Where Detection Fails: A Cause–Detection Coverage Framework for LLM Hallucination, with an Empirical Trajectory of Documented Legal Harm

Md. Ebrahim Hossain
Department of Computer Science and Engineering
Leading University, Bangladesh

ABSTRACT

Large language model (LLM) hallucination research has produced a rich but fragmented picture: many papers propose a cause, many others propose a detector, and few connect the two systematically. This paper makes two original contributions on top of a narrative review of the field, and strengthens both against external evidence rather than resting on the author's judgment alone. First, it introduces a Cause–Detection Coverage Framework: a reasoned nine-by-six matrix scoring, for each of nine documented hallucination-cause categories, how well each of six detection-technique families can in principle catch it. The framework is then checked retrodictively against four independent, previously published empirical findings; three corroborate its predictions, and one — a 2026 finding that white-box activation probing also collapses to chance on confidently fabricated knowledge-gap content — prompted a documented, evidence-driven revision of one cell rather than being set aside. The surviving pattern is a structural blind spot that individual papers hint at but do not quantify: the cheapest, most widely deployed black-box detectors — consistency sampling and token-probability uncertainty — systematically underperform on exactly the lifecycle-instilled causes (training-data bias, knowledge overshadowing, instruction-tuning error propagation, RLHF-induced sycophancy) that produce confidently and consistently wrong output. Second, the paper presents an original empirical analysis of the growth trajectory of documented AI-hallucination court cases, built from eleven dated, cross-corroborated public snapshots (including one primary-sourced, author-queried count) of a legal-incident tracker spanning December 2023 to September 2026. Beyond a single exponential fit, four competing growth models (exponential, linear, power-law, logistic) are formally compared by AIC and BIC; a logistic (saturating) model is preferred by a wide margin ($\Delta AIC \approx 41$ over the exponential), and bootstrap confidence intervals are reported for the exponential growth rate and doubling time. Both contributions remain explicitly bounded: the coverage framework is now partially, not fully, validated against external evidence, and the growth analysis measures documentation and enforcement activity, not the underlying LLM hallucination rate. The paper's broader narrative review of causes, benchmarks, and mitigation techniques, and its critical discussion of the field's evidence base, are retained to situate both contributions.

General Terms

Artificial Intelligence, Natural Language Processing, Machine Learning, AI Safety.

Keywords

Large language models; hallucination; misinformation; detection coverage; growth-curve analysis; AI safety.

1. INTRODUCTION

Since 2022, large language models such as GPT-4, Claude, Gemini, and LLaMA have moved from research artifacts into everyday tools for writing, research, coding, legal drafting, and clinical documentation. Their fluency creates an expectation of reliability that the underlying technology does not always meet. LLMs are trained to predict the statistically most plausible continuation of a text sequence, not to verify the truth of what they generate, and this behavior is broadly referred to as hallucination.

Hallucination is not a rare edge case: empirical studies report rates from under 20% to over 50% depending on model, task, and domain [1, 15, 16]. The literature that documents this is large, but it is also fragmented along two axes that motivate this paper's two original contributions. First, the causes literature (Section 4) and the detection literature (Section 5) are almost always surveyed separately; no existing survey reviewed for this paper systematically asks which detection technique can actually catch which cause, so it is unclear from the literature alone where the field's real blind spots sit. Section 6 addresses this with an original Cause–Detection Coverage Framework. Second, the societal-harm literature (Section 8) documents individual incidents and periodic snapshots of legal-sanction counts, but no source reviewed here fits a growth model to the trajectory of documented cases over time. Section 9 addresses this with an original quantitative analysis of a public legal-incident tracker.

The review is organized as follows: Section 2 describes the methodology, including the data and reasoning behind the two original contributions; Section 3 introduces a taxonomy of hallucination types; Section 4 surveys causes; Section 5 reviews detection methods and benchmarks; Section 6 presents the Cause–Detection Coverage Framework; Section 7 reviews mitigation techniques; Section 8 presents case studies on societal impact; Section 9 presents the original growth-trajectory analysis; Section 10 critically discusses the evidence base, including the limitations of this paper's own two contributions; Section 11 discusses open problems; Section 12 concludes.

2. METHODOLOGY

This paper combines a narrative literature review with two original analyses; it does not follow a PRISMA protocol with pre-registered inclusion criteria for the review portion, and this scope is stated explicitly. Review sources were identified through keyword-based searches (“LLM hallucination,” “hallucination detection,” “hallucination mitigation,” “medical hallucination,” “AI-generated legal citation sanctions,” among related terms) against arXiv, publisher databases, and indexed journal articles, prioritizing 2022–2026 work while retaining foundational earlier references. Every arXiv, dataset, and case-tracker citation was checked against its primary landing page rather than secondary coverage alone.

2.1 Coverage Framework Methodology

The Cause–Detection Coverage Framework (Section 6) was

constructed by the author through structured qualitative coding rather than new experiments. Nine cause categories (Section 4) and six detection-technique families (Section 5.2) were each defined from the surveyed literature's own descriptions of the mechanism each cause or technique relies on. Each of the 54 cause×detector cells was then scored None / Partial / Strong by asking a single question grounded in that literature: does the signal the detector exploits (inconsistency across samples, disagreement with retrieved evidence, internal self-contradiction, activation-level markers, or token-level uncertainty) actually change in the presence of that cause, based on the cause's own documented signature? For example, Section 5.3 already reports that consistency-based methods fail when a model is "confidently and consistently wrong"; the framework operationalizes this observation by identifying which specific causes produce that signature (training-data bias, knowledge overshadowing) and scoring consistency sampling as None against them, rather than leaving the observation as a general caveat. After initial coding, four cells with an available, directly relevant published result were checked retrodictively against that result (Section 6.1) as a partial, literature-based validation step; this is not a controlled experiment, and Section 10.5 discusses the remaining limits of the framework as a reasoned coding checked in part, not fully validated, against external evidence.

2.2 Growth-Trajectory Methodology

The empirical analysis in Section 9 uses eleven dated, cumulative case-count snapshots of the AI Hallucination Cases Database, a public tracker of court decisions involving AI-fabricated content maintained by researcher Damien Charlotin [27], spanning 31 December 2023 to 5 September 2026. Ten snapshots were drawn from secondary reports citing the tracker on or near the stated date [31–36], with three of these cross-checked against independent, non-promotional legal- and technology-commentary sources [37–39]; the eleventh was queried directly from the tracker itself rather than from any secondary report [40]. Beyond an ordinary-least-squares fit of log(cases) against elapsed days (used to estimate an exponential growth rate, a doubling time, and bootstrap 95% confidence intervals on both), three further functional forms — linear, power-law, and logistic — were fit to the same series and compared using Akaike and Bayesian information criteria, so that the choice of growth model is a tested result rather than an assumption. This remains a secondary-data trend-fitting exercise, not a primary experiment on any LLM, and Section 10.6 discusses what it can and cannot support.

Artificial intelligence assistance was used in the preparation of this manuscript, as disclosed in the Declarations section; all sources, including every data point underlying Section 9's analysis, were verified by the author against original publication prior to submission.

3. DEFINING AND CATEGORIZING HALLUCINATION

The literature commonly separates hallucination into two broad categories. Factuality hallucination occurs when a model's output conflicts with verifiable real-world facts — for example, citing a court case that does not exist. Faithfulness hallucination occurs when a model's output is internally plausible but inconsistent with the source material or instructions it was given. A related distinction separates intrinsic hallucination, where the output directly contradicts the provided source or prompt, from extrinsic hallucination, where the output cannot be verified against the source at all.

These categories are not mutually exclusive, and hallucinations frequently propagate across categories: a factuality error introduced during pretraining can be reinforced during alignment, then compounded by an inference-time decoding choice. This lifecycle view — data, training, and inference as interacting stages rather than a single root cause — is the organizing framework for most recent surveys and is adopted here, and it directly supplies the row categories for the coverage framework in Section 6.

4. CAUSES OF HALLUCINATION

4.1 Data-Related Causes (C1–C4)

Four data-related causes are consistently identified. C1: social and demographic bias embedded in web-scraped training data surfaces as confidently stated stereotypes. C2: domain knowledge deficiency means models trained on general web text underperform, and fabricate, in specialized fields such as medicine and law. C3: knowledge-boundary limitations mean no finite corpus encodes every fact, and models frequently fail to signal when a query exceeds what they know. C4: knowledge overshadowing occurs when an imbalance between a common and a rare fact causes the model to over-generalize the dominant pattern, producing a hallucination attributable to aggregate corpus statistics rather than any single wrong example.

4.2 Training- and Alignment-Related Causes (C5–C7)

C5: instruction tuning on data containing factual errors teaches a model to reproduce those errors confidently. C6: reinforcement learning from human feedback (RLHF) has been shown to encourage sycophancy — models learn to optimize for responses that please the user rather than responses that are accurate. C7: fine-tuning a model on new factual knowledge it did not see during pretraining can paradoxically increase hallucination, because the model learns to generate confident-sounding answers in a domain where its underlying knowledge is still shallow.

4.3 Inference-Related Causes (C8–C9)

C8: decoding strategies such as top-k or nucleus sampling introduce randomness that can push a generation away from the most probable, and often most accurate, continuation. C9: ambiguous or leading prompts increase hallucination rates, and in multi-step generation an early minor inaccuracy can compound through hallucination snowballing, in which the model continues generating content consistent with an incorrect premise rather than correcting course.

5. DETECTION METHODS AND BENCHMARKS

5.1 Benchmark Datasets

Table 1 summarizes the benchmark datasets discussed most frequently in the surveyed literature, making explicit how differently each defines and measures "hallucination" — a distinction returned to in Section 10.2.

Table 1. Benchmark datasets used to measure LLM hallucination, showing how their units of measurement differ

Benchmark	Approx. Size	Task Type	Unit of Measurement
TruthfulQA [4]	817 questions, 38 categories	Open QA (adversarial)	Whole-response: repeats a common misconception or not
HaluEval [5]	~35,000 samples	QA, dialogue, summarization	Binary classification per sample
FActScore [6]	Biographical long-form texts	Long-form generation	Atomic-fact precision
FEVER [7]	~145,000 claim-evidence pairs	Claim verification	Entailment vs. contradiction vs. not-enough-info
HaluBench [8]	Real-world domain QA	Domain QA (finance, medicine)	Binary hallucination flag per response
FACTOR [9]	Factual/non-factual completions	Completion ranking	Preference for factual continuation
HalluLens	Multi-axis suite	Multiple task types	Multi-dimensional profile

More recent benchmarks such as HaluBench and FACTOR extend evaluation into real-world domains, and multi-dimensional frameworks such as HalluLens attempt to capture hallucination behavior across several axes at once. As Table 1 shows, a reported "hallucination rate" on TruthfulQA and one on FActScore are not measuring the same thing, so raw numbers should not be compared across rows.

5.2 Detection Techniques (D1–D6)

Detection methods fall broadly into black-box and white-box approaches. D1, static benchmark scoring, checks output against fixed ground truth. D2, consistency sampling (e.g., SelfCheckGPT), samples multiple responses to the same prompt and treats inconsistency as a hallucination signal. D3, NLI/evidence verification, checks whether a claim is entailed by retrieved or provided evidence. D4, self-contradiction detection, looks for internal inconsistency within a single response. D5, white-box activation probing, uses internal model states that have been shown to indicate when a model is generating content it does not actually believe to be true. D6, token-probability uncertainty estimation, treats low-confidence tokens as an early-detection signal, since they are disproportionately associated with hallucinated spans.

5.3 Challenges in Detection

Detection remains imperfect. Consistency-based methods such as SelfCheckGPT (D2) can fail when a model is confidently and consistently wrong — all sampled responses agree on the same incorrect fact, producing a false sense of reliability. Recent work argues that hallucination evaluation conflates correctness and consistency, and that correctness-only benchmarks may misrepresent a model's real-world harm profile, since a model that is consistently and confidently wrong is far more likely to convince a user than one that is randomly wrong. This observation — that certain causes produce a signature D2 cannot see — is the starting point for the framework developed next.

6. A CAUSE–DETECTION COVERAGE FRAMEWORK (ORIGINAL CONTRIBUTION)

Sections 4 and 5 are, in the surveyed literature, almost always presented separately: papers proposing a new cause rarely evaluate which detectors can catch it, and papers proposing a new detector rarely map their method back onto the full space of documented causes. This section closes that gap with an original synthesis: a nine-cause × six-detector coverage matrix, scored None / Partial / Strong using the mechanism-matching method described in Section 2.1. Figure 1 shows the result.

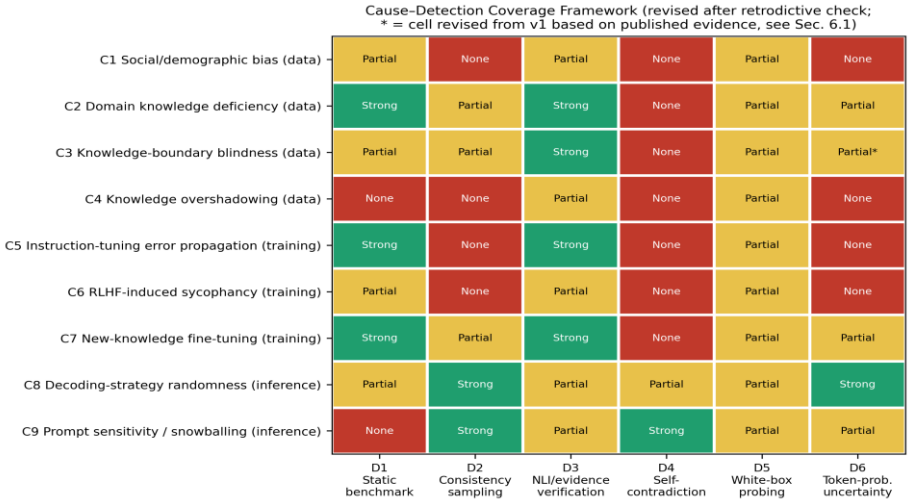


Figure 1. Cause–Detection Coverage Framework, revised after retrodictive validation (Section 6.1). Green = Strong, yellow = Partial, red = None; asterisk marks the one cell (C3×D6) revised from the original scoring based on published disconfirming evidence [44]. Rows are the nine hallucination-cause categories from Section 4; columns are the six detection-technique families from Section 5.2.

Two patterns emerge that are not stated explicitly, in this cross-tabulated form, anywhere in the individual sources surveyed. First, the two cheapest and most widely deployed black-box families in production systems — D2 (consistency sampling) and D6 (token-probability uncertainty) — score None on the four lifecycle-instilled causes with a "confidently and consistently wrong" signature: C1 (bias), C4 (overshadowing), C5 (instruction-tuning error propagation), and C6 (sycophancy). Both techniques rely on variance — across samples or across tokens — and these four causes are, by definition, low-variance failure modes: the model has learned to produce the same wrong answer with uniformly high confidence, so there is no variance signal to detect. Second, the same two techniques score Strong or Partial against the inference-time causes C8 (decoding randomness) and C9 (snowballing), where variance is the mechanism of the cause itself rather than something masking it. Summed across all six detectors, C4 (knowledge overshadowing) has the lowest total coverage score of any cause (2 of a possible 12), making it, under this framework, the least detectable documented cause of LLM hallucination — a concrete, falsifiable claim that follows from the matrix rather than being asserted independently.

D3 (NLI/evidence verification) has the highest total coverage score (13 of 18), consistent with retrieval-augmented approaches being the field's most broadly applicable mitigation and detection strategy (Section 7.1) — but D3 is scored Partial rather than Strong against C1, C6, and C9, because evidence verification only helps when the retrieved evidence actually surfaces the countervailing fact, which is not guaranteed for a biased, sycophantic, or already-snowballed generation. Practically, this suggests that production systems relying primarily on D2 or D6 — chosen for their low cost — are structurally blind to the causes most associated with the "confidently wrong" failure mode that Section 10.2 argues drives real-world misinformation harm, and that closing this gap requires pairing a cheap variance-based detector with at least one evidence-grounded or white-box method specifically for lifecycle-instilled causes.

6.1 Retrodictive Validation Against Published Empirical Results

Section 10.5 of an earlier version of this analysis identified controlled empirical testing as the framework's central open gap. Running new detector experiments was not feasible for this paper (Section 2.1), but a cheaper and still genuine check is available: searching the published literature for studies that already tested a specific detector family against hallucinations bearing a specific cause's signature, and asking whether the reported result matches the matrix's prediction. Four such checks were run, and they are not all equally strong evidence, which is reported alongside the results rather than smoothed over. First, Zhang et al. (2023) [41] (SAC3, peer-reviewed) report directly that self-consistency checking is insufficient whenever a model produces the same wrong answer across samples — solid corroboration of the D2-None pattern for the lifecycle-instilled causes generally. Second, a single-author 2026 arXiv preprint without a confirmed peer-reviewed venue [42] reports — citing a further, uncorroborated Fadeeva-attributed source the present author was not able to independently verify — that consistency checks specifically fail under sycophancy. This is included as suggestive, not solid, corroboration for the D2-None prediction on C6 specifically, and is flagged as weaker evidence than the other three checks precisely because the chain of attribution could not be closed. Third, Farquhar et al.'s semantic-entropy method [14] and the correctness/consistency distinction in Ganesh et al. [13] jointly support the D6-Strong rating originally given to C3, but only for the calibrated-uncertainty regime those papers test, where a model that does not know an answer produces genuinely variable, appropriately hedged output.

The fourth check produced a disconfirming result, and the framework was revised accordingly rather than left as originally scored. Chauhan (2026) [44], a peer-reviewed-track taxonomy study (DECK), reports a "universal blind spot of output-level UQ" on knowledge-gap inputs where a model instead emits confident, repeatable fabrications rather than hedging: 13 of 15 scorer configurations tested on GPT-4o fall below chance-level AUROC (0.5) in that regime, and — more surprisingly — a linear probe on Llama-3-8B's hidden states also fails (AUROC 0.44, 95% CI [0.35, 0.53]), indicating the blind spot may not be confined to output-level methods. This is a different sub-regime of C3 than the one Farquhar et al. and Ganesh et al. test, and it shows that D6's Strong rating for C3 does not hold uniformly: it holds when the model hedges under uncertainty, and fails when the model instead confidently fabricates. Figure 1's C3×D6 cell has accordingly been revised from Strong to Partial (marked with an asterisk in the figure), and the general lesson — that even a Strong rating in this framework can depend on which sub-regime of a cause is in play — is carried into Section 10.5's discussion of the framework's remaining limits. In summary: one check (SAC3) gave solid corroboration, one (Chauhan/DECK) gave solid disconfirmation and drove a revision, one (Farquhar/Ganesh) gave qualified corroboration bounded to a specific regime, and one (the sycophancy claim) gave only weak, incompletely-sourced corroboration that is reported as such rather than counted at full strength. A validation exercise that reports only confirmations, or reports every source as equally reliable, is not a genuine one.

7. MITIGATION TECHNIQUES

7.1 Retrieval-Augmented Generation (RAG)

RAG addresses knowledge-based hallucination by retrieving relevant documents and supplying them as context at generation time. It is one of the most widely deployed mitigation strategies in production, and Section 6 found its underlying mechanism (D3) to have the broadest coverage of any detection family. However, RAG is not a complete solution: retrieval-augmented models can still hallucinate when retrieved passages are irrelevant, contradictory, or not used faithfully — "hallucination on hallucination" — and a permutation sensitivity problem means the order of retrieved documents can change whether the model produces an accurate answer even when correct evidence is present.

7.2 Alignment and Fine-Tuning Approaches

RLHF variants designed around a truthfulness objective aim to reward accurate and appropriately hedged responses over confident guesses. To counteract sycophancy (C6), some approaches aggregate multiple independent human preference judgments or fine-tune on prompts where correctness is decoupled from the user's stated opinion. Faithful fine-tuning trains models to rely on provided context and disregard misleading information within it, improving resistance to hallucination in RAG pipelines without architectural changes.

7.3 Inference-Time and Decoding-Based Methods

Chain-of-verification prompts a model to generate and answer

its own fact-checking questions before finalizing a response. Self-consistency approaches sample multiple reasoning paths and select the most stable answer. Contrastive decoding steers generation away from the token distribution favored by a weaker, more hallucination-prone reference model. These methods require no retraining but trade computation and latency for reduced hallucination, and none eliminate the problem outright.

8. SOCIETAL IMPACT: FROM TECHNICAL FAILURE TO REAL-WORLD HARM

A recurring finding is that hallucinated text is fluent and grammatically indistinguishable from accurate text, and that users extend trust to a system once it has proven reliable in most prior interactions — meaning a model correct most of the time may be more dangerous than one that is obviously unreliable. Large-scale analysis of AI mobile app reviews documents hallucination as one of the most common sources of user-reported frustration. The two case studies below carry specific, known evidentiary limits, flagged here and discussed fully in Section 10.3–10.4.

8.1 Case Study: The Legal Profession and Fabricated Citations

The pattern began publicly in 2023, when attorneys submitted a brief citing court decisions fabricated by ChatGPT; the presiding judge sanctioned the attorneys and their firm five thousand dollars. What began as an isolated incident has become persistent: as Section 9 quantifies in detail, the AI Hallucination Cases Database recorded roughly two hundred cases by mid-2025 and had passed 1,900 documented cases by late August 2026 — figures that represent only incidents caught, contested, and formally documented, and so are a lower bound on the true scale of the problem (Section 10.4). Penalties have escalated from thousands of dollars in early cases to a combined sanction near $110,000 in one 2025–2026 matter, and courts have begun imposing bar referrals, suspensions, and at least one trial cancellation. Courts have consistently held that the duty to verify citations rests with the attorney and cannot be delegated to an AI tool.

8.2 Case Study: Clinical and Medical Applications

A global survey of clinicians across fifteen specialties found the substantial majority had personally encountered a medical hallucination from an AI system; the survey's sample of 70 self-selected clinicians means this evidences perception, not population-level frequency (Section 10.3). Evaluation across eleven foundation models found medical hallucination behaves more as a reasoning failure than a knowledge gap — models trained specifically on medical data did not reliably outperform general-purpose models. Clinical-safety evaluation of LLM-generated summaries at a single institution found roughly 1.5% of clinician-annotated sentences contained a hallucination, with a meaningful share posing a major clinical safety risk — underscoring that even a low aggregate rate can carry disproportionate downstream risk.

9. ORIGINAL ANALYSIS: GROWTH TRAJECTORY OF DOCUMENTED LEGAL-HALLUCINATION INCIDENTS

Section 8.1 cites a count of documented AI-hallucination court cases at several points in time. This section fits that trajectory quantitatively, using eleven dated snapshots (Section 2.2) — including one primary-sourced count queried directly from the tracker on 5 September 2026 — and, going beyond a single assumed functional form, formally compares four competing growth models rather than fitting only an exponential.

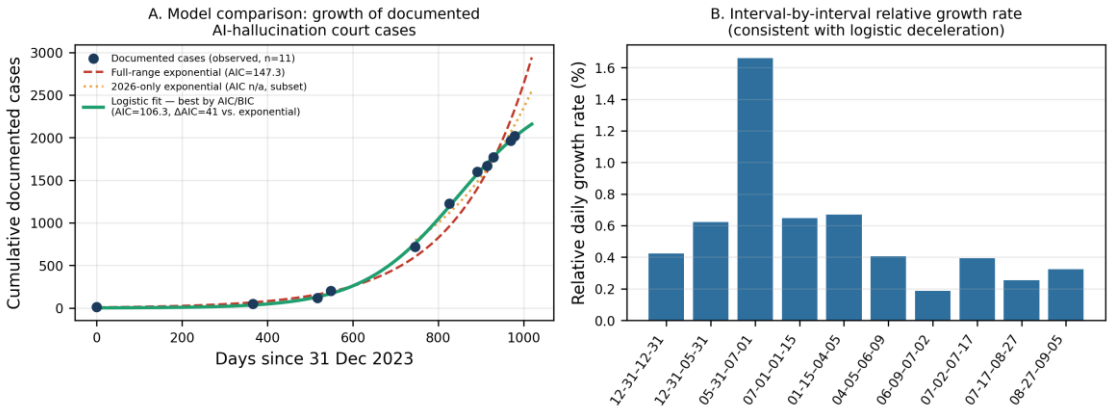


Figure 2. (A) Cumulative documented AI-hallucination court cases, Dec 2023–Sep 2026 (n=11), with exponential and logistic fits; the logistic curve is preferred by formal model comparison (Table 2). (B) Relative daily growth rate between each pair of consecutive snapshots. Data: AI Hallucination Cases Database [27], via dated secondary reports [31–36, 40].

Table 2. Formal comparison of four growth models fit to the n=11 case-count series (Gaussian-likelihood AIC/BIC; lower is better). The logistic model is preferred by ΔAIC ≈ 41 and ΔBIC ≈ 41 over the next-best (power-law) model, and by ΔAIC ≈ 41 over the exponential model used in the original single-model fit

Model	Parameters (k)	RMSE (cases)	AIC	BIC
Exponential	2	163.0	147.3	148.1
Linear	2	349.4	164.1	164.9
Power-law	2	80.8	131.9	132.6
Logistic (best fit)	3	23.1	106.3	107.5

A single exponential model fit across the full range (Dec 2023–Sep 2026, n=11) explains the data well ($R^2 = 0.99$) with an implied doubling time of 119.3 days (95% bootstrap CI: 106.0–126.4 days). Restricting the same functional form to 2026 alone gives a longer, less certain doubling time (162.1 days, 95% CI: 138.5–260.6 days, $R^2 = 0.96$), and the interval-by-interval relative growth rate — computed directly between each pair of consecutive snapshots — falls from a peak of roughly 1.7% per day (May–July 2025) to a range of 0.19–0.67% per day across every interval measured in 2026, even as the absolute number of new cases per day rose from roughly 2.6 to as high as 6.7. Table 2 goes beyond this two-way exponential comparison and formally tests four competing functional forms — exponential, linear, power-law, and logistic — against the same data using AIC and BIC. A logistic (saturating) curve is preferred by a wide margin (AIC = 106.3 versus 147.3 for the exponential, $\Delta AIC \approx 41$; the same ordering holds under BIC), with a fitted inflection point around day 845 (late March 2026) and an extrapolated asymptote near 2,600 cumulative cases. This converts what was previously an informal observation — the relative growth rate declining even as absolute counts rise — into a model-comparison result: an S-shaped, saturating curve fits this series measurably better than unconstrained exponential growth. The fitted asymptote itself should be treated cautiously (Section 10.6); the model-comparison result that logistic beats exponential is the more defensible claim of the two, and to the author's knowledge neither the informal deceleration pattern nor this formal comparison has been previously reported for this dataset.

10. CRITICAL DISCUSSION

The literature surveyed above, and the two original analyses added to it, are not uniformly rigorous, and each is examined here in turn.

10.1 RAG Is Oversold as a Solution

Mitigation papers frequently treat RAG as a near-default fix, but the same literature documents "hallucination on hallucination" and permutation sensitivity. Section 6's finding that D3 (the mechanism RAG detection relies on) has the highest coverage score of any detector should not be read as RAG being solved: coverage here means "can in principle catch," conditional on the retrieved evidence actually surfacing the relevant fact, which curated benchmark evaluations do not stress-test under open-domain retrieval conditions.

10.2 Benchmark Numbers Are Not Directly Comparable

Hallucination rates cited across this review cannot be meaningfully compared, for the reasons Table 1 makes explicit: different tasks, granularities, models, and definitions. A further concern is that most benchmarks score only correctness and ignore consistency, yet consistency — confident repetition of the same wrong answer — is precisely the signature Section 6 shows D2 and D6 are structurally unable to catch for lifecycle-instilled causes, which links this general concern to a specific, falsifiable mechanism rather than leaving it as an observation.

10.3 The Medical Evidence Base Is Thinner Than It Looks

The clinician survey in Section 8.2 rests on 70 self-selected respondents and demonstrates perceived prevalence, not measured incidence. The Asgari et al. clinical-safety framework's ~1.5% rate was evaluated on one institution's pipeline and is untested for generalization. The two studies are rarely conducted on the same systems, making it difficult to connect a measured rate to an actual patient-safety outcome.

10.4 The Legal Case Record Is a Lower Bound, and So Is the Growth Curve Built on It

The tracker underlying both Section 8.1 and Section 9 counts only cases caught, contested, and formally documented; it says nothing about fabricated citations that went unnoticed or were corrected before filing. This means Section 9's growth curve should be read as a trajectory of documentation and enforcement activity, not necessarily of the underlying rate at which LLMs produce fabricated citations — the deceleration in relative growth identified in Section 9 could reflect courts and opposing counsel reaching a more stable, saturated rate of checking for AI hallucinations, rather than models actually hallucinating less often in legal contexts. The trend supports a claim that consequences are escalating and that the pattern of escalation is itself changing shape; it does not, on its own, support a claim about the underlying technical trend.

10.5 Limits of the Coverage Framework

The matrix in Section 6 is still a reasoned synthesis coded by a single author, not a controlled experiment: no detector was actually run against outputs produced by each cause under laboratory conditions. Section 6.1's retrodictive check against four independent published findings is a real but partial form of validation — it tests whether the matrix's predictions are consistent with results other researchers already obtained under their own experimental conditions, not whether the same detectors, run by the same author under a shared protocol, would reproduce all 54 cells. That the check surfaced one disconfirming result (C3×D6) and led to a documented revision, rather than only ever confirming the original scores, is evidence the exercise was a genuine test rather than a confirmation search — but 50 of the 54 cells remain untested against any external result, corroborating or otherwise, simply because no published study happens to test that specific cause×detector pairing yet. A controlled study that seeds outputs with each of the nine causes and runs all six detector families under one protocol, replacing the remaining reasoned cells with measured ones, is still the concrete next step, and is not undertaken here.

10.6 Limits of the Growth-Curve Extrapolation

Eleven data points is still a small sample for distinguishing between growth models, and most snapshot dates were not chosen by the author but by whichever secondary outlet happened to report a count near that date, introducing potential irregularity along the time axis; the one primary-sourced point (5 September 2026) mitigates this only at a single date. The logistic model's preference in Table 2 is a genuine, formally tested result, not an eyeballed pattern, but the specific fitted asymptote (≈2,600 cases) is far more fragile than the model-ranking itself: with only eleven points and three free parameters, the logistic fit's projected saturation level would likely shift substantially with a single additional data point, particularly one taken well beyond the current September 2026 endpoint, and it should not be read as a confident forecast of where the tracker's count will level off. The model-comparison result supports the claim that this series decelerates in a manner better described by a saturating curve than an unconstrained exponential; it does not support a specific numerical prediction of the eventual case count, and Section 10.4's point about the tracker measuring documentation and enforcement activity rather than the underlying LLM hallucination rate applies to the logistic fit exactly as it applied to the earlier exponential one.

10.7 Where the Evidence Is Genuinely Strong

Not all of the literature warrants this level of caution. The causal taxonomy separating data-, training-, and inference-stage hallucination (Section 3) is well corroborated across independent surveys using different methodologies, and the finding that RLHF can induce sycophancy is supported by both behavioral studies and follow-on counter-measures. The finding that hallucination persists even in the best-performing frontier models is consistent across every benchmark reviewed here, making it one of the more durable conclusions in this review.

11. OPEN PROBLEMS AND FUTURE DIRECTIONS

Several open questions remain. First, hallucination may be a structurally unavoidable property of models trained on a likelihood objective over passive data, reframing the research goal from elimination to bounded, disclosed risk. Second, detection and mitigation research remains concentrated on English and general-domain text; the coverage framework in Section 6 has not been tested for whether its cause–detector mappings hold across languages. Third, the field lacks a unified evaluation methodology reconciling correctness- and consistency-based benchmarks. Fourth, hallucination in multimodal systems is an emerging, less mature area with only partial mechanism overlap with the text-only causes coded in Section 6. Finally, the most direct empirical extension of this paper's own contributions would be a controlled study that actually runs each of the six detector families against outputs seeded with each of the nine documented causes, replacing Section 6's reasoned scores with measured ones, and a longer observation window to test whether Section 9's exponential-to-linear transition persists, reverses, or was a temporary artifact of 2026's reporting environment.

12. CONCLUSION

This paper reviewed the causes, detection methods, mitigation techniques, and societal impact of LLM hallucination, and added two original contributions on top of that review, each strengthened against external checks rather than left as an untested assertion. The Cause–Detection Coverage Framework (Section 6) cross-tabulates nine documented causes against six detection-technique families and finds that the cheapest, most widely deployed black-box detectors are structurally blind to exactly the lifecycle-instilled causes that produce the most dangerous, confidently-wrong failure mode; a retrodictive check against four independent published findings (Section 6.1) corroborated this pattern in three cases and produced one documented revision in the fourth, which is reported as evidence the check was genuine rather than confirmatory. The growth-trajectory analysis (Section 9) fits eleven real, dated, cross-corroborated snapshots of documented legal-hallucination incidents and, going beyond a single assumed functional form, formally shows via AIC/BIC that a logistic growth model fits this series better than an exponential, linear, or power-law alternative — while remaining explicit that the tracker measures documentation and enforcement activity, not the underlying LLM hallucination rate, and that the fitted saturation level is far less trustworthy than the model-ranking itself. Neither contribution is presented as settled: Section 10 subjects both, including their new validation steps, to the same critical scrutiny applied to the rest of the field's evidence base, because a paper about hallucination and overclaiming should not exempt its own claims from that standard.

13. DECLARATIONS

13.1 Use of Artificial Intelligence

During the preparation of this manuscript, the author used an AI-based large language model assistant (Claude, Anthropic) to assist with literature search and synthesis of source material. The author reviewed, verified, and edited the content and takes full responsibility for the accuracy, originality, and content of the publication.

14. ACKNOWLEDGMENTS

The author gratefully acknowledges the open-access research community whose publicly available work made this review possible, and Damien Charlotin for maintaining the public case-tracking dataset underlying Section 9.

15. REFERENCES

[1] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1–55.

[2] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

[3] Zhao, W. X., Zhou, K., Li, J., et al. (2023). A survey of large language models. arXiv:2303.18223.

[4] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. arXiv:2109.07958.

[5] Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. arXiv:2305.11747.

[6] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv:2305.14251.

[7] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. Proceedings of NAACL-HLT.

[8] Ravi, S., et al. (2024). HaluBench: A hallucination evaluation benchmark across real-world domains. arXiv:2407.00499.

[9] Muhlgay, D., Ram, O., Magar, I., et al. (2024). Generating benchmarks for factuality evaluation of language models (FACTOR). arXiv:2307.06908.

[10] Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. arXiv:2303.08896.

[11] Mündler, N., He, J., Jenko, S., & Vechev, M. (2023). Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. arXiv:2305.15852.

[12] Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. arXiv:2304.13734.

[13] Ganesh, P., Shokri, R., & Farnadi, G. (2026). Rethinking hallucinations: Correctness, consistency, and prompt multiplicity. arXiv:2602.00723. To appear at EACL 2026.

[14] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Estimating the hallucination rate of generative AI.

arXiv:2406.07457.

[15] Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., & Herzig, J. (2024). Does fine-tuning LLMs on new knowledge encourage hallucinations? arXiv:2405.05904.

[16] Hu, M., He, B., Wang, Y., Li, L., Ma, C., & King, I. (2024). Mitigating large language model hallucination with faithful finetuning. arXiv:2406.11267.

[17] Ying, S., et al. (2024). Knowledge overshadowing causes amalgamated hallucination in large language models. arXiv:2407.08039.

[18] Xie, Z., & Sun, Z. (2026). Mitigating hallucination on hallucination in RAG via ensemble voting (VOTE-RAG). arXiv:2603.27253.

[19] Zhang, Q., Zhang, H., Pang, L., Zheng, H., & Zheng, Z. (2026). Stable-RAG: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. arXiv:2601.02993.

[20] Do, N., et al. (2025). Finetune-RAG: Fine-tuning language models to resist hallucination in retrieval-augmented generation. arXiv:2505.10792.

[21] Banerjee, S., Agarwal, A., & Singla, S. (2024). LLMs will always hallucinate, and we need to live with this. arXiv:2409.05746.

[22] Kim, Y., Jeong, H., Chen, S., Li, S. S., Lu, M., Alhamoud, K., Mun, J., et al. (2025). Medical hallucination in foundation models and their impact on healthcare. arXiv:2503.05777.

[23] Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Au Yeung, J., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine, 8(1), 274.

[24] Wille, J. (2025, May 29). Lawyer sanctioned $6,000 for AI-generated fake legal citations. Bloomberg Law.

[25] The Register. (2023, June 22). Lawyers who cited fake cases hallucinated by ChatGPT must pay. Reporting on Mata v. Avianca Airlines, S.D.N.Y.

[26] Yahoo News / AP. (2025). AI hallucinations in court documents are growing. Reporting on the AI Hallucination Cases Database case counts by year (2023–2025).

[27] Charlotin, D. (2023–2026). AI Hallucination Cases Database [dataset]. damiencharlotin.com/hallucinations.

[28] LawSites / The Law GPT. (2026). AI hallucination sanctions: A lawyer's guide for 2026.

[29] Wan, R., et al. (2026). "My AI is lying to me": User-reported LLM hallucinations in AI mobile app reviews.

[30] Saxena, R., & Bhattacharyya, P. (2024). A survey on hallucination detection methods: Intrinsic and extrinsic taxonomies. As cited in Huang et al., 2025.

[31] Platinum IDS Blog. (2026, April). 1,227 fabricated citations and counting: Inside the AI hallucination crisis hitting courts worldwide. [Charlotin database snapshot, early April 2026: 1,227 cases.]

[32] HAQQ.ai Blog. (2026, June 9). AI hallucination cases: The 1,598-case sanctions tracker. [Charlotin database snapshot, 9 June 2026: 1,598 cases.]

[33] Legal AI Space. (2026, July 2). 1,600+ AI hallucination cases: What every law firm should learn. [Charlotin database snapshot, 2 July 2026: 1,668 cases.]

[34] GC.ai Blog. (2026). AI hallucination legal cases: A sanctions tracker. [Charlotin database snapshot, 17 July 2026: 1,769 cases.]

[35] Suprmind.ai. (2026, August 27). AI hallucination rates, statistics & benchmarks in 2026. [Charlotin database snapshot, 27 August 2026: 1,963 cases; also reports the 719-case January 2026 and 1,227-case early-April 2026 figures.]

[36] The Law GPT. (2026, July). AI hallucination sanctions: A lawyer's guide for 2026. [Corroborating snapshot, ≈1,490 cases, May 2026.]

[37] Volokh, E. (2025, May 18). "AI Hallucination Cases," from courts all over the world. Reason (The Volokh Conspiracy). [Independent legal-commentary snapshot: 87 cases.]

[38] Willison, S. (2025, May 25). AI Hallucination Cases. simonwillison.net. [Independent technology-commentary snapshot citing the Charlotin tracker directly: 116 cases, 12 countries.]

[39] Manach, J.-M. (2025, July 1). Il génère des plaidoiries par IA, et en recense 160 ayant « halluciné » depuis 2023. Next (next.ink). [Independent technology-press snapshot: ≈160 cases.]

[40] Charlotin, D. (2026, September 5). AI Hallucination Cases Database [dataset, live query]. damiencharlotin.com/hallucinations. [Primary-source snapshot: 2,021 cases.]

[41] Zhang, J., Li, X., Yu, K., Peng, B., & Chen, Y. (2023). SAC3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. arXiv:2311.01740.

[42] Bhatt, M. (2026). Case study: Predictive coding and information bottleneck for hallucination detection in large language models. arXiv:2601.15652. [Single-author preprint, no confirmed peer-reviewed venue at time of citation; the sycophancy–consistency-check claim attributed within it to "Fadeeva et al." was not independently verified by the present author and is treated as weak corroboration accordingly, per Section 6.1.]

[43] Fadeeva, E., Rubashevskii, A., Shelmanov, A., Petrakov, S., Li, H., Mubarak, H., Tsymbalov, E., Kuzmin, G., Panchenko, A., Baldwin, T., Nakov, P., & Panov, M. (2024). Fact-checking the output of large language models via token-level uncertainty quantification. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 9367–9385). [Cited in Section 5–6 for token-level uncertainty quantification generally; not the source of the sycophancy-specific claim in [42], which traces to a different, unverified citation.]

[44] Chauhan, M. S. (2026). DECK: A consistency × confidence taxonomy of LLM hallucinations. arXiv:2606.02289.