1)}) - p_t(v^{(2)}) \qquad (2)$$

Joining these with attention dispersion, step divergence and similarity to stored failures into a feature vector f_t, the risk score (3) is

$$R_t = \sigma(\mathbf{w}^{\mathrm{T}} f_t + b) \qquad (3)$$

with parameters learned from labelled generations by binary cross-entropy. Two thresholds $\tau_1 < \tau_2$ define the policy: continue when $R_t < \tau_1$, pause and re-check when $\tau_1 \le R_t < \tau_2$, and rewrite or flag when $R_t \ge \tau_2$. Answers judged low-risk therefore pass without added delay.

A. Signal Groups

Token-level signals indicate uncertainty. If a token-level signal suggests uniformity or a low distinction between options then it indicates the model is guessing. Focus-level signals indicate grounding. If the focus shows little focus or erratic focus then it indicates the model is diverging from its source. Reasoning-level signals indicate coherence. If there is an abrupt transition between reasoning steps then it indicates a lack of coherence. These signals individually are weak but PCFL works on the premise that when taken together they make a temporal pattern which can give more information than any individual interpretation.

B. Adaptive Reasoning Behaviour

In low-risk situations nothing is done. In medium cases, we might re-derive the current step of the model (by re-generating, for example) and the extra cost is only incurred on the ambiguous parts. For high risk scenarios we might re-write the span, check against retrieval if this is available, or mark the passage as ambiguous. As only a minority of steps will get to the higher levels of the system we expect the average overhead to be small.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE II

COMPARISON OF HALLUCINATION-HANDLING APPROACHES

Approach	Acts	Signal used	Strength	Weakness
Sampling consistency [7]	After generation	Agreement between samples	Portable; needs no outside data	Delay and compute from repeated runs
Token validation [10]	During generation	Low-confidence tokens plus checking	Limits spread of errors	Needs external validators
Hidden-state probing [4]	After or during	Internal activations	Informative about truthfulness	Needs internal access; narrow tests
Self-evaluation [5]	After generation	The model's own estimate	Simple; partly calibrated	Quality varies by task
Retrieval grounding [8]	Before generation	Retrieved documents	Supplies evidence	Cannot repair reasoning faults
PCFL (proposed)	During generation	Token, attention, reasoning cues and failure memory	Looks ahead; self-contained; supports adaptive response	Conceptual; untested

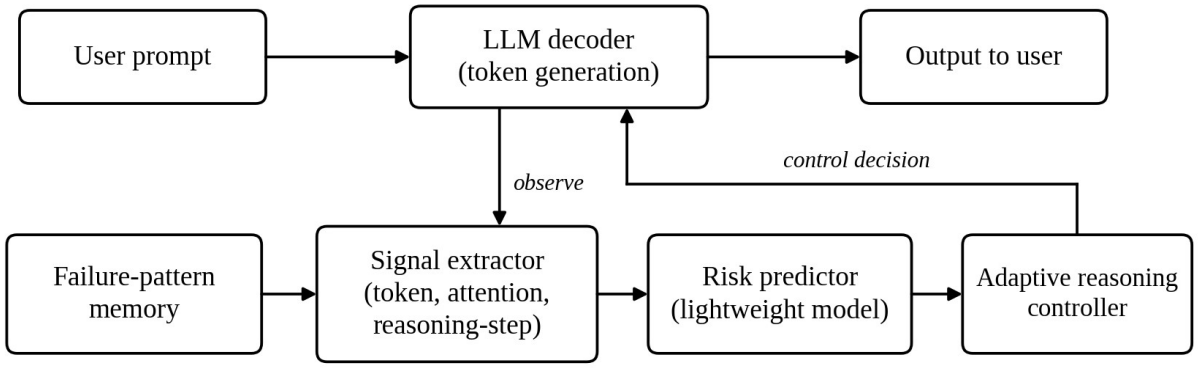


Fig. 2. Conceptual PCFL framework. A monitor observes the decoder, scores failure risk at each step, and returns a control decision.

C. Proposed Evaluation Plan

There could be four steps to running a follow-up study: 1) Collect responses from an open source question answering and summarization model and label them as faithful or hallucinated. 2) Track the three signals during decoding and use some of the labelled dataset to train the risk model. 3) Compare the performance of the model against held out data with sampling consistency and token verification using the metrics described above. 4) Test for transferable learning by applying the trained model to a different domain and language. Using ablation studies to remove one signal at a time can identify which signal contains the most predictive power.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

D. Potential Applications

If this is accurate it can be applied to many real-world applications such as clinical information assistants, legal drafting assistants, tutoring, financial reporting and enterprise search. This will ensure that alerts and additional checking will only occur in high risk areas.

E. Ethical and Practical Considerations

The framework also includes transparency and privacy. Thresholds must be logged and auditable, the memory must not lead to biases being present which could mask errors in certain languages or topics and prompts must be logged in accordance with data protection laws. This is a conceptual framework and is designed to provide guidance for future experiments. The main limitations of this framework are ensuring that the internal states can be accessed (which may not be available in closed models), the time it takes to gather labelled failure history data and careful threshold tuning.

VIII. Conclusion

We surveyed 10 papers addressing various aspects of LLM hallucination, including taxonomies, post-hallucination detection, calibration, hidden state analysis, token-level verification and retrieval grounding. Existing approaches are either post-hoc, require external information or fail to provide a prediction from internal signals. One possible concept-based solution, PCFL, is described. It combines token, attention mechanism and reasoning clues with a failure memory and adaptive control layer. We will implement a prototype on open source models where we have access to their internals. We will compare it against the baselines (sampling and token verification) and investigate how different clues benefit different tasks and languages. Finally, we will seek to reduce monitoring cost, combine it with retrieval grounding and also generalise to multimodal systems.

References

[1] L. Huang et al., “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” *ACM Trans. Inf. Syst.*, Nov. 2024, doi: 10.1145/3703155.

[2] Y. Zhang et al., “Siren’s song in the AI ocean: A survey on hallucination in large language models,” *arXiv preprint arXiv:2309.01219*, 2023.

[3] Z. Ji et al., “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.

[4] A. Azaria and T. Mitchell, “The internal state of an LLM knows when it’s lying,” in *Findings Assoc. Comput. Linguistics: EMNLP 2023*, Singapore, Dec. 2023, pp. 967–976.

[5] S. Kadavath et al., “Language models (mostly) know what they know,” *arXiv preprint arXiv:2207.05221*, 2022.

[6] J. Kaddour et al., “Challenges and applications of large language models,” *arXiv preprint arXiv:2307.10169*, 2023.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

[7] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Singapore, Dec. 2023, pp. 9004–9017.

[8] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.

[9] Y. Bang et al., "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," in *Proc. Int. Joint Conf. Natural Lang. Process. Asia-Pacific Chapter Assoc. Comput. Linguistics (IJCNLP-AACL)*, Nusa Dua, Bali, Nov. 2023, pp. 675–718.

[10] N. Varshney et al., "A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation," *arXiv preprint arXiv:2307.03987*, 2023.