American Impact Review

Review Article

Why Large Language Models Cannot Be Certified for Safety-Critical Systems

Akbar Sayakov*

Base80, AI Systems Architect, San Francisco, CA, USA

ORCID: 0009-0006-0106-0496

* Corresponding author: akbarsayakov@gmail.com

OPEN ACCESS

Citation:
Akbar Sayakov (2026). Why Large Language Models Cannot Be Certified for Safety-Critical Systems. *Am. Impact Rev.* 10.66308/air.e2026066

Received: July 16, 2026
Accepted: August 9, 2026
Published: August 10, 2026

DOI:
10.66308/air.e2026066

ISSN: 3071-124X

Abstract

Large language models (LLMs) are moving rapidly into domains governed by functional-safety certification: aviation, road vehicles, medicine, and critical infrastructure. Certification regimes such as IEC 61508, DO-178C, and ISO 26262 combine system-level risk targets with process-, traceability-, configuration-, and evidence-based obligations, in mixes that differ by regime but share a demand for verifiable, bounded behavior. This review synthesizes three literatures that rarely meet: statistical learning theory on hallucination, empirical measurements of LLM error rates in high-stakes domains, and the standards and regulatory documents that define certification. The convergent finding is that the mismatch between the two worlds is structural rather than incidental. Impossibility results establish a nonzero error floor for calibrated probabilistic generators under stated conditions; reported benchmark metrics in legal, medical, and agentic tasks are not directly convertible to certification targets, and no published deployment supplies the system-level hazard and exposure model that a compliance demonstration would require; and every major mitigation family (retrieval augmentation, guardrails, formal verification, uncertainty quantification, and neurosymbolic hybrids) is documented to narrow, but not close, the gap. We formalize error compounding over execution horizons, tabulate the documented limits of each mitigation, and examine why plausibility cannot substitute for assurance. Two coherent exits emerge: statistical acceptance criteria in standards for bounded tasks, and architectures that confine the LLM to an untrusted-proposer role inside a deterministic, independently verifiable execution envelope. Certifying the envelope rather than the model is, on the current evidence, the most defensible path consistent with both the mathematics and the standards.

Keywords: large language models, hallucination, functional safety, certification, IEC 61508, DO-178C, safety-critical systems, deterministic execution

1. Introduction

1.1 Motivation and context

Between 2023 and 2026, large language models crossed a threshold that separates consumer novelty from regulated infrastructure. The European Union's Artificial Intelligence Act entered into force with an explicit list of high-risk application areas, including safety components of critical infrastructure, medical devices, and systems used in the administration of justice, and with Article 15 requiring that high-risk AI systems achieve "an appropriate level of accuracy, robustness, and cybersecurity" that must be declared and demonstrated (European Parliament and Council, 2024). In the United States, the National Institute of Standards and Technology issued a dedicated risk-management framework for artificial intelligence while candidly recording

the immaturity of current AI risk-measurement practice (NIST, 2023). Sector regulators moved in parallel and with notable caution: the European Union Aviation Safety Agency published a human-centric AI roadmap and the first usable guidance for machine-learning applications in aviation (EASA, 2023; EASA, 2024), and the Federal Aviation Administration issued a roadmap for AI safety assurance built around an explicitly incremental introduction of AI capabilities (FAA, 2024).

The pressure to put language models inside these regulated loops is not hypothetical. LLM-based assistants are already drafting legal filings, summarizing clinical evidence, generating maintenance documentation, and orchestrating multi-step software workflows. The documented failures are equally concrete. In *Mata v. Avianca*, a federal court sanctioned attorneys who filed a brief containing judicial opinions fabricated by a chatbot (Mata v. Avianca, 2023). In *Moffatt v. Air Canada*, a tribunal held an airline liable for misinformation produced by its own customer-service chatbot, rejecting the argument that the bot was a separate entity responsible for its own statements (Moffatt v. Air Canada, 2024). Systematic audits have found that public LLMs hallucinate on 58 % to 88 % of direct, verifiable questions about randomly sampled federal court cases (Dahl et al., 2024), and that 91.4 % of the bibliographic references produced by a leading chatbot in systematic-review tasks were classified as hallucinated under the study's criteria (Chelli et al., 2024). Community efforts such as the AI Incident Database now catalog these failures precisely so that they are not repeated blindly (McGregor, 2021).

1.2 Research questions and scope

This review asks a deliberately narrow question. Can a system that places a large language model inside its critical execution path, where the model's output directly determines a safety-relevant action, satisfy the existing certification regimes for safety-critical software? We decompose the question into three parts. First, what do certification regimes actually require of a software artifact (Section 2 and Section 7)? Second, what does the published literature establish about the error behavior of LLMs, theoretically (Section 4) and empirically (Section 5)? Third, do the documented mitigation strategies change the answer (Section 6)?

Three scope restrictions apply throughout. The claim concerns LLMs *inside* the certified boundary, not offline uses such as drafting assistance reviewed by a qualified human who retains full responsibility for the output. The claim concerns *current* transformer-based autoregressive architectures and the *current* generation of quantified certification regimes; Section 8.3 discusses how the argument behaves as either side evolves. Finally, the claim is about certification in the strict sense of demonstrating conformity to quantified failure targets and traceability obligations, not about the looser sense in which any risk-managed deployment is sometimes called "certified."

1.3 Contributions

The review makes four contributions. First, it connects three literatures (statistical learning theory, functional-safety engineering, and the philosophy of language technology) that address the same question with almost no cross-citation. Second, it assembles a corpus of 55 sources, including the standards and regulatory documents themselves, checks each against primary records, and extracts the quantitative claims that survive comparison with the primary texts (Section 3 describes the protocol). Third, it formalizes the horizon argument: the elementary but underappreciated arithmetic by which per-response error rates that look tolerable in a chat window become near-certain failure over the execution horizons of industrial automation (Section 4.2, Figure 1). Fourth, it develops a structural diagnosis (Section 8) and a taxonomy of the two coherent responses, with an explicit treatment of the objections a skeptical reviewer should raise (Section 8.3).

1.4 Organization of the paper

Section 2 provides background on certification regimes and on the probabilistic foundations of LLMs, and situates this review among existing surveys. Section 3 describes the review methodology. Section 4 synthesizes the theoretical results on hallucination inevitability and error compounding. Section 5 collects the empirical evidence across domains and benchmarks. Section 6 reviews mitigation strategies and their documented limits. Section 7 analyzes how standards bodies and regulators are responding. Section 8 discusses the epistemic dimension, states the structural diagnosis, answers anticipated objections, and maps the two pathways forward. Section 9 acknowledges the limitations of the review itself, Section 10 outlines future research directions, and Section 11 concludes.

2. Background

2.1. Safety-critical certification regimes

Functional-safety certification is best understood as a *grammar* for justified confidence in software behavior, developed over four decades in response to accidents. Its foundational instrument, IEC 61508, defines four Safety Integrity Levels (SIL 1 through SIL 4) and ties each level to quantified targets for the probability of dangerous failure, expressed as average probability of dangerous failure on demand (PFD_{avg}) for low-demand functions and probability of dangerous failure per hour (PFH) for high-demand or continuous functions (IEC, 2010). The essential feature is not the specific numbers but the *form* of the obligation: the developer must demonstrate, with evidence acceptable to an assessor, that a specific quantified reliability target is met by the delivered system.

The avionics counterpart, DO-178C, assigns software to Design Assurance Levels (DAL A through E) according to the severity of the failure condition it can contribute to, and attaches to each level a set of objectives covering planning, requirement capture, verification, and configuration management (RTCA, 2011). Two of its obligations matter especially for the present argument. Requirement traceability demands a demonstrable link from every applicable software requirement to the code that implements it and the tests that verify it, in both directions: no unreviewed functionality may exist in the delivered binary. Structural coverage analysis, up to Modified Condition/Decision Coverage (MC/DC) at DAL A, demands that testing exercise the internal structure of the code to a defined criterion, so that the evidence speaks about the artifact as built, not merely about a sample of its behavior. The automotive standard ISO 26262 organizes the same discipline around Automotive Safety Integrity Levels (ASIL A through D) across the vehicle lifecycle (ISO, 2018). Table 1 summarizes the instruments that recur throughout this review.

Table 1. Certification and assurance instruments discussed in this review.

Instrument	Domain	Core assurance mechanism	Unit of assurance
IEC 61508 (2010)	Electrical/electronic/programmable safety systems	SIL 1 - 4 tied to quantified dangerous-failure targets	PFD_{avg} / PFH
DO-178C (2011)	Airborne software	DAL A-E objectives; requirement traceability; structural coverage incl. MC/DC	Qualitative level; quantitative targets set at aircraft level in companion documents
ISO 26262 (2018)	Road vehicles	ASIL A-D; V-model verification across lifecycle	Qualitative integrity level
ISO/PAS 8800 (2024)	Road vehicles, AI-specific	AI safety lifecycle; data and operational-design-domain arguments	Process conformity
ANSI/UL 4600 (2023)	Autonomous products	Safety case: claim-argument-evidence	Argumentation adequacy
FDA PCCP guidance (2024)	AI-enabled medical device software	Predetermined change control plan	Pre-specified change envelope
EASA Concept Paper Issue 02 (2024)	Aviation ML (Levels 1 - 2)	W-shaped learning-assurance process	Learning-assurance objectives

The instruments of Table 1 combine system-level risk targets with process-, traceability-, configuration-, and evidence-based obligations, and the exact mix differs by regime. What recurs across the rows, and defines "certified" for this review, is an auditable combination of a *quantified or argued bound* on dangerous failure, *traceability* between requirements and implemented behavior, and an *identified, version-controlled configuration* under assessment. These are the properties that the literature reviewed in Sections 4 through 6 shows current LLMs struggle to provide.

2.2. Probabilistic foundations of large language models

A large language model defines a probability distribution over token sequences. Given an input context x and previously generated tokens $y_{<t}$, the model produces the next token by sampling from a learned conditional distribution, so that the probability of a complete response $y = (y_1, \ldots, y_T)$ factorizes autoregressively:

$$p_\theta(y|x) = \prod_{t=1}^{T} p_\theta\left(y_t|x, y_{<t}\right) (1)$$

Decoding introduces a further stochastic layer. Under temperature sampling, the probability of emitting vocabulary item v at step t is a rescaled softmax over the model's logits z :

$$p_T(y_t = v) = \frac{exp(z_v/T)}{\sum_u exp(z_u/T)} (2)$$

Two consequences follow directly from this construction and frame everything that follows. First, the output is a *draw from a distribution*, not a deterministic function of a specification: identical inputs may yield different outputs under sampling. DO-178C-style traceability runs from requirements through design and code to verification evidence; for a generative model the analogous chain terminates in undifferentiated weights, and while greedy decoding can make outputs reproducible, reproducibility of a learned sampling map is not

a specification of its correctness over an open input domain. Second, the model's fidelity is an aggregate statistical property of a training corpus, not a per-requirement engineering artifact. The survey literature has developed a detailed taxonomy of the resulting failure modes, distinguishing intrinsic hallucination (output contradicting the source) from extrinsic hallucination (output unverifiable against the source) (Ji et al., 2023), and extending the taxonomy to instruction-following and factuality failures specific to modern LLMs (Huang et al., 2025).

2.3. Related surveys and the position of this review

Several substantial surveys cover the hallucination phenomenon itself: its taxonomy in natural language generation (Ji et al., 2023), its principles, detection methods, and mitigation benchmarks in LLMs (Huang et al., 2025), and the emerging fact-checking and factuality-evaluation toolchain. On the certification side, surveys exist on machine-learning certification in avionics (Vidot et al., 2021, preprint) and on the open challenges of certifying airborne AI (Luettig et al., 2024), and the working groups behind the forthcoming aeronautical ML standard have published their development lifecycle together with a coverage analysis of existing guidance (Kaakai et al., 2022). What has been missing is a synthesis that reads these two literatures against each other, together with the philosophical work on what kind of epistemic artifact an LLM is (Section 8.1), and asks whether the certification question has an answer in principle rather than merely a backlog of engineering work. That synthesis is the purpose of this review; the closest antecedents are the analysis of ISO 26262 techniques against machine-learned components (Salay et al., 2018), which reached a structurally similar conclusion for pre-LLM machine learning, and the position papers accompanying the EASA and FAA roadmaps (EASA, 2023; FAA, 2024).

3. Review Methodology

3.1 Design and search strategy

This is a narrative review with systematic elements, not a full systematic review; Section 9 discusses the difference and its consequences. The corpus was assembled by structured searches along seven predefined directions corresponding to the analytical needs of the argument: (i) statistical and theoretical origins of hallucination; (ii) the text and structure of safety-certification standards; (iii) regulator and industry work on certifying machine learning in aviation; (iv) reliability of multi-step LLM agents and error compounding; (v) mitigation approaches and their documented limits; (vi) epistemological and philosophical analyses of LLM output; and (vii) documented real-world failures and their legal and regulatory consequences. Searches covered arXiv, the ACM Digital Library, Springer, Oxford University Press, IEEE, Nature, PNAS, official standards catalogs (IEC, ISO, UL, RTCA), regulatory repositories (EASA, FAA, FDA, NIST, EUR-Lex), and court records, with a cutoff of 3 August 2026.

3.2. Inclusion criteria and verification protocol

Sources were retained if they satisfied all of the following: (a) the publication exists at a resolvable location (publisher page, DOI, arXiv identifier, or official repository); (b) bibliographic metadata (title, authors, year, venue) could be confirmed against the primary location; and (c) the specific claim for which the source is cited in this review appears in the source itself, in the abstract or body text, rather than in secondary commentary. Bibliographic metadata and cited claims were checked in two passes against primary sources: an initial screening pass followed by a claim-by-claim comparison of each bibliographic record and each quantitative statement with the primary text. Candidates whose headline claims could not be matched to the primary text

were corrected where the discrepancy was minor and attributable (three cases, all involving overstated or mis-scoped claims rather than nonexistent publications) or excluded otherwise. Digital object identifiers were resolved and title-matched through the Crossref and DataCite registries where available; sources without a registered DOI (standards, regulatory documents, court records, and certain proceedings papers) were checked against official publisher or repository records and are cited by their official designations.

3.3 Corpus overview

The final corpus comprises 55 sources: 7 on hallucination fundamentals (including two impossibility theorems published at STOC and in *Nature*), 8 standards and regulatory instruments, 9 on machine-learning certification in aviation, 8 on agent reliability and error compounding, 8 on mitigation families and their limits, 8 philosophical and epistemological analyses, and 7 documented incidents, audits, and regulatory frameworks. The distribution is deliberately balanced: no single literature dominates the argument, and the standards and regulatory documents are treated as primary sources in their own right rather than through secondary commentary. Where a source is a preprint rather than a peer-reviewed publication, this status is flagged at the point of citation, and no load-bearing quantitative claim in this review rests on a preprint alone.

4. Theoretical Limits of Probabilistic Generation

4.1 Hallucination as a theorem, not a defect

The central development of 2024 - 2026 in the theory of language models is the demonstration that hallucination is a *provable consequence* of the statistical objective, not an artifact of poor engineering. Kalai and Vempala (2024) proved, in a peer-reviewed result at the ACM Symposium on Theory of Computing, that any language model satisfying a natural statistical calibration condition must hallucinate on "arbitrary" facts (facts whose truth cannot be inferred from patterns in the training data) at a rate bounded below by the fraction of such facts that appear exactly once in the training corpus, a quantity estimated by the classical Good-Turing method. Writing $\hat{s}$ for that singleton fraction, the bound takes the schematic form

$$err_{gen} \gtrsim \hat{s} - o(1), \quad (3)$$

and, under the stated calibration condition, it does not depend on model architecture or training-data quality: even an error-free corpus does not remove it, because the pressure toward hallucination comes from calibration itself, the very property that makes a probabilistic model's confidence meaningful.

The successor result strengthens the diagnosis. Kalai, Nachum, Vempala, and Zhang reduce valid generation to a binary classification problem ("Is-It-Valid") and show that the generative error rate is bounded below by approximately twice the achievable misclassification rate on that problem (Kalai et al., 2026):

$$err_{gen} \gtrsim 2 \cdot err_{\mathrm{IIV}} - o(1). \quad (4)$$

The *Nature* version adds a socio-technical mechanism for the persistence of the phenomenon: dominant benchmark metrics score binary accuracy, and under binary scoring a model that guesses confidently strictly outperforms one that admits uncertainty, so the evaluation ecosystem itself trains overconfident error into the deployed population of models (Kalai et al., 2026). Independently, Xu, Jain, and Kankanhalli (2024, preprint) formalized hallucination in computability-theoretic terms and showed that for any fixed computable language model there exists a computable ground-truth function on which it errs, so no single model can serve as an infallible general problem solver, however trained; and Banerjee, Agarwal, and Singla (2025) argued that each

stage of the LLM pipeline, from training-data compilation through fact retrieval and intent classification to text generation, contributes its own irreducible hallucination channel. The theoretical picture is corroborated at the detection layer: Farquhar et al. (2024), in *Nature*, operationalized "confabulation" via semantic entropy over meaning-equivalence classes of sampled answers and achieved useful but far-from-perfect discrimination (AUROC 0.790 across models and tasks), which is what one expects if erroneous generation is woven into the sampling process rather than being a rare malfunction.

For certification, the significance of these results lies in their form rather than their magnitude. They rule out the assumption of error-free open-ended generation under their stated conditions (calibration, rare or arbitrary facts, general-purpose use), but they do not by themselves quantify the probability of a dangerous failure in a declared operational domain, and their authors are explicit that restricted architectures and post-processing can help on repeated and systematic facts. What the theorems remove is the default assurance argument: a developer cannot claim that residual generation error will be engineered to zero, because for the general, open-ended case a nonzero floor is established mathematically under the theorems' conditions, and a quantified safety claim must therefore be built on something other than the model's own correctness.

4.2. The arithmetic of execution horizons

The second theoretical strand concerns what happens when single-response error rates are embedded in multi-step execution, which is precisely the deployment mode of industrial automation. Let a task require n dependent steps and let ε be the per-step probability of an erroneous step, assumed independent for the moment. The probability that the entire chain executes without error is

$$R(n) = (1-\varepsilon)^n, F(n) = 1-(1-\varepsilon)^n, \quad (5)$$

which decays exponentially in the horizon n . The illustration circulating in the literature makes the point vividly: a per-token error probability as small as 0.01 compounds to a 0.86 probability of at least one error over 200 tokens, since $1-0.99^{200} \approx 0.86$ (Bachmann and Nagarajan, 2024, who present the calculation as a recap of the pre-existing "snowballing errors" criticism before contributing a distinct training-time failure mode; the underlying bound originates in the imitation-learning literature). That older bound deserves precise statement: Ross, Gordon, and Bagnell showed that naive behavior cloning, which trains a predictor on ground-truth prefixes and deploys it on its own outputs, incurs total cost growing as $O(T^2\varepsilon)$ over a horizon of T decisions, because early errors shift the state distribution away from anything seen in training; interactive corrective training (DAgger) restores the linear $O(T\varepsilon)$ regime but requires an oracle that can label the model's own error states (Ross et al., 2011):

$$J_{clone}(\hat{\pi}) \leq J(\pi^*) + O(T^2\varepsilon) \text{ vs. } J_{\text{DAgger}}(\hat{\pi}) \leq J(\pi^*) + O(T\varepsilon). \quad (6)$$

Autoregressive generation is structurally a behavior-cloning deployment: the model was trained on human-written prefixes and generates on its own. Empirical work confirms the mechanism in modern architectures: transformer performance on compositional tasks degrades sharply with the depth of the required reasoning graph, consistent with compounding rather than uniform per-instance difficulty (Dziri et al., 2023), and controlled long-horizon studies find that models *self-condition* on their own earlier mistakes, so that observed dependence between steps makes matters worse than the independence baseline of Equation (5), not better (Sinha et al., 2026).

Figure 1 plots Equation (5) for per-step error rates from the empirically reported range down to levels far below any current measurement. The geometry is the argument: to keep a 100-step execution inside an overall

failure budget of 10^{-5} requires per-step error near 10^{-7} , far below any error rate reported in the studies of Section 5, and the requirement tightens as horizons lengthen. Certification-grade targets themselves are defined in different units and are deliberately not plotted (Sections 5.3 and 8.3).

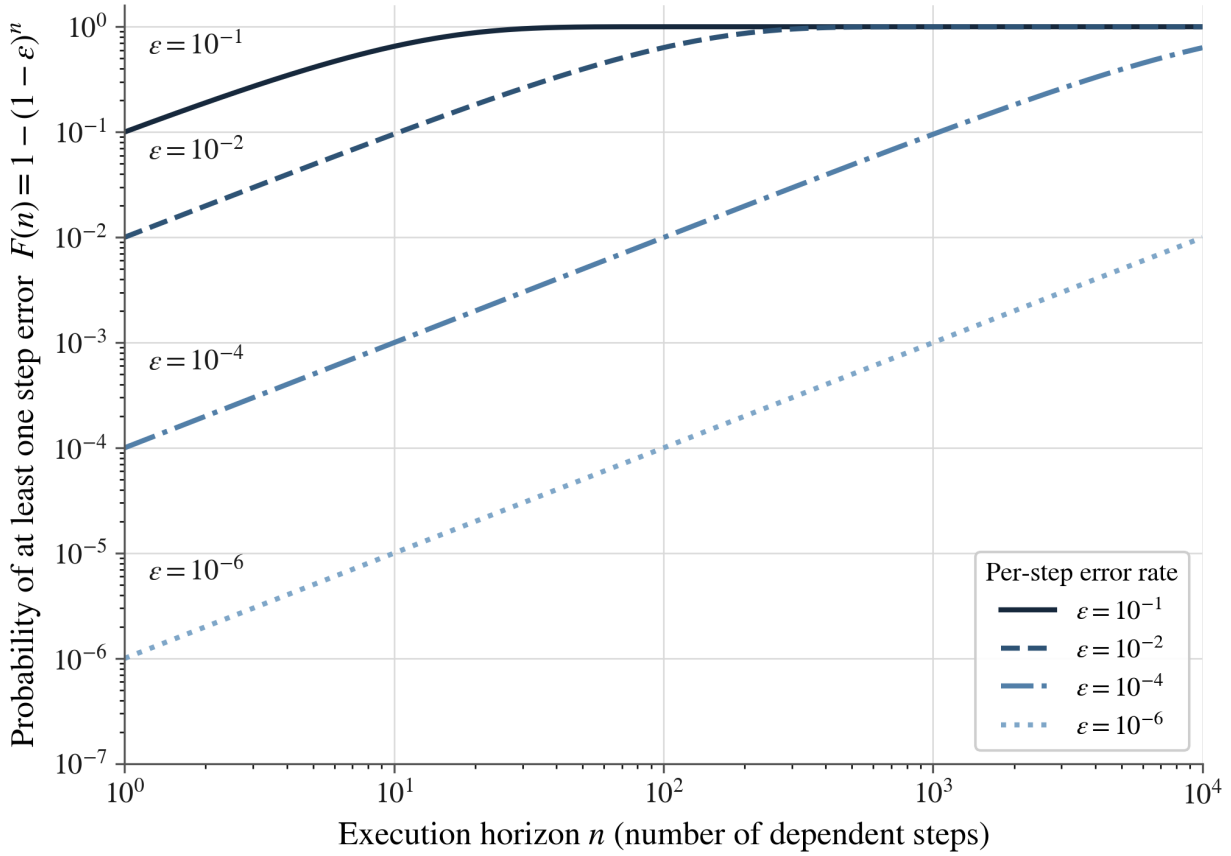

Figure 1. Probability of at least one uncorrected step error under the i.i.d. assumption, $F(n) = 1 - (1 - \varepsilon)^n$, over execution horizon n for per-step error rates ε from 10^{-1} to 10^{-6} , on logarithmic axes. The figure illustrates the arithmetic of uncorrected error accumulation only. Certification targets (PFD_{avg}, PFH) are defined in different units and over different reference frames and are therefore not plotted; Sections 5.3 and 8.3 discuss why the two families of quantities are not directly comparable. Step-independence is a modeling assumption; measured dependence (self-conditioning) worsens the picture (Sinha et al., 2026).

4.3. The specification-completeness problem

The third theoretical obstacle is prior to any question of reliability: certification verifies an artifact *against a specification*, and for open-ended language tasks the specification does not and cannot exist in the required form. DO-178C's traceability obligation presupposes an enumerable set of requirements that jointly define all intended behavior; MC/DC coverage presupposes that exercising the code's decision structure is a meaningful proxy for exercising the requirement space. For a model whose input domain is unrestricted natural language and whose intended function is "respond correctly and appropriately," neither presupposition holds. The point was established for conventional machine learning before LLMs existed: Salay, Queiroz, and Czarnecki's analysis of ISO 26262 found that roughly 40 % of the unit-level software techniques assessed in Part 6 of the standard are inapplicable to machine-learned components, precisely because training sets substitute for specifications and network weights substitute for reviewable design (Salay et al., 2018). The EASA guidance responds to the same difficulty for narrow models by demanding a closed operational design domain and a

frozen model as preconditions of learning assurance (EASA, 2024); open-ended generation, whose value proposition is exactly its unbounded input space, cannot satisfy the precondition by construction.

5. Empirical Evidence Across Domains

5.1. Measured error rates in professional domains

The theoretical floor of Section 4 would matter little if measured error rates were already near certification targets. They are not: read simply as magnitudes, the measured rates sit orders of magnitude above the quantified targets of Section 2.1, with the caveats on comparability detailed in Section 5.3. In the largest systematic audit of legal hallucination, public models asked direct, verifiable questions about randomly sampled federal court cases hallucinated between 58 % (ChatGPT 4) and 88 % (Llama 2) of the time, accepted false legal premises embedded in questions (contra-factual bias), and could not reliably predict their own hallucinations (Dahl et al., 2024). In medicine, a comparative analysis of reference accuracy found that 39.6 % (GPT-3.5), 28.6 % (GPT-4), and 91.4 % (Bard) of generated citations were classified as hallucinated under the study's field-level criteria (Chelli et al., 2024). These are not adversarial prompts; they are the ordinary work products of the professions concerned.

Two further measurements matter because they test the *best available* configurations rather than raw chat models. A peer-reviewed audit of commercial legal research tools built on retrieval-augmented generation, marketed with explicit "hallucination-free" claims, found hallucinated or misgrounded output in 17 % to 33 % of responses across the leading products (Magesh et al., 2025). And an industrial study of retrieval-augmented structured-output generation, deployed for enterprise workflow automation, reduced hallucination from roughly 21 % without retrieval to below 8 % with it (Béchard and Marquez Ayala, 2024). The latter figures are measured per generated workflow step and table name rather than per response, and are, to our knowledge, among the lowest hallucination indicators reported in a published production setting; they are nevertheless of order 10^{-2} , while the quantified failure targets of Section 2.1 are set orders of magnitude lower.

5.2. Reliability of multi-step agents

Benchmark results for agentic deployment, where the model plans and executes sequences of actions, exhibit exactly the horizon sensitivity that Equation (5) predicts. On WebArena, a realistic web-task environment, the best end-to-end agent at publication achieved a 14.41 % success rate against a human baseline of 78.24 % (Zhou et al., 2024). On SWE-bench, resolving real GitHub issues, the strongest model at publication resolved 1.96 % of instances (Jimenez et al., 2024). On τ-bench, which simulates tool-using customer-service agents under realistic multi-turn interaction, average success was well below reliable-service levels, and the consistency metric pass^8 (the probability that the agent succeeds on all eight independent replays of the same task) fell below 25 % in the retail domain, demonstrating that even tasks the agent can complete are not completed *repeatably* (Yao et al., 2025).

The longitudinal trend is real but does not change the conclusion within any planning horizon relevant to standards bodies. The METR study of task-completion time horizons found that the task length at which frontier models achieve 50 % success has been doubling roughly every seven months, reaching roughly 110 minutes of human-equivalent work for the strongest model in the published version; critically, the horizon at which the same models achieve 80 % success is four to six times shorter, and certification-relevant reliability levels are off the measured scale entirely (Kwa et al., 2025, published at NeurIPS 2025). Controlled long-horizon execution studies confirm the mechanism: per-step accuracy gains translate into disproportionate horizon gains, yet absolute end-to-end reliability collapses as horizons grow, with models degrading further

by conditioning on their own earlier errors (Sinha et al., 2026, published at ICLR 2026).

5.3 Quantifying the gap

Table 2 consolidates heterogeneous reliability indicators; the metrics differ by construction, and none converts directly to the per-demand or per-hour quantities in which certification targets are stated (Section 8.3 addresses the mismatch directly). For orientation: the lowest indicator in Table 2 (below 8 %, measured per structured-output step) is separated by more than three orders of magnitude from the least demanding continuous-mode band of IEC 61508 (PFH at SIL 1). These benchmark metrics are not directly convertible to PFD_{avg} or PFH without an explicit system-level hazard, exposure, demand-rate, and mitigation model; the comparison is an illustration of distance, not a claimed measurement of it. What the numbers do establish is the absence of any published evidence that a quantified certification target has been met by an LLM component.

Table 2. Heterogeneous empirical reliability indicators for LLM systems in high-stakes settings. Values are as reported (or rounded) in the cited studies; configurations marked RAG include retrieval augmentation. The metrics differ by construction (hallucination rates, end-to-end success, replay consistency, time horizons): rows are not mutually comparable, and none converts directly to a per-demand failure probability.

Domain and task	System(s)	Measured rate	Metric	Source
Legal Q&A on random federal cases	ChatGPT 4 / ChatGPT 3.5 / PaLM 2 / Llama 2	58 % - 88 %	Hallucination rate	Dahl et al. (2024)
Commercial legal research (RAG)	Leading legal AI products	17 % - 33 %	Hallucinated or misgrounded responses	Magesh et al. (2025)
Medical reference generation	GPT-3.5 / GPT-4 / Bard	39.6 % / 28.6 % / 91.4 %	References classified as hallucinated under the study's criteria	Chelli et al. (2024)
Enterprise structured outputs (RAG)	Production workflow system	≈21 % to <8 %	Hallucination per generated workflow step	Béchard & Marquez Ayala (2024)
Autonomous web tasks	Best agent at publication	14.41 % success (human: 78.24 %)	End-to-end task success	Zhou et al. (2024)
Real software-repository issues	Strongest model at publication	1.96 % resolved	Issue resolution rate	Jimenez et al. (2024)
Tool-using service agents	Frontier agents	pass^8 < 25 % (retail domain)	Success on all 8 replays	Yao et al. (2025)
Long-task time horizon	Frontier models, 2019 - 2025	≈110 min at 50 % success (strongest model); 80 %-horizon 4 - 6× shorter	Human-equivalent task length	Kwa et al. (2025)

6. Mitigation Strategies and Their Documented Limits

6.1 Retrieval-augmented generation

Retrieval augmentation grounds generation in documents fetched at inference time, attacking the parametric-memory failure mode directly. Its measured effect is a substantial reduction, never an elimination: the

production study above achieved the reduction from ≈21 % to below 8 % in structured outputs (Béchard and Marquez Ayala, 2024), while the audit of commercial legal tools shows that mature, professionally deployed RAG systems still produce hallucinated or misgrounded answers in one response out of every three to six, even as their marketing claimed the problem solved (Magesh et al., 2025). The survey literature systematizes why: retrieval can fail, retrieved passages can be misread, and the generator remains free to interpolate beyond its evidence (Huang et al., 2025). RAG relocates trust from parametric memory to a retrieval pipeline; it does not change the generator's statistical nature, which is what Equation (3) and Equation (4) bound.

6.2 Guardrails and output filtering

Guardrail frameworks interpose programmable checking between model and user. NeMo Guardrails, the most widely documented toolkit, allows developers to express dialog-level rails and content constraints, and its authors present it as an engineering layer for controllability, without formal guarantees of constraint satisfaction (Rebedea et al., 2023). The theoretical ceiling for this family was articulated by Glukhov et al. (2024) in an ICML position paper: semantic censorship of LLM output, deciding whether generated content carries a prohibited meaning, inherits undecidability from the generality of the languages involved, and "can be perceived as an undecidable problem"; the authors conclude that filtering should be treated as a security problem (attack surface management) rather than as a solvable ML-metrics problem. A filter that is itself heuristic cannot upgrade a probabilistic generator into a bounded-failure component; it composes two fallible layers.

6.3. Formal verification of neural networks

Formal methods are certification's native language, so their status for neural networks is decisive for any hope of "verifying the model." The field's foundational result is sobering in the present context. Reluplex demonstrated that verifying properties of ReLU networks is possible in practice for networks of a few hundred nodes, while establishing that the underlying decision problem is NP-complete (Katz et al., 2017). The international verification competition VNN-COMP has since matured the tooling substantially, yet its own retrospective records benchmark networks that remain orders of magnitude smaller than production language models, and the properties verified are local (robustness within a norm ball around given inputs) rather than global or semantic (Brix et al., 2023). No published technique verifies a semantic property, such as "the generated answer is factually correct," for a transformer at LLM scale; the gap between what formal verification can state and what certification would need it to state is qualitative, not a matter of compute.

6.4. Uncertainty quantification and abstention

A different strategy accepts that generation errs and tries to *know when*. Semantic entropy estimates uncertainty over meaning-equivalence classes of sampled answers and detects confabulations with AUROC 0.790 across tasks and models (Farquhar et al., 2024): valuable, probabilistic, and post hoc, and therefore a monitoring signal rather than a guarantee. The strongest formal result in this family is conformal factuality: by calibrating on exchangeable data and progressively filtering sub-claims, one can guarantee that the retained output is correct with probability at least $1 - \alpha$:

$$\Pr\left(\text{all retained claims are correct}\right) \geq 1 - \alpha, \quad (7)$$

with published evaluations achieving 80 % - 90 % guarantee levels at acceptable output-retention rates (Mohri and Hashimoto, 2024). Two features keep even this result short of certification. The guarantee is statistical

over an exchangeable population of queries, not a worst-case bound over an operational design domain, and it is purchased by abstention: as α tightens toward certification levels, the system converges toward refusing to answer, which in a safety function is itself a failure mode with its own budget. Uncertainty quantification is, on the published evidence, the mathematics of *managed* unreliability, not of certified reliability.

6.5 Neurosymbolic architectures

The neurosymbolic program proposes hybrid systems in which neural components propose and symbolic components reason, aiming to combine learning with the auditability of logic (d'Avila Garcez and Lamb, 2023). Its relevance here is double-edged and instructive. To the extent a symbolic layer *checks* neural proposals against formal constraints, the resulting guarantee attaches to the symbolic layer, and the neural model is demoted to an untrusted generator of candidates, which is precisely the architectural exit developed in Section 8.4. But as a claim about making the neural component itself trustworthy, the program remains open: its own literature lists provably sound knowledge extraction from large networks among unsolved problems, and the third-wave survey is explicit that sound, explainable neurosymbolic computation is a research agenda rather than an available mechanism (d'Avila Garcez and Lamb, 2023). Table 3 consolidates the mitigation landscape.

Table 3. Mitigation families, the type of guarantee each provides, and the documented limit.

Family	Mechanism	Guarantee type	Documented limit	Key sources
Retrieval augmentation	Ground generation in retrieved evidence	Empirical reduction	17 - 33 % residual in commercial legal tools; <8 % per structured-output step in the best published production case	Magesh et al. (2025); Béchard & Marquez Ayala (2024)
Guardrails / filtering	Programmable rails; output classification	Procedural, no formal bound	Semantic filtering argued undecidable; rails are engineering controls	Rebedea et al. (2023); Glukhov et al. (2024)
Formal verification	SMT / branch-and-bound over network	Deterministic, but local properties	NP-complete; verified nets $\ll$ LLM scale; properties are robustness, not truth	Katz et al. (2017); Brix et al. (2023)
Uncertainty quantification	Semantic entropy; conformal filtering	Statistical ($1-\alpha$), post hoc	Detection AUROC 0.790; guarantee via abstention; exchangeability assumption	Farquhar et al. (2024); Mohri & Hashimoto (2024)
Neurosymbolic hybrids	Symbolic layer over neural proposer	Soundness as goal	Provably sound extraction from large nets an open problem	d'Avila Garcez & Lamb (2023)

The pattern across all five rows is uniform. Each mitigation converts "wrong answer" into "less frequently wrong answer" or "no answer." None converts a probabilistic generator into a component with a demonstrable worst-case failure bound, which is the currency in which certification trades.

7. Standards and Regulatory Responses

7.1. AI-aware extensions of certification frameworks

Standards bodies have not been idle, and the shape of their response is itself evidence about the problem. ISO/PAS 8800 extends automotive functional safety with an AI-specific lifecycle covering data management, operational design domains, and AI-specific verification; the document itself concedes that "it is not possible to provide detailed requirements … to achieve an acceptably low level of residual risk," locating its contribution in process discipline rather than quantified acceptance (ISO, 2024). UL 4600 embraces safety-case argumentation for autonomous products, admitting statistical and simulation evidence but leaving the burden of a convincing quantitative argument on the applicant (UL, 2023). The FDA's predetermined change control plan guidance permits AI-enabled device software to evolve after clearance, but only within a bounded envelope of planned modifications whose development, validation, and impact-assessment methodology is specified and reviewed in advance (FDA, 2024, reissued August 2025); the mechanism presupposes a pre-specifiable change space, which open-ended model evolution does not offer. In aviation, the EASA concept paper defines a W-shaped learning-assurance process for Level 1 - 2 machine learning, anchored on a frozen model, a closed operational design domain, and demonstrated generalization; the CoDANN projects worked through what such arguments look like for perception networks (EASA and Daedalean, 2020; EASA and Daedalean, 2024; EASA, 2024), and the FAA's roadmap commits to incremental introduction (FAA, 2024).

Two facts about these instruments matter for the thesis. First, every one of them presupposes an identified, version-controlled configuration or a controlled, pre-authorized change process: an enumerable ODD, a fixed or envelope-bounded model version. Open-ended generative deployment, as practiced over continuously updated models and APIs, violates the presupposition, not the fine print. Second, where narrow machine learning has actually been brought inside a certification argument, the mechanism was architectural: a published DAL C assurance case study for a runway-sign classifier reached its assurance level not by verifying the network to DAL C but by composing two dissimilar DAL D networks with a conventional safety monitor, an explicitly architectural mitigation that treats each network as individually untrustworthy (Dmitriev et al., 2023). The strongest published assurance case in the corpus is thus an early instance of the envelope pattern of Section 8.4; it documents a certifiable design, not a completed product certification.

7.2 The measured readiness gap

Where the coverage of existing guidance has been measured against the needs of ML assurance, the result is quantified incompleteness. The joint EUROCAE WG-114 / SAE G-34 working group, developing the forthcoming SAE ARP6983 / EUROCAE ED-324 standard (designated AS6983 in the group's earlier publications), surveyed eight existing standards and guidance documents against its baseline of learning-assurance objectives and found that only three exceeded 50 % coverage, with none approaching completeness (Kaakai et al., 2022). The independent survey literature reaches the same verdict qualitatively: existing means of compliance do not yet accommodate machine-learned components, and the open problems are foundational rather than clerical (Vidot et al., 2021, preprint; Luettig et al., 2024). These findings concern *narrow* supervised models, the easy case. Generative systems entered the regulatory perimeter only in mid-2026: the proposed Issue 03 of the EASA concept paper, released for consultation on 3 June 2026, extends the guidance toward Level 3 advanced automation and brings additional AI techniques, including generative models, into scope, and it does so on terms that corroborate this review's diagnosis, keeping AI constituents that touch the most severe failure conditions outside current means of compliance and relying on architectural mitigation around the learned component (EASA, 2026).

7.3. Legal accountability in the absence of certification

Where engineering assurance is absent, liability fills the vacuum, and early case law shows the direction of travel. The *Moffatt* tribunal's reasoning is the significant part: the airline was responsible for all information on its digital properties, however generated, and could not delegate epistemic responsibility to a stochastic component (Moffatt v. Air Canada, 2024). *Mata* established that professional duties of verification survive intact when the fabricating tool is an LLM (Mata v. Avianca, 2023). At the regulatory layer, the EU AI Act's Article 15 accuracy-declaration obligation will force quantitative claims about high-risk systems into legally reviewable form (European Parliament and Council, 2024), while NIST's framework records that the measurement science for such declarations is immature (NIST, 2023). Legal scholarship has begun to map the residual gap under the heading of "careless speech," noting that no existing doctrine imposes on model providers a general duty that their systems tell the truth (Wachter et al., 2024). The combined picture is a regime in which deployers are already being held answerable for outputs that *no available process can certify*, which is precisely the tension this review diagnoses.

8. Discussion

8.1. The epistemic dimension: plausibility is not assurance

The technical gap has an epistemological root, and the philosophical literature makes it precise in a way engineers should find congenial rather than ornamental. Frankfurt's category of *bullshit* denotes speech produced with indifference to truth, as distinct from lying, which requires tracking truth in order to negate it. Hicks, Humphries, and Slater argue on this basis that LLM output is bullshit in the technical sense: the generation process optimizes plausibility under a distribution, and any relation to truth is incidental to that objective (Hicks et al., 2024). The linguistic argument reaches the same place from the other side: a system trained only on form has no access to meaning (Bender and Koller, 2020), and scaling the training corpus manufactures fluency faster than it manufactures grounding (Bender et al., 2021). Whether some form of understanding nevertheless emerges at scale remains genuinely contested among serious researchers (Mitchell and Krakauer, 2023), and Shanahan's admonition to describe these systems in the exact vocabulary of what they do, next-token prediction under a learned distribution, is a piece of engineering hygiene as much as philosophy of mind (Shanahan, 2024).

The connection to certification is direct. A safety case is an *epistemic artifact*: a structured argument that transfers justified confidence from evidence about the artifact to a claim about its future behavior. Every mechanism in Table 1 (traceability, coverage, quantified targets) exists to make that transfer of warrant auditable. The philosophical results say that for LLM output the first link is missing: there is no truth-tracking mechanism inside the generator for evidence to attach to, so confidence in an individual output can only be established *outside* the model, by independent verification. Work on trust makes the same point in its own vocabulary: LLMs combine algorithmic opacity with a phenomenological transparency that hides the mediating system from the user, a combination that defeats ordinary calibration of trust (Heersmink et al., 2024); the epistemology of chatbot testimony concludes that believing a chatbot is not relevantly like believing a person or reading an instrument (Magnus, 2025). "Mostly right" is a perfectly good property for a drafting aid. As a load-bearing property in an assurance argument, it is a category error: assurance quantifies over *all* demands in the operational domain, and the mechanisms that would justify that quantification are absent by construction.

8.2 The structural diagnosis

Assembling Sections 4 through 7 yields a requirement-by-property analysis, presented in Table 4. The point of the table is that the two columns conflict *pairwise and independently*: fixing any one row leaves the others standing, which is what distinguishes a structural mismatch from an engineering backlog.

Table 4. Certification requirements versus documented LLM properties.

Certification requirement	Documented LLM property	Basis (this review)
Reproducible behavior traceable to reviewed requirements	Sampled generation, Eq. (1)-(2); greedy decoding restores repeatability but not specified correctness	§2.2
Complete specification of intended function	Open-ended natural-language input/output space; ~40 % of assessed ISO 26262 Part 6 unit-level techniques inapplicable to ML	§4.3; Salay et al. (2018)
Requirement-to-implementation traceability	Behavior realized in undifferentiated weights; multi-level opacity	§6.3; Heersmink et al. (2024)
Quantified worst-case failure bound	Theorem-level error floor (Eq. 3 - 4); measured rates 10^{-2} - 10^{0}	§4.1; §5
Bounded, frozen configuration under assessment	Freezing and versioning are possible but forfeit the improvement path; deployed practice tracks continuously updated models and APIs; PCCP-style envelopes presuppose a bounded, pre-specified change space	§7.1
Verification evidence about the artifact's structure (e.g., MC/DC)	Structural coverage undefined for attention over 10^{9+} parameters; formal tools verify local robustness at far smaller scale	§6.3

8.3. Anticipated objections and responses

The table also clarifies why partial progress on any single row does not move the conclusion. A hypothetical decoding scheme that made sampling deterministic would leave the specification, traceability, and error-floor rows untouched; a future formal-verification breakthrough at scale would verify robustness properties of a model that still cannot be specified against an open input domain; and a frozen, versioned model would still carry the theorem-level floor of Section 4.1. Structural, in the sense intended here, means exactly this pairwise independence of the obstructions.

"The title overclaims: 'cannot' is doing too much work." The claim is scoped in Section 1.2: current autoregressive LLMs, inside the certified boundary, under quantified certification regimes. Within that scope, "cannot" is the correct modality, because the obstruction rests on impossibility results (Section 4.1) and on the specification problem (Section 4.3), not on a performance shortfall. Outside that scope the review itself charts what "can" looks like (Section 8.4).

"Equation (5) assumes independent steps; real pipelines have error correction." Independence is a baseline, not a hidden premise. The measured direction of dependence is adverse: models condition on their own earlier errors (Sinha et al., 2026), and the imitation-learning analysis shows why deployment on self-generated state distributions is the unfavorable regime (Ross et al., 2011). Where explicit correction layers exist, they are the mitigations of Section 6, whose measured residuals are in Table 3; and the deflationary reading of the compounding argument (that it presupposes a next-token predictor that already errs at some

small per-token rate, and that the failure might be repaired by backtracking or verification wrappers) is due to Bachmann and Nagarajan (2024) and is cited here rather than suppressed. The empirical agent results of Section 5.2 bypass the modeling dispute entirely: end-to-end measured reliability is low, whatever the correct error model.

"Per-response error rates and per-hour failure targets are incommensurable." Correct, and acknowledged wherever the comparison appears (Figure 1 caption, Section 5.3): benchmark error rates are not convertible to PFD_{avg} or PFH without a system-level hazard, exposure, and demand-rate model. The argument does not rest on a conversion. It rests on an absence of evidence: no published measurement, under any mapping its authors defend, shows an LLM component meeting a quantified dangerous-failure target, while nothing in the published record suggests proximity to any such target. A unit dispute relocates the gap; it does not produce the missing evidence.

"The argument leans on preprints." The spine does not. Hallucination inevitability is peer-reviewed at STOC (Kalai and Vempala, 2024) and in *Nature* (Kalai et al., 2026); detection limits are in *Nature* (Farquhar et al., 2024); the legal and medical error audits are in *Journal of Legal Analysis*, *Journal of Empirical Legal Studies*, and *JMIR* (Dahl et al., 2024; Magesh et al., 2025; Chelli et al., 2024); the compounding bound is in the peer-reviewed imitation-learning literature (Ross et al., 2011); the standards analyses are in *SAE International Journal of Aerospace* and peer-reviewed venues (Kaakai et al., 2022; Dmitriev et al., 2023). Preprints (Xu et al., 2024; Vidot et al., 2021) are flagged at each citation and carry corroborating, not load-bearing, weight; the remaining theoretical, empirical, and standards anchors are peer-reviewed publications or official documents.

"Humans are not certified either, yet we let them fly aircraft." Certification regimes govern *systems*; human fallibility is managed by a disjoint apparatus of licensing, supervision, and liability that presupposes accountable agents with duties of care. The moment an LLM is embedded in software inside the certified boundary, it is governed by software assurance, and the early case law of Section 7.3 indicates that existing professional and organizational duties remain applicable to those outputs, on the facts of the cases decided so far, rather than excusing them as quasi-human error.

"Narrow ML components have been certified; LLMs will follow the same path." The strongest published assurance case in the corpus points the other way. The runway-sign classifier study presented a DAL C-certifiable architecture built from individually untrusted networks under a conventional monitor, inside a closed ODD with a frozen model (Dmitriev et al., 2023). Every element of that recipe (closed ODD, frozen weights, external deterministic check) is exactly what open-ended generation lacks by definition; what generalizes from the precedent is the envelope, not the model-side assurance.

"Models keep improving; the gap will close on its own." The improvement trend is real and measured, and it is a horizon trend, not an error-floor trend: doubling the 50 %-success task length every seven months (Kwa et al., 2025) says nothing about driving per-demand dangerous-failure probability through six orders of magnitude, and model-parameter scaling alone does not evade the theorems of Section 4.1 under their stated assumptions, though the magnitude of the error floor depends on the data distribution. Extrapolating the measured curves reaches "an agent that can attempt a week-long task" long before it reaches "a component with a demonstrable 10^{-5} failure bound."

8.4 Two coherent exits

If the diagnosis is structural, coherent responses change one of the two structures. **Exit A: the standards move.** Certification could develop explicitly statistical acceptance criteria for machine-learned components: population-level bounds over declared ODDs, conformal-style guarantees with audited calibration data, and abstention budgets treated as a first-class failure mode. ISO/PAS 8800, UL 4600's safety cases, and the EASA

W-process are early movements in this direction (Section 7.1), and the open-rubric evaluation reform proposed in the *Nature* hallucination paper is the benchmark-layer counterpart (Kalai et al., 2026). The limit of Exit A is already visible in its instances: statistical acceptance works where the ODD closes (perception over a defined sensor domain) and has no purchase on open-ended generation, where the population over which the statistics would range cannot be characterized (Section 4.3).

Exit B: the architectures move. The convergent pattern across the strongest results in this corpus is to stop asking the model to be trustworthy and to build a deterministic envelope that is. The components are all published: the model as untrusted proposer with a symbolic or programmatic checker holding the guarantee (d'Avila Garcez and Lamb, 2023); conformal filtering supplying calibrated abstention (Mohri and Hashimoto, 2024); semantic-entropy monitors as runtime detectors (Farquhar et al., 2024); programmable rails as the engineering substrate (Rebedea et al., 2023); dissimilar redundancy under a conventional monitor as the published assurance-case precedent (Dmitriev et al., 2023); and the W-shaped process as the assurance wrapper for whatever narrow learned components remain (EASA, 2024). In this architecture the certified object is the envelope: its interlocks, its checkers, its fallback channels are conventional software amenable to conventional assurance, and the LLM inside is formally a source of *suggestions* whose acceptance is gated by verifiable logic. The division of labor mirrors the epistemology of Section 8.1: confidence is manufactured outside the generator, where truth-tracking mechanisms exist.

9. Limitations of This Review

Four limitations bound the conclusions. First, this is a narrative review with systematic elements, not a registered systematic review: the seven search directions were chosen analytically, no PRISMA flow accounting was performed, and selection bias cannot be excluded with systematic-review rigor, although the verification protocol of Section 3.2 controls the complementary risk of citation error. Second, the field moves quickly: the corpus closes on 3 August 2026, the agent benchmarks of Section 5.2 date from 2024 - 2025, and specific numbers will age, though the theoretical results of Section 4.1 and the structural analysis of Section 8.2 do not depend on them. Third, the empirical studies aggregated in Table 2 use heterogeneous definitions of hallucination and error, which is why this review synthesizes them qualitatively, by order of magnitude, rather than by meta-analysis; the order-of-magnitude framing is robust to definitional variance, but finer comparisons are not licensed. Fourth, the corpus is English-language, and certification practice in other jurisdictions may differ in ways the reviewed documents do not capture.

10. Future Research Directions

Five directions follow from the diagnosis. (1) *Specification languages for language tasks*: formal notations that can close an operational design domain over a constrained subset of natural-language interaction, making Exit A's statistics well-posed for at least bounded conversational functions. (2) *Verified constrained decoding*: decoding algorithms whose outputs provably remain within a machine-checkable grammar or logic, moving part of the guarantee inside generation itself and reducing the envelope's rejection load. (3) *Assurance-grade measurement protocols*: standardized, adversarially audited procedures for estimating per-demand error rates of LLM components with confidence bounds a regulator could accept, the missing measurement science that NIST identifies (NIST, 2023). (4) *Envelope engineering as a discipline*: reference architectures, patterns, and worked certification cases for deterministic envelopes around probabilistic proposers, generalizing the dissimilar-redundancy case study (Dmitriev et al., 2023) beyond perception. (5) *Abstention-first evaluation*: adoption of explicit-rubric benchmarks that reward calibrated refusal (Kalai et al., 2026), so that the deployed population of models is at least trained toward the behavior that safety architectures need at their interfaces.

11. Conclusion

This review set out to determine whether large language models can be certified for safety-critical systems and answered in the negative, on structural grounds assembled entirely from the published record. The certification regimes that govern safety-critical software impose differing combinations of system-level risk targets and specification, traceability, configuration-control, and evidence obligations (Section 2, Table 1). Against each demand, the literature supplies a documented obstruction: hallucination floors established as theorems under stated conditions (Section 4.1), an input space that defeats specification completeness (Section 4.3), exponential erosion of reliability over execution horizons (Section 4.2, Figure 1), reported reliability indicators that no published hazard-and-exposure mapping connects to a certification target (Section 5, Table 2), mitigations that reduce without bounding (Section 6, Table 3), and an epistemic character of generated text that severs the warrant transfer on which assurance cases run (Section 8.1). The standards bodies' own instruments, read closely, corroborate the diagnosis: every AI-aware extension presupposes the identified, version-controlled artifact that open-ended generative practice does not offer, and the strongest published assurance case in the corpus succeeded by architecturally distrusting its learned components.

The constructive conclusion is symmetrical. Where the operational domain can genuinely be closed, standards can and should develop genuine statistical acceptance criteria. Where it cannot, the engineering pattern that the strongest published results converge on is the deterministic envelope: certify the verifiable structure that surrounds the model, gate the model's proposals through logic that can be traced and tested, budget abstention as a failure mode, and let the language model do what a probabilistic instrument does well, generate candidates, while never holding the safety claim itself. On the published evidence, that division of labor is the arrangement for which a defensible assurance route has so far been demonstrated, and the one most consistent simultaneously with the mathematics of probabilistic generation and the grammar of certification.

Declarations

Funding. The author received no external funding for this work.

Conflicts of interest. The author declares no conflicts of interest.

Data availability. All sources reviewed are publicly available at the locations cited in the References; no new data were generated.

Corresponding author. Akbar Sayakov, Base80.ai, San Francisco, CA, USA. Email: [to be inserted].

References

1. Bachmann, G., & Nagarajan, V. (2024). The pitfalls of next-token prediction. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 2296 - 2318.

2. Banerjee, S., Agarwal, A., & Singla, S. (2025). LLMs will always hallucinate, and we need to live with this. In Lecture Notes in Networks and Systems. Springer. https://doi.org/10.1007/978-3-031-99965-9_39

3. Béchard, P., & Marquez Ayala, O. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track) (pp. 228 - 238). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-industry.19

4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (pp. 610 - 623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

5. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185 - 5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

6. Brix, C., Müller, M. N., Bak, S., Johnson, T. T., & Liu, C. (2023). First three years of the international verification of neural networks competition (VNN-COMP). International Journal on Software Tools for Technology Transfer, 25(3), 329 - 339. https://doi.org/10.1007/s10009-023-00703-4

7. Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. Journal of Medical Internet Research, 26, e53164. https://doi.org/10.2196/53164

8. Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. Journal of Legal Analysis, 16(1), 64 - 93. https://doi.org/10.1093/jla/laae003

9. d'Avila Garcez, A., & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. Artificial Intelligence Review, 56(11), 12387 - 12406. https://doi.org/10.1007/s10462-023-10448-w

10. Dmitriev, K., Schumann, J., Bostanov, I., Abdelhamid, M., & Holzapfel, F. (2023). Runway sign classifier: A DAL C certifiable machine learning system. In 2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC). IEEE. https://doi.org/10.1109/DASC58513.2023.10311228

11. Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Le Bras, R., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2305.18654

12. European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

13. European Union Aviation Safety Agency. (2023). EASA Artificial Intelligence Roadmap 2.0: A human-centric approach to AI in aviation (Version 2.0). https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-roadmap-20

14. European Union Aviation Safety Agency. (2024, March). EASA concept paper: Guidance for Level 1 & 2 machine learning applications (Issue 02). https://www.easa.europa.eu/en/downloads/139504/en

15. European Union Aviation Safety Agency. (2026). EASA concept paper: Guidance for artificial intelligence applications (Proposed Issue 03, public consultation). https://www.easa.europa.eu/sites/default/files/dfu/easa_concept_paper_guidance_for_artificial_intelligence_applications_proposed_issue_03.pdf

16. European Union Aviation Safety Agency, & Daedalean AG. (2020). Concepts of design assurance for neural networks (CoDANN) (Public report extract, Innovation Partnership Contract P-EASA.IPC.004, Version 1.0). European Union Aviation Safety Agency. https://www.easa.europa.eu/en/document-library/general-publications/concepts-design-assurance-neural-networks-codann

17. European Union Aviation Safety Agency, & Daedalean AG. (2024). Concepts of design assurance for neural networks (CoDANN) II: Public extract (Version 1.1, with Appendix B; original report published May 2021). https://www.easa.europa.eu/en/document-library/general-publications/concepts-design-assurance-neural-networks-codann-ii

18. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625 - 630. https://doi.org/10.1038/s41586-024-07421-0

19. Federal Aviation Administration. (2024). Roadmap for artificial intelligence safety assurance (Version I). U.S. Department of Transportation. https://www.faa.gov/media/82891

20. Glukhov, D., Shumailov, I., Gal, Y., Papernot, N., & Papyan, V. (2024). Position: Fundamental limitations of LLM censorship necessitate new approaches. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 15767 - 15787. https://proceedings.mlr.press/v235/glukhov24a.html

21. Heersmink, R., de Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. Ethics and Information Technology, 26(3), Article 41. https://doi.org/10.1007/s10676-024-09777-3

22. Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. Ethics and Information Technology, 26, Article 38. https://doi.org/10.1007/s10676-024-09775-5

23. 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. https://doi.org/10.1145/3703155

24. International Electrotechnical Commission. (2010). Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 1: General requirements (IEC Standard No. 61508-1:2010, Ed. 2.0). https://webstore.iec.ch/en/publication/5515

25. International Organization for Standardization. (2018). Road vehicles - Functional safety - Part 1: Vocabulary (ISO Standard No. 26262-1:2018). https://www.iso.org/standard/68383.html

26. International Organization for Standardization. (2024). Road vehicles - Safety and artificial intelligence (ISO/PAS 8800:2024). https://www.iso.org/standard/83303.html

27. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

28. Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? In The Twelfth International Conference on Learning Representations (ICLR 2024). https://arxiv.org/abs/2310.06770

29. Kaakai, F., Dmitriev, K., Adibhatla, S., Baskaya, E., Bezzecchi, E., Bharadwaj, R., Brown, B., Gentile, G., Gingins, C., Grihon, S., & Travers, C. (2022). Toward a machine learning development lifecycle for product certification and approval in aviation. SAE International Journal of Aerospace, 15(2), 127 - 143. https://doi.org/10.4271/01-15-02-0009

30. Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2026). Evaluating large language models for accuracy incentivizes hallucinations. Nature, 653(8116), 1047 - 1051. https://doi.org/10.1038/s41586-026-10549-w

31. Kalai, A. T., & Vempala, S. S. (2024). Calibrated language models must hallucinate. In Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC 2024) (pp. 160 - 171). Association for Computing Machinery. https://doi.org/10.1145/3618260.3649777

32. Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. In R. Majumdar & V. Kunčak (Eds.), Computer aided verification: CAV 2017 (Lecture Notes in Computer Science, Vol. 10426, pp. 97 - 117). Springer. https://doi.org/10.1007/978-3-319-63387-9_5

33. Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., et al. (2025). Measuring AI ability to complete long software tasks. Advances in Neural Information Processing Systems 38 (NeurIPS 2025). https://proceedings.neurips.cc/paper_files/paper/2025/hash/85069585133c4c168c865e65d72e9775-Abstract-Conference.html

34. Luettig, B., Akhiat, Y., & Daw, Z. (2024). ML meets aerospace: Challenges of certifying airborne AI. Frontiers in Aerospace Engineering, 3, Article 1475139. https://doi.org/10.3389/fpace.2024.1475139

35. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies, 22(2), 216 - 242. https://doi.org/10.1111/jels.12413

36. Magnus, P. D. (2025). On trusting chatbots. Episteme, 22(4), 906 - 916. https://doi.org/10.1017/epi.2024.29

37. Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023). https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/

38. McGregor, S. (2021). Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15458 - 15463. https://doi.org/10.1609/aaai.v35i17.17817

39. Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. Proceedings of the National Academy of Sciences, 120(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

40. Moffatt v. Air Canada, 2024 BCCRT 149 (Can. B.C. Civ. Resol. Trib. Feb. 14, 2024). https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html

41. Mohri, C., & Hashimoto, T. (2024). Language models with conformal factuality guarantees. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 36029 - 36047. https://proceedings.mlr.press/v235/mohri24a.html

42. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

43. Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C., & Cohen, J. (2023). NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In Y. Feng & E. Lefever (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 431 - 445). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-demo.40

44. Ross, S., Gordon, G. J., & Bagnell, J. A. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011), Proceedings of Machine Learning Research, 15, 627 - 635. https://proceedings.mlr.press/v15/ross11a.html

45. RTCA. (2011). Software considerations in airborne systems and equipment certification (RTCA DO-178C). RTCA, Inc.

46. Salay, R., Queiroz, R., & Czarnecki, K. (2018). An analysis of ISO 26262: Machine learning and safety in automotive software (SAE Technical Paper 2018-01-1075). SAE International. https://doi.org/10.4271/2018-01-1075

47. Shanahan, M. (2024). Talking about large language models. Communications of the ACM, 67(2), 68 - 79. https://doi.org/10.1145/3624724

48. Sinha, A., Arun, A., Goel, S., Staab, S., & Geiping, J. (2026). The illusion of diminishing returns: Measuring long horizon execution in LLMs. International Conference on Learning Representations (ICLR 2026). https://doi.org/10.48550/arXiv.2509.09677

49. UL Standards & Engagement. (2023). ANSI/UL 4600: Standard for evaluation of autonomous products (3rd ed.). https://www.shopulstandards.com/ProductDetail.aspx?productId=UL4600_3_S_20230317

50. U.S. Food and Drug Administration. (2024). Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Final guidance (issued December 2024; reissued August 18, 2025). U.S. Department of Health and Human Services. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence

51. Vidot, G., Gabreau, C., Ober, I., & Ober, I. (2021). Certification of embedded systems based on machine learning: A survey (arXiv:2106.07221). arXiv. https://doi.org/10.48550/arXiv.2106.07221

52. Wachter, S., Mittelstadt, B., & Russell, C. (2024). Do large language models have a legal duty to tell the truth? Royal Society Open Science, 11(8), Article 240197. https://doi.org/10.1098/rsos.240197

53. Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv. https://doi.org/10.48550/arXiv.2401.11817

54. Yao, S., Shinn, N., Razavi, P., & Narasimhan, K. (2025). τ-bench: A benchmark for tool-agent-user interaction in real-world domains. International Conference on Learning Representations (ICLR 2025). https://doi.org/10.48550/arXiv.2406.12045

55. Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., & Neubig, G. (2024). WebArena: A realistic web environment for building autonomous agents. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). https://arxiv.org/abs/2307.13854