AI Reliability Gap: Why Large Language Models Fail in Safety-Critical Systems

March 2026

Praneeth Vadlapati
Independent Researcher
praneethv@arizona.edu
ORCID: 0009-0006-2592-2564

Abstract:

Large Language Models (LLMs) have transitioned rapidly from research contexts into deployment across high-stakes domains, including healthcare, defense, and autonomous decision-making systems. Despite impressive performance on standardized benchmarks, a growing body of empirical evidence indicates that LLMs do not yet satisfy the reliability standards required for critical applications. This paper addresses a consequential gap in the existing literature that no prior work systematically defines operational reliability for LLMs in critical deployment contexts or provides a structured framework for analyzing how reliability failures manifest across distinct failure classes. To address this gap, this paper introduces the LLM Operational Reliability Failure Taxonomy (ORFT), a framework comprising eight empirically grounded failure classes. It is applied to examine the structural gap between measured performance and operational reliability. It further argues that current evaluation methodologies have not yet demonstrated the capacity to capture reliability in critical real-world operational conditions with direct consequences for human safety and economic integrity. The analysis leads to the conclusion that frontier AI systems have not yet achieved the reliability standards required for autonomous deployment in life-critical or mission-critical environments and identifies the research and governance investments necessary to advance this trajectory.

Keywords: Large Language Models, LLMs, AI Reliability, Critical Applications, AI in Defense, AI in Healthcare, AI Benchmarks, Hallucination, Robustness, Artificial Intelligence, Generative AI

1. Introduction

The past several years have witnessed the rapid scaling of Large Language Models into general-purpose AI systems capable of performing tasks across science, engineering, medicine, logical reasoning, and creative work [1], [2]. Prevailing development strategies, which include scaling model size, increasing training data volume, and applying reinforcement learning from human feedback, have produced models that exhibit substantial competence on standardized benchmarks [3]. Governments, defense agencies, and healthcare institutions have commenced deployment of, or are actively evaluating, AI systems for consequential tasks, including clinical decision support, intelligence analysis, and autonomous mission planning [4], [5].

Yet a foundational question has received insufficient systematic attention about whether LLMs are operationally reliable. Performance and reliability are distinct properties. A system may be capable since it can produce correct answers under controlled evaluation conditions, while being unreliable since its failure distribution is unpredictable, non-stationary, and disproportionately concentrated in precisely the conditions where reliability matters most. Recent empirical work has demonstrated that larger, more instruction-tuned models do not necessarily secure low-difficulty areas, and that scaled-up models tend to produce apparently plausible yet incorrect answers more frequently than their predecessors, including on complex questions that human supervisors frequently overlook [3]. This constitutes a reliability failure, not merely a performance limitation.

This distinction carries significant consequences in critical applications. In a clinical setting, a model that achieves ninety-five percent accuracy on benchmark questions but produces confident, incorrect answers in the remaining cases without any explicit indication of uncertainty may cause direct patient harm. In a defense context, an agentic AI system that performs adequately under standard conditions but that exhibits reasoning collapse, adversarial manipulation, or temporal drift under operational stress poses qualitative risks that extend well beyond simple inaccuracy [4], [6]. The International AI Safety Report 2026, produced by more than 100 independent experts across 30 countries, explicitly identifies that AI agents and multi-agent systems remain prone to basic errors and that reliability problems are compounded when human oversight is limited [6].

Recent surveys of LLMs offer comprehensive reviews of architecture, fine-tuning, benchmarking, and application domains [1], [2]. These surveys provide a strong foundation for understanding LLM capabilities. However, a critical gap persists in the literature that no existing work systematically defines operational reliability for LLMs in critical deployment contexts, distinguishes it from adjacent properties such as alignment, safety, and robustness, or provides a structured framework for analyzing how reliability failures manifest across distinct failure classes in high-stakes domains. Reliability as a primary analytical property, together with the question of whether existing models have demonstrated the predictable failure distributions required for safe deployment, remains an area that has not yet received the dedicated, systematic treatment that the stakes of deployment demand. This gap in the literature motivates the present study and the analytical framework it introduces.

This paper makes four principal contributions. First, it provides a precise operational definition of LLM reliability, formalized as the probability that a system performs its required function without failure over a specified number of future tasks under a given operational profile, rigorously distinguished from the adjacent properties of alignment, safety, and robustness, and grounded in established principles from software reliability engineering and domain-specific regulatory standards. Second, it introduces the LLM Operational Reliability Failure Taxonomy (ORFT), a structured analytical framework comprising eight failure classes systematically derived from prior work in LLM risk classification [20], [28] and from evidence in recent empirical literature. ORFT is designed to provide researchers, engineers, and regulators with a principled vocabulary and analytical structure for characterizing, assessing, and addressing reliability failures in critical LLM deployments. It is further intended as an extensible foundation that future work may refine, expand, or adapt to emerging deployment contexts. Third, it applies the ORFT framework to benchmark analysis, examining why standard evaluation practices, and in particular multiple-choice question formats, have not demonstrated the capacity to measure operational reliability in healthcare and defense settings, and identifying the structural limitations that reliability-oriented evaluation must overcome. Fourth, it derives governance implications from the preceding analysis, examining why current mitigation approaches have not yet closed the reliability gap, and concluding with research and regulatory recommendations to enable reliable LLM deployment in critical applications.

The domains examined in this paper, healthcare and defense, are chosen because they represent the clearest cases in which reliability failures produce irreversible consequences, and because active deployment is already underway despite the absence of demonstrated operational reliability [4], [5], [6].

2. Defining Reliability in the LLM Context

The literature on LLMs addresses several distinct properties that, for critical deployment, benefit from careful disambiguation. The following definitions are adopted for this study.

Alignment refers to the degree to which a model's outputs correspond to human intent across the full distribution of inputs the system will encounter [6]. Safety refers to the model's tendency to avoid outputs that are harmful, dangerous, or unethical, including outputs that might arise from misuse or adversarial

prompting [6]. Robustness refers to the model's resistance to performance degradation under perturbations, such as changes in phrasing, distributional shift, noisy inputs, or adversarial manipulation [7], [8]. Reliability refers to the probability that a system performs its required function without failure over a specified number of future tasks under a given operational profile [9].

The last definition, drawn from software reliability engineering and formalized in the HIP-LLM framework, is the most demanding of these properties [9]. Reliability is not a property of aggregate benchmark performance. Instead, it is a property of the distribution of failures across the operational context in which a system is deployed. A model that achieves ninety percent accuracy on a benchmark has not thereby demonstrated any particular level of reliability, because benchmark accuracy does not constrain where within the input distribution the remaining ten percent of failures occur, how confident the model is when it fails, or whether failures are concentrated in the most critical regions of the operational domain.

This distinction between statistical accuracy and operational reliability is central to the analysis presented in this paper. Critical domains such as defense and medicine impose regulatory and engineering standards specifically because they require predictable failure distributions, not merely high average performance. Standards such as MIL-STD-882 for system safety in defense and FDA requirements for Software as a Medical Device are premised on the ability to characterize, contain, and predict failure modes. Large Language Models, as probabilistic next-token predictors, do not yet satisfy these requirements [4], [6].

3. Background: LLM Performance and the Reliability Gap

3.1 The Promise of LLMs in Critical Domains

Large Language Models have demonstrated notable performance across a range of tasks relevant to critical domains. In healthcare, models have achieved passing scores on the United States Medical Licensing Examination and exhibited preliminary efficacy in radiology reporting, clinical note summarization, and differential diagnosis support [1]. In defense and intelligence analysis, LLMs are being evaluated for their capacity to synthesize large volumes of unstructured data, support decision-making under time pressure, and operate as components within autonomous systems [4], [5]. These demonstrated capabilities have accelerated both interest in and institutional adoption of LLM-based systems across high-stakes sectors.

3.2 The Reliability Gap

Despite these capabilities, recent empirical work consistently reveals a gap between benchmark performance and operational reliability. The following findings are particularly germane to this analysis.

Larger, more instruction-tuned models are more likely to produce confident, plausible-sounding, but factually incorrect answers than smaller models, and these errors are more difficult for human supervisors to detect and correct [3]. Ranking platforms for LLMs are susceptible to minor changes in the underlying crowdsourced evaluation data, casting uncertainty on the stability of comparative performance assessments [10]. Models incorporate clinically irrelevant information, such as typographical errors, inconsistent formatting, and informal language, into medical treatment recommendations, leading to errors that raise patient safety concerns in clinical contexts [11]. AI-based therapy tools have shown performance gaps with respect to patients with severe psychiatric conditions and have not consistently responded appropriately to suicidal ideation, with these patterns persisting across model scales and generations [12]. AI agents operating within multi-agent architectures exhibit reliability and control challenges that are qualitatively distinct from those observed in single-model deployments [6].

These findings collectively indicate a consistent pattern in which performance on benchmark tasks does not reliably predict performance in operational contexts, and the failure modes that emerge in deployment are precisely those with the most significant consequences in critical applications.

3.3 Distinguishing Benchmark Performance from Operational Reliability

A central argument of this paper is that existing LLM evaluations measure performance, which is the ability to produce correct outputs under favorable and controlled conditions, rather than reliability, which is the probability of correct operation across the full operational distribution, including adversarial and degraded conditions.

This gap has two structural causes. First, benchmark datasets are primarily drawn from the same distribution as training data, which means that high scores may reflect memorization or surface-level pattern matching rather than genuine generalization to novel inputs [7], [13]. Second, the dominant evaluation format, which is the multiple-choice question, has not been shown to adequately assess the open-ended, naturalistic reasoning required in critical applications. Studies in medical benchmarking have found that LLMs perform substantially better on multiple-choice formats than on free-response versions of equivalent questions, with average performance declines exceeding thirty-nine percentage points on open-ended equivalents. Moreover, models achieve above-chance accuracy even when the question stem is fully masked, suggesting that benchmark scores may reflect format sensitivity and surface-level pattern matching rather than verified substantive medical knowledge [14].

The LLM Risk Assessment Framework addresses this problem in the context of systems engineering by classifying LLM-based applications along dimensions of autonomy and impact, enabling organizations to determine appropriate validation strategies and required levels of human oversight for each deployment context [15]. However, such frameworks remain at an early stage of development and have not yet been systematically applied in healthcare or defense settings.

4. A Taxonomy of LLM Failure Classes in Critical Applications

The LLM Operational Reliability Failure Taxonomy (ORFT) introduced in this paper comprises eight failure classes, derived from prior work on LLM risk taxonomies [20], [28], and examined here as particularly consequential in critical deployment contexts. These classes are not mutually exclusive. Real-world failures frequently involve combinations of multiple classes occurring simultaneously.

4.1 Epistemic Hallucination

Large Language Models generate factually incorrect content with apparent fluency and apparent confidence [16], [17]. Hallucination is not an anomalous occurrence but a systematic property of probabilistic text generation. In critical domains, hallucinated facts, including incorrect drug dosages, fabricated intelligence assessments, and non-existent legal citations, can lead directly to harmful decisions. This risk is compounded by the fact that hallucinated outputs are frequently indistinguishable from correct outputs, making human verification difficult even for domain experts [16].

4.2 Overconfidence Failure

Large Language Models frequently assign high implicit confidence to incorrect outputs [3], [9]. This failure class is particularly consequential in critical contexts because it can induce over-reliance by human operators. Research on AI overreliance demonstrates that when AI systems present outputs with apparent confidence, whether through tone, formatting, or explicit numerical estimates, human supervisors frequently accept them without adequate scrutiny, even when they possess domain expertise that would

otherwise allow them to detect the error [18]. This creates a failure mode in which the AI system does not merely produce an error but may contribute to the human responsible for oversight accepting an incorrect output.

4.3 Abstention Failure

Large Language Models sometimes decline to answer questions within their legitimate operational scope, including sensitive but necessary topics in clinical or military contexts [3]. Abstention is an appropriate behavior when a model genuinely lacks the information required to answer reliably. It becomes a failure when the model abstains in conditions where a correct answer is both needed and available. Notably, scaled-up and instruction-tuned models are less likely to refuse questions outright but more likely to substitute a confident incorrect answer rather than an appropriate refusal [3], a shift that presents greater reliability challenges than outright refusal would in safety-critical contexts.

4.4 Prompt Fragility

Small changes in the phrasing of a question can produce substantially different model outputs, even when the semantic content of the two phrasings is equivalent [7], [8], [11]. Studies of LLM robustness to paraphrased benchmark questions find that while model rankings remain relatively stable across paraphrasings, absolute performance scores decline significantly [8]. In operational settings, where prompts are produced by users with varying communication styles, levels of domain expertise, and language backgrounds, this fragility produces a system whose outputs cannot be reliably anticipated. An MIT study found that nonclinical features of patient messages, including typographical errors, excess whitespace, and informal language, reduced the accuracy of LLM medical recommendations in ways that would be clinically unacceptable [11].

4.5 Temporal Drift

LLM performance does not constitute a fixed or stable property. Model accuracy changes over time as developers apply fine-tuning updates, modify system prompts, adjust guardrail configurations, or release successor model versions [6]. In a deployment context where a system must continuously satisfy a specified reliability threshold, temporal drift can cause a validated system to fall below the required threshold without warning or documentation. The rapid pace of model releases compounds this problem, the absence of standardized change documentation practices, and the lack of continuous post-deployment monitoring infrastructure in most current LLM deployments.

4.6 Reasoning Collapse

Under conditions that require extended chains of inference, including long contexts, multi-step planning tasks, and complex logical structures, LLMs can produce incoherent, repetitive, or truncated reasoning outputs [19]. This constitutes a class of robustness failures characterized by inconsistent performance across minor variations in task presentation [19]. In real-time operational contexts with latency constraints, reasoning collapse may manifest as timeouts, incomplete responses, or outputs that appear syntactically coherent but are logically disconnected from the input. This failure class is particularly consequential in agentic deployments where the model's reasoning output directly drives downstream actions.

4.7 Agentic Escalation

When LLMs are deployed as autonomous agents with access to external tools, such as web browsers, application programming interfaces, code execution environments, or physical actuators, errors can produce irreversible consequences [6]. A single incorrect action by an LLM agent may trigger downstream effects in connected systems that are irreversible. The International AI Safety Report 2026 notes that AI agents and multi-agent systems exhibit reliability and control problems that do not arise in single-model

deployments, and that human oversight of these systems is simultaneously more critical and more difficult to maintain [6]. In defense applications, the potential for agentic AI to trigger lethal kinetic effects makes this failure class categorically distinct from errors in purely advisory systems.

4.8 Adversarial Manipulation

Large Language Models are vulnerable to prompt injection attacks, in which malicious inputs embedded in the operational environment, including documents, web pages, or user messages, cause the model to deviate from its intended instructions [4], [6], [20]. In defense and intelligence contexts, adversaries can deliberately craft inputs to manipulate LLM behavior, potentially causing the system to withhold critical information, generate false assessments, or execute unauthorized actions. This failure class is particularly resistant to mitigation because it exploits the same general-purpose natural language understanding that constitutes the primary functional basis of LLM utility.

Table 1. LLM Failure Classes in Critical Applications

Failure Class	Description	Critical Impact
Epistemic Hallucination	Fabrication of plausible but factually incorrect content	Erroneous medical or military decisions based on false information
Overconfidence Failure	High-confidence presentation of incorrect outputs	Human operators accept errors without scrutiny
Abstention Failure	Inappropriate refusal to respond when a response is needed	System paralysis at critical decision points
Prompt Fragility	Output instability across semantically equivalent phrasings	Unpredictable system behavior in diverse operational conditions
Temporal Drift	Performance changes across model versions and time	Undetected reliability regression following post-deployment updates
Reasoning Collapse	Incoherence, repetition, or truncation under complex reasoning demands	Failure during time-critical operational tasks
Agentic Escalation	Autonomous execution of unsafe or unintended actions	Irreversible operational or physical consequences
Adversarial Manipulation	Prompt injection causing deviation from intended instructions	Command hijacking by deliberate adversaries

5. Why Benchmarks Have Not Demonstrated Operational Reliability

5.1 The Structure of Standard Benchmarks

The dominant paradigm for LLM evaluation uses static datasets of multiple-choice questions to produce scalar accuracy scores [7], [13], [14]. Benchmarks such as MMLU, ARC-C, and HellaSwag are administered in standardized formats and scored against fixed ground-truth labels. This paradigm was designed to measure performance growth, and it has been effective for that purpose, as benchmark scores have reliably tracked improvements in LLM performance across successive model generations.

However, the design of these benchmarks has not been shown to measure reliability adequately. Reliability requires characterizing the failure distribution across the complete operational profile, which is the range of inputs a system will encounter in actual deployment. Static benchmarks sample from a fixed and pre-

specified distribution that does not capture the diversity of real-world operational inputs. They do not include adversarial inputs, degraded conditions, or low-frequency scenarios that occur rarely in aggregate but are frequent enough across high-volume deployments to produce significant harm [7].

5.2 Benchmark Saturation and Residual Failures

As model performance improves, benchmark scores approach ceiling levels and benchmarks are retired or replaced. This practice systematically discards evidence of residual failure, as errors that persist in saturated benchmarks are attributed to annotation noise rather than investigated as indicators of unreliable model behavior [21]. The concept of platinum benchmarks, which are evaluation datasets on which a reliable model should achieve perfect accuracy, allowing remaining errors to be unambiguously attributable to the model rather than the benchmark, has been proposed as a corrective [21]. Evaluations of current frontier models on carefully cleaned versions of existing benchmarks reveal that even state-of-the-art systems exhibit systematic failures on simple tasks. These failures are ordinarily obscured by the noise present in standard benchmark datasets [21].

5.3 The Multiple-Choice Format Problem in Clinical Domains

The problem of benchmark validity is particularly acute in medical AI evaluation, where multiple-choice question formats are the dominant assessment method. Performance on multiple-choice formats is substantially higher than on free-response versions of equivalent medical questions, with one study documenting an average absolute performance gap of 39.43 percentage points across three frontier models [14]. Models also perform above chance even when the question stem is fully masked, demonstrating that benchmark scores reflect format exploitation and surface-level pattern matching rather than substantive medical knowledge [14]. The majority of medical LLM evaluations have relied on multiple-choice datasets, such as MedQA, MedMCQA, and PubMedQA, which have not been shown to capture the open-ended, multi-step reasoning required in actual clinical practice [22], [23], [24].

5.4 Instability of LLM Ranking Platforms

The instability of LLM ranking platforms further undermines the evidentiary basis for reliability claims in practice. An MIT study found that removing a small fraction of the crowdsourced data underlying widely used online ranking platforms can significantly alter the resulting model rankings [10]. If the evaluation infrastructure used to determine which models are most capable is itself unstable, then deployment decisions based on those rankings have no reliable empirical foundation.

5.5 Unproven Reliability Under Real-World Input Conditions

A further structural limitation of existing benchmark evaluations is that they do not demonstrate reliable LLM performance under the input conditions routinely encountered in real-world critical deployments. Standard benchmarks present clean, well-formed, grammatically correct prompts in high-resource languages, typically English, and at moderate context lengths. Such conditions bear little correspondence to the heterogeneous inputs that operational systems must process. Empirical studies have demonstrated that even minor spelling errors and typographical variations in prompts degrade performance across multiple model families and task types, with absolute accuracy reductions of up to 13.6 percentage points observed for individual models on reasoning benchmarks [25]. Grammatical errors in user inputs introduce analogous fragility, though the severity varies by task and model [7]. Multilingual prompts present a distinct and compounding challenge as LLMs trained predominantly on English exhibit substantial performance disparities when queried in lower-resource languages, with leading models showing accuracy gaps exceeding 30 percentage points between high-resource and low-resource language conditions [26]. Context overload constitutes a further underexamined source of reliability failure. Research has established that model performance degrades systematically as input length increases, with information positioned in the

middle of long contexts disproportionately lost relative to content at the beginning or end of the context window, even in models explicitly designed for long-context operation [27]. Beyond these individually documented failure conditions, existing evaluations provide no evidence of reliable performance across the full combinatorial space of real-world scenarios, including mixed-language inputs, noisy or colloquial phrasing, domain-specific jargon, incomplete sentences, and many other input variations that arise naturally in clinical, defense, and other high-stakes environments. The absence of such evidence is not a minor gap in the evaluation literature. It represents a fundamental unresolved question about whether the performance observed in controlled benchmark settings generalizes to the conditions under which reliability actually matters.

6. Why Current Mitigations Have Not Yet Closed the Reliability Gap

Several technical approaches have been proposed to improve LLM reliability, including retrieval-augmented generation, guardrail systems, fine-tuning, and multi-agent oversight architectures. Each addresses a subset of the failure classes described above and constitutes a meaningful methodological contribution. However, each also leaves substantive reliability gaps unaddressed.

6.1 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) reduces epistemic hallucination by grounding model responses in retrieved external documents, thereby reducing the model's dependence on parametric memory [16]. However, RAG introduces distinct failure modes, including retrieval errors, document-level hallucination in which the model misrepresents the content of correctly retrieved documents, and sensitivity to the quality and currency of the retrieval corpus. The retrieval corpus itself changes over time, introducing temporal drift at the data layer. In dynamic operational environments where intelligence databases, clinical guidelines, and operational parameters change rapidly, maintaining a current, accurate retrieval corpus constitutes a significant and ongoing operational burden.

6.2 Guardrail Systems

Guardrail systems filter model inputs and outputs to prevent harmful, out-of-scope, or otherwise unacceptable responses. Guardrails address safety and alignment concerns more directly than reliability concerns. They do not reduce hallucination rates, prompt fragility, or reasoning collapse, as they intercept outputs after generation rather than improving the underlying generation process. Moreover, guardrail systems are themselves software artifacts that require validation, versioning, and maintenance. In rapidly evolving deployments, guardrail systems may fall out of alignment with model behavior following model updates, and their effectiveness against adversarial manipulation is limited, as sophisticated prompt injection attacks can cause models to satisfy guardrail constraints while violating operational intent [6], [20].

6.3 Fine-Tuning

Domain-specific fine-tuning improves average performance on targeted task distributions and can reduce certain forms of hallucination in specialized domains [15]. However, fine-tuning does not eliminate any of the eight failure classes described in this paper. It may reduce failure rates in the training distribution while leaving failure rates unchanged, or even elevated, in the tail of the operational distribution, which carries the most significant consequences in critical applications. Fine-tuning also introduces temporal drift at the model level. Each new fine-tuned model version must be re-validated, imposing ongoing evaluation costs that are difficult to sustain in operational environments with rapidly evolving task requirements.

6.4 Multi-Agent and Human-in-the-Loop Oversight

Multi-agent architectures, in which LLM agents monitor and validate each other's outputs, have been proposed as a mechanism for improving reliability. However, the International AI Safety Report 2026 explicitly identifies multi-agent systems as introducing distinct reliability and control challenges [6]. Errors can propagate across agents, and the behavior of multi-agent systems introduces additional complexity in prediction and characterization compared to individual models. Human-in-the-loop oversight is a complementary approach, but the overconfidence failure class undermines it. When AI systems present outputs with apparent confidence, human supervisors are less likely to apply appropriate scrutiny [18]. Studies indicate that even domain experts may not consistently identify AI errors when they are presented in fluent, confident language [3].

6.5 Structural Gaps in the Reliability Assurance Ecosystem

Beyond specific technical mitigation, the ecosystem for LLM reliability assurance contains structural gaps that cannot be addressed by any single technical intervention. There are no open, standardized benchmarks designed specifically to measure operational reliability rather than average performance in critical application domains [21]. A dedicated regulatory framework for LLM reliability, equivalent in scope to existing FDA requirements for Software as a Medical Device or MIL-STD requirements for defense systems, has yet to be established. The workforce with expertise spanning both AI system engineering and critical domain requirements represents an area of ongoing professional development and institutional capacity-building [4]. New model versions are released on timelines that make thorough operational validation challenging, and the typical development cycle of training, fine-tuning, and deployment has not yet incorporated the reliability characterization that critical applications require [6]. AI supervision systems capable of detecting accuracy failures in production, as distinct from safety policy violations, remain an active area of research and development.

7. Case Analysis: Healthcare and Defense

7.1 Healthcare

The healthcare sector represents a domain where LLM deployment is already active, encompassing clinical decision support, documentation, patient communication, and diagnostic assistance, and where reliability failures carry direct patient safety consequences. The deployment landscape is further shaped by the fact that healthcare is a heavily regulated domain, in which medical devices must meet FDA requirements for Software as a Medical Device, and clinical AI systems are subject to institutional review processes that continue to evolve alongside the rapid pace of LLM development and iteration.

The reliability failures most consequential for healthcare include epistemic hallucination producing fabricated drug interactions or non-existent clinical trial citations, prompt fragility causing performance variation driven by nonclinical features of patient messages [11], overconfidence failure in which AI systems present incorrect diagnostic outputs with apparent confidence thereby reducing physician scrutiny, and temporal drift in which model behavior changes following updates in ways not communicated to clinical users. AI mental health tools have exhibited additional failure modes, including performance gaps with respect to patients with severe psychiatric conditions and insufficient responsiveness to suicidal ideations, patterns that have not shown consistent improvement across model scales and generations [12].

The absence of open-ended, ecologically valid evaluation benchmarks for clinical AI means that regulatory and institutional decisions regarding model deployment are based on multiple-choice performance that has not been shown to reflect real-world performance reliably [14], [22], [23], [24]. This represents a documented gap in the evaluation infrastructure for clinical AI systems that warrants dedicated attention.

7.2 Defense

In defense contexts, reliability failures can result in consequences ranging from operational ineffectiveness to irreversible lethal harm. Large Language Models are currently being evaluated and deployed for tasks including intelligence synthesis, course-of-action analysis, logistics optimization, and autonomous system control [4], [5]. Each of these applications involves at least a subset of the eight failure classes described in this paper, and applications involving autonomous control potentially involve all of them.

Anthropic, one of the leading frontier AI developers, has explicitly stated that frontier AI systems do not yet meet the reliability requirements for fully autonomous weapons systems and that the necessary safeguards remain absent [4]. The International AI Safety Report 2026 notes that AI agents, which are increasingly central to defense system architectures, remain prone to basic errors and that human oversight is more difficult to maintain in multi-agent configurations [6]. Research examining AI applications in defense mental health contexts has identified reliability as a first-order concern for AI systems supporting service members in high-stress operational environments [5].

Adversarial manipulation constitutes a qualitatively distinct concern in defense contexts. Unlike healthcare, where prompt injection attacks are primarily theoretical risks, defense deployment contexts involve deliberate adversaries with both the performance and the motivation to craft inputs that manipulate AI behavior. The demonstrated vulnerability of LLMs to adversarial prompting, which current guardrail approaches have not yet fully mitigated, represents a fundamental obstacle to reliable deployment in contested operational environments [6], [20].

8. Toward a Reliability Standard for LLMs in Critical Applications

The preceding analysis leads to a set of research and governance priorities that this paper identifies as necessary prerequisites for reliable LLM deployment in critical domains.

8.1 Reliability-Oriented Benchmarking

Existing benchmarks must be supplemented with reliability-specific evaluations that use open-ended formats not susceptible to format exploitation, assess performance across semantically equivalent paraphrases to characterize prompt fragility, include adversarial and degraded-condition inputs, measure confidence calibration rather than accuracy alone, and are designed such that perfect performance is achievable in principle, enabling the detection of residual failures at saturation [21]. The HIP-LLM framework provides a methodological foundation for this work by defining reliability as the probability of success on future tasks under a specified operational profile [9].

8.2 The CRIT-LLM Benchmark

This paper proposes the CRIT-LLM Benchmark as a reliability-oriented evaluation instrument designed specifically to assess LLM performance in critical deployment contexts. Unlike existing benchmarks that measure average performance under favorable conditions, CRIT-LLM is structured around six evaluation components, each targeting a distinct dimension of operational reliability identified within the ORFT framework. The first component consists of adversarial prompts, crafted inputs that probe model behavior under deliberate manipulation, including prompt-injection attacks and inputs designed to elicit policy violations or factual fabrication. The second component evaluates paraphrase robustness by testing model outputs across semantically equivalent reformulations of the same query, thereby characterizing the degree of prompt fragility present in a given system. The third component assesses long-context performance by presenting models with extended inputs that require integrating information distributed across the full context window, with particular attention to degradation in the middle-context region documented in prior

literature [27]. The fourth component evaluates multilingual reliability by administering equivalent prompts across a representative set of high-resource and low-resource languages, measuring the performance disparity attributable to language rather than task difficulty. The fifth component introduces noisy inputs, including typographically degraded, grammatically irregular, and domain-jargon-laden prompts that reflect the realistic heterogeneity of operational inputs in clinical, defense, and infrastructure settings. The sixth component evaluates agent task reliability by assessing model performance within multi-step agentic workflows that require sequential planning, tool use, and error recovery, targeting the failure classes of reasoning collapse and agentic escalation. By combining these components into a unified evaluation instrument, CRIT-LLM enables the construction of failure profiles across the full ORFT taxonomy rather than the scalar accuracy scores produced by conventional benchmarks.

8.3 The Operational Reliability Score

This paper further proposes the Operational Reliability Score (ORS) as a composite metric for summarizing LLM reliability across the ORFT failure taxonomy. Existing evaluation metrics, including benchmark accuracy, pass rates, and human preference scores, measure aggregate performance without characterizing how failures are distributed across the input space or how confident the model is at the point of failure. The ORS is designed to address this limitation by aggregating evidence across four components. The first is a weighted accuracy score computed over the CRIT-LLM benchmark components, with component weights determined by the severity of failure consequences in the target deployment domain. The second is a confidence calibration index that quantifies the degree to which model confidence estimates correspond to empirical accuracy, penalizing systems that produce high-confidence incorrect outputs. The third is a failure concentration measure that characterizes whether failures are distributed uniformly across the input space or disproportionately concentrated in high-consequence input regions, such as adversarial or safety-relevant queries. The fourth is a temporal stability coefficient that tracks ORS values across successive model versions and updates, providing a quantitative indicator of the degree of temporal drift present in a deployed system. A higher ORS reflects a more operationally reliable system. A system that achieves a high scalar accuracy score but exhibits poor confidence calibration or failure concentration in high-stakes input regions will produce a substantially lower ORS, thereby signaling deployment risk that conventional metrics would not detect. The ORS is intended to serve as the primary reliability reporting metric for CRIT-LLM evaluations and as the basis for regulatory reliability thresholds in critical domain deployment frameworks.

8.4 The LLM Reliability Stress Test Suite

Complementing the CRIT-LLM Benchmark and the ORS metric, this paper proposes the LLM Reliability Stress Test Suite (LRSTS) as a standardized collection of targeted test protocols designed to probe individual ORFT failure classes in isolation. Whereas CRIT-LLM evaluates reliability holistically across an integrated benchmark, LRSTS provides modular test instruments that practitioners can apply selectively to diagnose specific reliability concerns in a given deployment context. The LRSTS comprises five example test protocols. The first, paraphrase stability testing, administers multiple semantically equivalent reformulations of each test query and measures the variance in model outputs across reformulations. A system exhibiting low paraphrase stability produces outputs that differ substantially across phrasings of the same question, indicating prompt fragility that would render it unreliable across the varied natural-language inputs of operational users. The second, typo robustness testing, introduces controlled orthographic degradation into prompts and measures the resulting accuracy decrement. This protocol is calibrated against the empirically documented sensitivity of LLMs to typographical variation [25], and models are assessed against a minimum robustness threshold that reflects the expected frequency of typographic error in the target operational environment. The third, adversarial injection testing, embeds prompt-injection sequences within otherwise legitimate inputs and assesses the extent to which the model's behavior deviates from the intended instructions. Test cases are drawn from documented attack patterns that are scaled to reflect the adversarial sophistication plausible in the target deployment domain, with defense contexts requiring higher

adversarial intensity than civilian applications. The fourth, long-context reasoning test presents models with inputs of increasing length and assesses performance on queries that require integrating information distributed at varying positions within the context window. Performance is measured as a function of both input length and information position, enabling the characterization of context-dependent reliability degradation and informing safe maximum context limits for each deployment. The fifth, abstention calibration testing, presents models with queries for which a correct answer is unavailable, outside the model's demonstrable competence, or would require information not provided in the context. The test measures the degree to which models appropriately decline to answer, rather than producing confident incorrect outputs, and quantifies the calibration between stated uncertainty and empirical accuracy. Together, the LRSTS protocols are intended to function as a pre-deployment checklist for critical applications, providing evidence that a system has been assessed against principal failure classes most consequential to the target domain.

8.5 Domain-Specific Operational Profiles

Reliability cannot be assessed in the abstract. It must be evaluated relative to a specified operational profile. Critical domain regulators, including the FDA for medical devices and relevant defense acquisition authorities for military systems, should develop operational profiles that specify the input distributions, performance requirements, and failure mode tolerances appropriate to each deployment context. LLM reliability assessment should be conducted against these domain-specific profiles rather than against generic academic benchmarks.

8.6 Continuous Post-Deployment Monitoring

Given the demonstrated problem of temporal drift, reliability assessment cannot be treated as a one-time pre-deployment activity. Operational LLM systems in critical domains require a continuous monitoring infrastructure capable of detecting performance changes, distributional shifts, and emerging failure modes. Such infrastructure does not currently exist at a production scale and requires substantial investment in both methodology and operational tooling.

8.7 Regulatory Frameworks

Critical domain regulators should develop AI-specific reliability standards equivalent in scope to existing software safety frameworks, including MIL-STD-882 for defense system safety and IEC 62304 for medical software lifecycle processes. These standards should specify minimum reliability thresholds, required validation methodologies, change management obligations, and post-market surveillance requirements. The LLM Risk Assessment Framework provides a starting point for systems engineering contexts [15].

8.8 Human Oversight Calibration

Given that overconfidence failure systematically undermines human oversight of AI systems, critical deployments should incorporate explicit interventions designed to counteract over-reliance. Such interventions include presenting AI outputs with explicit, calibrated uncertainty quantification, training human operators to apply domain-appropriate scrutiny to AI outputs irrespective of apparent confidence, and designing human-AI interaction protocols that require active human verification rather than passive acceptance of AI recommendations [18].

9. Conclusion

Large Language Models represent a genuinely transformative technology with substantial potential for beneficial application in healthcare, defense, and other critical domains. This paper does not contend that this potential is illusory or that LLM-based systems should be categorically excluded from critical domains. It submits that current LLMs, evaluated against operational reliability standards rather than benchmark accuracy, are not yet sufficiently reliable for autonomous deployment in contexts where failures produce irreversible harm.

The eight failure classes constituting the LLM Operational Reliability Failure Taxonomy (ORFT), namely epistemic hallucination, overconfidence failure, abstention failure, prompt fragility, temporal drift, reasoning collapse, agentic escalation, and adversarial manipulation [20], [28], are empirically documented, are structurally rooted in the probabilistic nature of LLM text generation, and are not adequately addressed by currently available mitigation approaches. Standard benchmarking practices, dominated by multiple-choice evaluation formats, have not been shown to adequately characterize these failure modes or measure operational reliability as it manifests in real-world deployment.

The path toward reliable deployment of LLMs in critical applications requires sustained investment in reliability-oriented benchmarking, domain-specific operational profiles, continuous post-deployment monitoring, appropriate regulatory frameworks, and human oversight mechanisms that explicitly account for the overconfidence problem. Until this investment is made and the resulting reliability evidence accumulated, the deployment of autonomous LLM agents in life-critical or mission-critical applications should be approached with substantial caution, and meaningful human oversight should be maintained at every consequential decision point.

The ORFT framework, together with the CRIT-LLM Benchmark, the Operational Reliability Score, and the LLM Reliability Stress Test Suite introduced in this paper, is intended as a starting point rather than a definitive account. Each of these instruments is designed to be extensible, and future work is explicitly encouraged to refine, expand, and empirically validate them as the field matures. Several directions merit dedicated investigation. With respect to evaluation frameworks, the CRIT-LLM Benchmark components should be expanded through systematic empirical work that calibrates each component against documented failure rates in operational deployments, establishing evidence-based input distributions, difficulty gradients, and minimum sample sizes sufficient to characterize failure distributions with statistical confidence. The ORS weighting scheme requires empirical validation to determine whether the proposed component weights produce scores that correlate with independently observed deployment reliability, and domain-specific weight configurations should be developed for healthcare, defense, and other critical sectors. Future work may also extend the LRSTS by developing additional test protocols targeting failure classes not yet covered, including temporal drift testing that tracks ORS values across successive model updates and multilingual abstention calibration that assesses refusal behavior across language conditions. The ORFT itself may be refined by adding new failure classes as novel deployment contexts introduce failure modes not yet observed, and cross-domain validation in sectors beyond healthcare and defense, including critical infrastructure, legal systems, and financial decision-making, represents a further productive avenue. Finally, the relationship between the ORFT failure classes and the technical properties of specific model architectures, training regimes, and fine-tuning strategies warrants systematic empirical investigation, with the goal of identifying which design choices most effectively reduce failure rates within each class and under what operational conditions those reductions hold.

References

[1] A. Rane et al., “A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications,” Electronics, vol. 14, no. 18, p. 3580, 2025. doi: 10.3390/electronics14183580.

[2] T. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020. doi: 10.48550/arXiv.2005.14165.

[3] L. Zhou et al., “Larger and more instructable language models become less reliable,” Nature, vol. 634, pp. 61–68, 2024. doi: 10.1038/s41586-024-07930-y.

[4] Anthropic, “Statement: Department of War,” Anthropic, 2025. [Online]. Available: https://www.anthropic.com/news/statement-department-of-war

[5] M. Finn and J. L. Murray, “AI for Defence: Readiness, Resilience and Mental Health,” RUSI Journal, vol. 169, no. 6, 2024. doi: 10.1080/03071847.2024.2424780.

[6] Y. Bengio et al., “International AI Safety Report 2026,” DSIT 2026/001, Department for Science, Innovation and Technology, Feb. 2026. [Online]. Available: https://internationalaisafetyreport.org

[7] R. Lunardi et al., “On Robustness and Reliability of Benchmark-Based Evaluation of LLMs,” arXiv preprint arXiv:2509.04013, 2025. doi: 10.48550/arXiv.2509.04013.

[8] S. M. Mousavi et al., “Garbage In, Reasoning Out? Why Benchmark Scores are Unreliable and What to Do About It,” arXiv preprint arXiv:2506.23864, 2025. doi: 10.48550/arXiv.2506.23864.

[9] R. Aghazadeh-Chakherlou et al., “HIP-LLM: A Hierarchical Imprecise Probability Approach to Reliability Assessment of Large Language Models,” arXiv preprint arXiv:2511.00527, 2024. doi: 10.48550/arXiv.2511.00527.

[10] A. Zewe, “Study: Platforms that rank the latest LLMs can be unreliable,” MIT News, Feb. 9, 2026. [Online]. Available: https://news.mit.edu/2026/study-platforms-rank-latest-llms-can-be-unreliable-0209

[11] A. Zewe, “LLMs factor in unrelated information when recommending medical treatments,” MIT News, Jun. 23, 2025. [Online]. Available: https://news.mit.edu/2025/llms-factor-unrelated-information-when-recommending-medical-treatments-0623

[12] Stanford University, “New study warns of risks in AI mental health tools,” Stanford News, Jun. 2025. [Online]. Available: https://news.stanford.edu/stories/2025/06/ai-mental-health-care-tools-dangers-risks

[13] A. Zewe, “Researchers discover a shortcoming that makes LLMs less reliable,” MIT News, Nov. 26, 2025. [Online]. Available: https://news.mit.edu/2025/shortcoming-makes-llms-less-reliable-1126

[14] I. Groh et al., “The pitfalls of multiple-choice questions in generative AI and medical education,” Scientific Reports, vol. 15, 2025. doi: 10.1038/s41598-025-26036-7.

[15] S. Weiss et al., “Generative AI in Systems Engineering: A Framework for Risk Assessment of Large Language Models,” arXiv preprint arXiv:2602.04358, 2026. doi: 10.48550/arXiv.2602.04358.

[16] A. Pesaranghader and E. Li, “Hallucination Detection and Mitigation in Large Language Models,” arXiv preprint arXiv:2601.09929, 2026. doi: 10.48550/arXiv.2601.09929.

[17] Z. Ji et al., “Survey on Hallucination in Natural Language Generation,” ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023. doi: 10.1145/3571730.

[18] Stanford HAI, “AI Overreliance Is a Problem. Are Explanations a Solution?” Stanford Human-Centered AI, Mar. 13, 2023. [Online]. Available: https://hai.stanford.edu/news/ai-overreliance-problem-are-explanations-solution

[19] P. Song, P. Han, and N. Goodman, “Large Language Model Reasoning Failures,” Transactions on Machine Learning Research, 2026. doi: 10.48550/arXiv.2602.06176.

[20] Z. Wang et al., “Risk Assessment and Security Analysis of Large Language Models,” arXiv preprint arXiv:2508.17329, 2025. doi: 10.48550/arXiv.2508.17329.

[21] J. Vendrow et al., “Do Large Language Model Benchmarks Test Reliability?” arXiv preprint arXiv:2502.03461, 2025. doi: 10.48550/arXiv.2502.03461.

[22] D. Jin et al., “What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams,” Applied Sciences, vol. 11, no. 14, p. 6421, 2021. doi: 10.3390/app11146421.

[23] A. Singhal et al., “Large Language Models Encode Clinical Knowledge,” Nature, vol. 620, pp. 172–180, 2023. doi: 10.1038/s41586-023-06291-2.

[24] K. Saab et al., “Capabilities of Gemini Models in Medicine,” Nature Medicine, 2024. doi: 10.1038/s41591-024-03423-7.

[25] E. Gan, Y. Zhao, L. Cheng, M. Yancan, A. Goyal, K. Kawaguchi, M.-Y. Kan, and M. Shieh, “Reasoning Robustness of LLMs to Adversarial Typographical Errors,” in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, Florida, USA, Nov. 2024, pp. 10449–10459. doi: 10.18653/v1/2024.emnlp-main.584.

[26] S. Romanou et al., “MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation,” arXiv preprint arXiv:2503.10497, 2025. doi: 10.48550/arXiv.2503.10497.

[27] N. F. Liu et al., “Lost in the Middle: How Language Models Use Long Contexts,” Transactions of the Association for Computational Linguistics, vol. 12, pp. 157–173, 2024. doi: 10.1162/tacl_a_00638.

[28] T. Cui et al., “Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems,” arXiv preprint arXiv:2401.05778, 2024. doi: 10.48550/arXiv.2401.05778.