

ORIGINAL RESEARCH

Open access

Ground, Qualify, Verify, and Abstain—A Safety Framework for Hallucination Control, Evidence Grounding, Uncertainty Communication, and Decision Boundaries in Generative AI

Ingrid Berg^{1*}, Thomas Müller², Naledi Mokoena³

Abstract

Generative artificial intelligence can produce fluent, context-sensitive, and apparently authoritative responses even when relevant knowledge is incomplete, evidence is unavailable, or generated claims are false. This creates an epistemic safety problem concerning not only what a system produces, but also what it appears to know, how its claims are supported, how uncertainty is expressed, and when answering should be limited. This article develops a conceptual safety framework organized around four operational principles: ground, qualify, verify, and abstain. The framework treats hallucination as a multidimensional risk involving fabrication, factuality failure, groundedness failure, unsupported claims, false confidence, source and context errors, reasoning errors, citation errors, overgeneralization, confabulation, speculative output, and incomplete evidence. Grounding links material claims to relevant and sufficiently reliable evidence; qualification communicates uncertainty, assumptions, scope, and evidential limitations; verification examines claims, sources, citations, and inferential consistency; and abstention enables refusal, deferral, escalation, or constrained output when evidence or system competence is insufficient. Human review, meaningful oversight, decision boundaries, failure containment, incident monitoring, and lifecycle updating operate across these principles rather than functioning as final guarantees. The main contribution is an integrated generative AI safety framework that connects output-level hallucination control with source quality, knowledge verification, uncertainty communication, review quality, response limitation, and downstream risk containment. The framework is intended to structure domain-specific assessment and future research. It does not independently establish factual reliability, ethical sufficiency, legal compliance, operational safety, certification, or deployment readiness.

Keywords Hallucination, Generative AI, Human review, Epistemic safety, Evidence grounding, Uncertainty communication

*Correspondence:

Ingrid Berg

ingrid.berg@nhh.no

¹ Department of AI Safety and Reliability, Norwegian School of Economics, Bergen, Norway

² Department of Computer Science and AI, University of Vienna, Vienna, Austria

³ Department of Information Systems, University of Cape Town, Cape Town, South Africa

Introduction

Generative AI systems can synthesize language, images, code, explanations, and recommendations across a widening range of tasks, creating substantial informational and organizational opportunities while also introducing distinctive reliability and governance risks [1]. Their outputs are generated through learned statistical relationships rather than through an independently warranted understanding of truth, evidence, or meaning. Consequently, coherent language and

broad apparent competence should not be interpreted as proof that a system possesses justified knowledge or dependable semantic understanding [2]. This distinction is central to epistemic safety because users may treat fluent responses as authoritative even when the underlying evidence is missing, uncertain, outdated, or incorrectly represented.

Trustworthy AI cannot be reduced to accuracy or usefulness alone. It encompasses interacting requirements concerning reliability, robustness, transparency, accountability, human agency, fairness, privacy, and appropriate governance [3]. These requirements are socio-technical because system behavior is influenced by data, model design, interfaces, organizational processes, user expectations, review arrangements, and the environments in which outputs are applied [4]. For generative AI, trustworthiness therefore depends partly on whether claims are grounded, uncertainty is made intelligible, verification is proportionate to the claim, human intervention is meaningful, and unsupported responses can be constrained before they influence consequential decisions.

Existing discussion often treats hallucination as a discrete model defect or a problem to be solved through better prompting, retrieval, training, or filtering. Such approaches may reduce particular errors, but they do not by themselves resolve the broader problem of what a generative system should claim to know. Epistemic safety requires attention to evidence provenance, source reliability, contextual applicability, uncertainty, user reliance, and downstream decision consequences. It also requires recognition that apparently correct outputs can remain unsafe when their evidential basis is obscure, their scope is overstated, or users cannot distinguish verified information from plausible synthesis.

This article proposes an original conceptual framework organized around four connected principles: **ground**, **qualify**, **verify**, and **abstain**. Grounding concerns the relationship between claims and relevant evidence; qualification concerns uncertainty, assumptions, scope, and evidential limitations; verification concerns claim checking, source assessment, and consistency; and abstention concerns refusal, deferral, escalation, or output restriction when response conditions are not satisfied. Human review, decision boundaries, failure containment, monitoring, and lifecycle governance connect these principles. The framework is intended to clarify safety requirements and research priorities rather than prescribe universal technical architectures, thresholds, governance instruments, or deployment decisions.

Hallucination as an Epistemic Safety Problem

Hallucination in natural-language generation commonly refers to content that is plausible in form but unsupported by the relevant input, source, or established knowledge [5]. In large language models, the problem includes multiple forms of factual, contextual, reasoning, citation, and faithfulness error arising from training conditions, decoding processes, knowledge limitations, retrieval defects, and interaction context [6]. A single definition is therefore insufficient for every use case. The relevant safety question is not merely whether an output contains an error, but whether the system presents a claim with a level of evidential authority, certainty, or completeness that the available information cannot justify.

Hallucination risk is intensified when the system cannot adequately distinguish between a supported answer and a linguistically coherent completion. Semantic uncertainty may help identify some confabulations by measuring variation in meaning across alternative generations, but it does not establish the factual truth of an answer or detect every unsupported claim [7]. Broad task evaluations likewise show that conversational systems can perform strongly in some contexts while remaining inconsistent or unreliable in others [8]. Apparent versatility must therefore be separated from dependable competence, especially when users may generalize success in familiar tasks to domains where the system has not demonstrated comparable reliability.

Treating hallucination as an epistemic safety problem broadens the analysis beyond factual inaccuracy. Fabrication concerns invented entities, events, evidence, or sources; groundedness concerns fidelity to supplied material; factuality concerns correspondence with externally checkable facts; and truthfulness concerns whether the response communicates warranted information rather than merely persuasive text. Unsupported claims may be true by coincidence yet remain epistemically defective because no adequate evidence connects the system's assertion to the proposition.

False confidence adds another layer of risk by presenting uncertain, incomplete, or speculative material in language that encourages inappropriate reliance.

The resulting harm pathway is relational rather than purely technical. A generated statement becomes consequential when users interpret it, integrate it into another process, communicate it onward, or use it to support a decision. Responsible and trustworthy AI therefore require controls over both content and the conditions under which content is presented. No grounding method, uncertainty estimator, verification process, human reviewer, or refusal rule can independently eliminate hallucination or establish safety. The proposed framework retains explicit no-hallucination-elimination and no-safety-guarantee boundaries so that partial safeguards are not misrepresented as comprehensive assurance.

Sources and Forms of Generative AI Error

The term *hallucination* is sometimes used so broadly that it obscures important distinctions between factual error, source unfaithfulness, unsupported inference, and other failures [9]. A safety-oriented classification should therefore begin with the relationship among the claim, its context, and the evidence expected to support it. A factuality error contradicts verifiable information; a groundedness error departs from provided or retrieved material; a truthfulness failure misrepresents what is warranted; and an unsupported claim lacks an adequate evidential basis even when it cannot immediately be shown to be false. Distinguishing these forms helps determine whether the appropriate response is grounding, qualification, verification, correction, or abstention.

Error can also arise from incentives embedded in model evaluation and interaction design. When systems are rewarded principally for supplying an answer, without adequate recognition of justified non-answering, they may be encouraged to guess rather than acknowledge insufficient knowledge [10]. False confidence, speculative completion, and overgeneralization can consequently emerge from pressures toward completeness and responsiveness. A model may answer beyond the available context, extend a limited observation into a general rule, or convert an ambiguous possibility into a definite statement. Epistemic safety therefore requires evaluation that considers appropriate abstention and qualification alongside answer accuracy.

Upstream data and knowledge conditions create additional risks. Recursive reliance on model-generated material can degrade informational distributions and progressively weaken model reliability, demonstrating why data provenance and source authenticity matter throughout the lifecycle [11]. Retrieval-supported systems can also fail when queries are poorly formed, relevant evidence is absent, sources are outdated, or retrieved material is unreliable. Citation errors may involve non-existent references, incorrect bibliographic details, or real sources that do not support the associated claim. These failures can be particularly misleading because the appearance of sourcing may increase user confidence even when the evidential connection is defective.

Calibration and knowledge distribution also shape error behavior. Research on monofacts and miscalibration indicates that limited exposure to particular facts and unreliable confidence estimates can contribute to hallucination risk [12]. Context errors arise when generally correct information is applied to an unsuitable case, while reasoning errors occur when valid premises are combined through invalid inference. Confabulation fills evidential gaps with coherent narrative, and incomplete evidence creates a misleading appearance of comprehensiveness. **Table 1** summarises sources and forms of generative AI error that create hallucination, factuality, grounding, uncertainty, and epistemic safety risks.

Table 1. Sources and Forms of Generative AI Error: Hallucination, Fabrication, Unsupported Claims, False Confidence, Factuality Errors, Groundedness Errors, Truthfulness Failures, Source Errors, Context Errors, Retrieval Errors, Reasoning Errors, Citation Errors, Overgeneralization, Confabulation, Speculative Output, Incomplete Evidence, and Hallucination Boundaries

Error source or form	Core safety question	How the error appears	Potential harm pathway	Affected use context	Detection or mitigation need	Boundary statement	Friction
Hallucination	Is the output adequately supported?	Plausible but unsupported content	Fluent error is accepted as knowledge	Knowledge-intensive generation	Grounding, verification, qualification	No single control eliminates hallucination	Uncertainty
Fabrication	Has an entity, event, result, or source been invented?	Non-existent facts or references	False evidence enters downstream work	Research and professional writing	Entity and source validation	Coherence does not establish existence	Verification
Unsupported claim	What evidence justifies the assertion?	Claim lacks traceable support	Unwarranted statement influences action	Evidence-sensitive tasks	Claim-to-evidence mapping	Possible truth does not equal justification	Grounding
False confidence	Does wording overstate certainty?	Definitive language under uncertainty	Users over-rely on weak output	Ambiguous or novel requests	Calibration and qualification	Expressed confidence may be unreliable	Qualification
Factuality error	Does the claim conflict with facts?	Incorrect dates, names, quantities, or relations	Incorrect belief or decision	Information and analytical use	External fact-checking	Verification can still miss errors	Verification
Groundedness error	Is the answer faithful to supplied evidence?	Response departs from context or sources	Source-backed appearance conceals inconsistency	Document-based generation	Source–claim alignment checking	Faithfulness does not prove source truth	Grounding
Truthfulness failure	Is the system communicating warranted information?	Misleading or distorted presentation	Users infer dependable knowledge	Conversational and advisory use	Claim and context review	Truthfulness exceeds factual accuracy	Context
Source error	Is the relied-upon source authentic and suitable?	Invented, stale, biased, or weak source	Poor evidence appears authoritative	Research and synthesis	Source-quality assessment	Source presence does not establish quality	Evidence quality
Context error	Is the information applicable to the actual case?	Correct general information is misapplied	Inappropriate transfer causes harm	Domain-specific requests	Scope and context confirmation	General validity does not ensure local fit	Domain

Retrieval error	Was relevant and reliable evidence retrieved?	Missing, irrelevant, or conflicting material	Generation is grounded in defective context	Retrieval-supported generation	Retrieval and coverage evaluation	Retrieval does not guarantee truth	F
Reasoning error	Does the conclusion follow from the evidence?	Invalid inference or contradiction	Accurate facts yield a false conclusion	Multi-step analysis	Consistency and logic checking	Correct evidence does not ensure correct reasoning	Ve
Citation error	Does the citation support the exact claim?	Invented, misplaced, or non-entailing citation	Unsupported output gains false authority	Academic and technical writing	Citation existence and entailment checking	Citation format does not prove support	Tr
Overgeneralization	Is limited evidence presented universally?	Narrow result becomes a broad rule	Contextually unsuitable application	Synthesis and advisory tasks	Scope qualification	Local evidence cannot prove universality	Qu
Confabulation	Is coherent narrative filling a knowledge gap?	Internally consistent unsupported explanation	Coherence hides absence of evidence	Open-ended dialogue	Uncertainty and grounding checks	Coherence is not justification	AI
Speculative output	Is conjecture distinguished from established knowledge?	Possibility is presented as fact	Predictions influence action	Forecasting and emerging topics	Explicit speculation labels	Qualification does not validate prediction	Q re c
Incomplete evidence	Are material sources or conditions absent?	Partial evidence appears comprehensive	Alternatives and limitations are overlooked	Research synthesis and complex decisions	Coverage review and escalation	Available evidence may be insufficient	e
Hallucination boundary	Can safeguards remove all residual error?	Errors persist despite controls	Safeguard confidence creates new exposure	All generative AI use	Monitoring and residual-risk documentation	Hallucination elimination is not established	M fra b

Evidence Grounding and Knowledge Verification

Evidence grounding connects generated claims to identifiable information that is relevant to the question and suitable for the intended context. Retrieval-augmented generation can provide access to external documents and more current knowledge, but its reliability remains dependent on query formulation, retrieval coverage, ranking, source quality, context construction, and the model's use of retrieved material [13]. Grounding should therefore be treated as a structured

evidence process rather than as the simple inclusion of retrieved text. A response may be faithful to a retrieved passage while remaining factually wrong because the source itself is unreliable, outdated, incomplete, or inapplicable.

Grounding requirements should distinguish evidence grounding, knowledge grounding, source grounding, grounded generation, and citation grounding. Evidence grounding asks what supports the claim; knowledge grounding asks whether an appropriate domain of knowledge has been consulted; source grounding asks whether identifiable sources are accurately represented; and grounded generation asks whether the response remains faithful to the supplied context. Factuality research and fact-checking scholarship indicate that verification can strengthen claim assessment, but difficulties remain in defining checkable claims, locating suitable evidence, resolving disagreement, and evaluating changing or context-dependent information [14]. These limitations require explicit no-grounding-sufficiency and no-verification-guarantee boundaries.

Claim verification should begin by decomposing an output into material assertions rather than treating the response as a single unit. Automated fact-checking research commonly separates claim identification, evidence retrieval, source assessment, and evidence-based judgment [15]. This structure is useful for generative AI because different sentences in one response may have different evidence needs. Descriptive claims may require direct documentation, causal claims may require stronger methodological support, predictions may require qualification, and recommendations may depend on context that the system cannot independently establish. Verification outputs should therefore permit categories such as supported, contradicted, mixed, unresolved, or outside the available evidence.

Cross-checking and consistency checking provide complementary safeguards. Cross-checking compares claims against independent evidence pathways, whereas consistency checking examines contradictions within the response, across repeated generations, or between the response and the supplied evidence. Surveys of automated fact-checking show that verification systems remain sensitive to dataset design, evidence availability, domain assumptions, and the relationship between retrieved material and the final judgment [16]. Agreement among sources may be misleading when those sources reproduce the same original error, and internal consistency may merely indicate that an unsupported narrative has been generated coherently.

Evidence traceability, knowledge provenance, and data provenance extend verification across the lifecycle. Traceability should permit reconstruction of the relationship among the request, sources, retrieved passages, generated claims, qualifications, and review decisions. Knowledge provenance concerns the origin, version, and transformation of relied-upon information, while data provenance concerns the lineage of training, retrieval, or updating material where such information is available. These practices may support auditability and incident analysis, but they cannot guarantee factuality. The framework therefore treats grounding and verification as mutually reinforcing controls whose adequacy must be evaluated for the specific claim, source environment, user context, and potential consequence.

Uncertainty Communication and Human Review

Uncertainty should be treated as a property requiring interpretation rather than as a single confidence value. Epistemic uncertainty concerns limitations in knowledge that may be reduced through additional evidence, whereas aleatoric uncertainty concerns variability that may remain even when knowledge improves [17]. This distinction can help determine whether a system should retrieve further evidence, qualify an answer, present alternatives, defer to expertise, or abstain. It does not follow, however, that a generative model can reliably identify every knowledge gap through its own confidence signals.

Uncertainty communication translates uncertainty estimates and evidential limitations into language that users can interpret. Linguistic calibration may reduce conversational overconfidence by aligning the strength of an assertion with the available confidence or support [18]. A qualified response may state assumptions, identify conflicting sources, distinguish established evidence from speculation, narrow the scope of a conclusion, or explain why a complete answer is not

warranted. Qualification should not be reduced to generic disclaimers, because repetitive warnings may be ignored and may fail to communicate the particular source of uncertainty.

The usefulness of uncertainty communication depends on how intended users interpret it. Evidence from high-consequence machine-learning contexts indicates that uncertainty information can support second opinions and review, but poorly presented signals may confuse users or create unwarranted reassurance [19]. Numerical confidence can appear more precise than the underlying estimate justifies, while verbal terms such as “likely” or “uncertain” may be understood differently across users and settings. Explanation uncertainty should also be considered because a plausible explanation of an answer may itself be unstable, incomplete, or unsupported.

Human review can supplement model outputs with contextual understanding, domain knowledge, responsibility, and access to evidence not available to the system. Human-in-the-loop arrangements nevertheless require defined interaction points, reviewer roles, decision authority, evidence access, feedback channels, and escalation pathways [20]. Human-on-the-loop supervision may be suitable for monitoring some operations, whereas material claims may require review before an output is used. Meaningful review requires sufficient time, competence, independence, and authority to challenge or reject the system rather than merely approve its recommendation.

Human involvement does not automatically prevent harm. Empirical research on trust in AI shows that users may over-trust or under-trust systems depending on perceived competence, transparency, anthropomorphic cues, prior experience, and task context [21]. Reviewers may display automation bias, accept fluent content without checking evidence, or become overloaded by excessive referrals. The framework therefore includes a no-uncertainty-guarantee boundary and a no-human-review-sufficiency boundary: communicated uncertainty does not guarantee calibrated reliance, and human presence does not independently establish correctness or safety.

Decision Boundaries and Failure Containment

A decision boundary specifies what a generative AI system may answer, what kind of support it may provide, and when a request must be limited or transferred. Machine-learning systems with a reject option demonstrate the general value of withholding predictions when confidence or coverage conditions are not satisfied [22]. In generative AI, comparable response options include a grounded answer, a qualified answer, a request for additional evidence, deferral, refusal, abstention, or escalation. Their appropriateness depends on evidence availability, system scope, expected consequences, and the capacity of an alternative reviewer.

Generative AI systems should not always answer. Low-evidence contexts, out-of-scope requests, unresolved factual conflicts, safety-critical contexts, and risk-sensitive decisions may justify response limitation even when fluent completion is possible. Abstention has epistemic value because it acknowledges that answering can be less responsible than withholding a claim, but abstention also has limitations and may shift responsibility rather than resolve it [23]. A refusal can be inappropriate, inconsistent, or excessively broad; deferral can fail when no competent reviewer is available; and delayed answers can themselves create harm.

Figure 1 illustrates the ground–qualify–verify–abstain safety logic for generative AI, showing how evidence availability, uncertainty, risk context, human review, and decision boundaries shape safe response behavior.

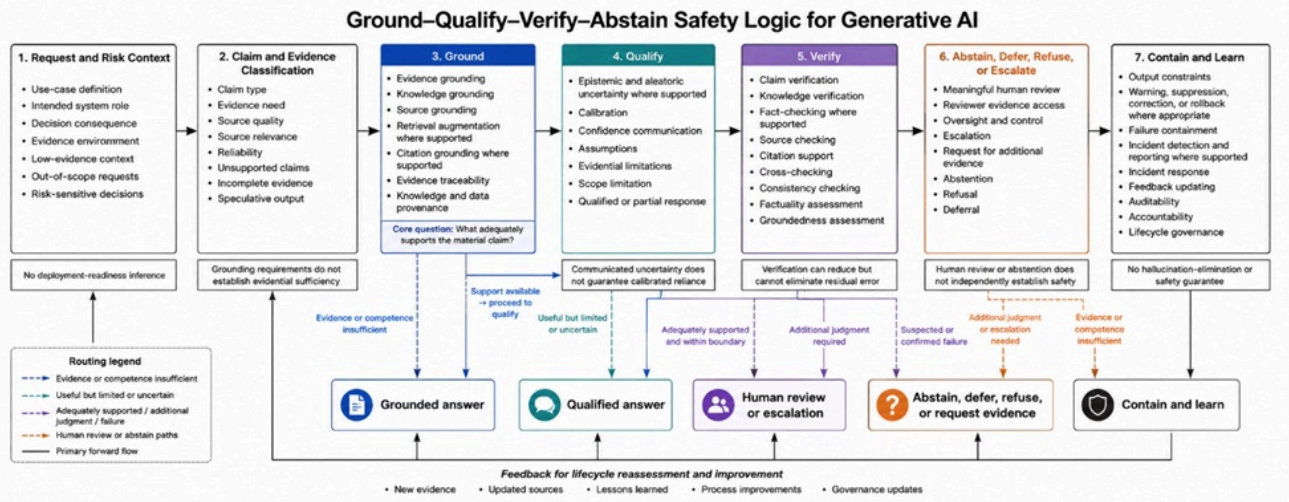


Figure 1. Ground-Qualify-Verify-Abstain Safety Logic for Generative AI

Alt-text description.

A safety logic diagram showing generative AI outputs routed through grounding, qualification, verification, human review, abstention, refusal, deferral, decision boundaries, and failure containment.

Failure containment is required because error prevention will remain incomplete. Safety assurance research shows that technical controls interact with ethical, organizational, and legal considerations, leaving gaps that cannot be closed by one mechanism [24]. Containment can include suppressing a suspected output, constraining its detail, attaching a correction, withdrawing distributed content where feasible, directing users to validated sources, suspending a function, or escalating an incident. These actions reduce propagation risk but cannot ensure that every recipient sees a correction or reverses a decision already influenced by the error.

Containment must continue across the machine-learning lifecycle. Assurance requirements include data and model traceability, change management, monitoring, evaluation after updates, incident detection, and reassessment when the operating context changes [25]. Incident reporting and response processes should identify what failed, what evidence was unavailable, how the output passed existing controls, and whether similar outputs may have been produced elsewhere. Feedback updating may revise retrieval sources, verification procedures, interface warnings, decision boundaries, or escalation triggers, but each update requires renewed evaluation because corrective action can introduce different failures.

Human-AI task allocation should also be treated as a boundary-design problem. Research on human-AI collaboration indicates that delegation can create cognitive challenges, including reduced attention, misplaced reliance, and difficulty integrating human and machine judgments [26]. High-risk cases should not be transferred to nominal reviewers without evidence access or real authority, and low-risk cases should not generate so many alerts that meaningful review becomes impossible. Table 2 maps evidence grounding, knowledge verification, uncertainty communication, human review, decision boundaries, abstention, and failure containment as core safety requirements for generative AI.

Table 2. Core Safety Requirements for Hallucination Control and Epistemic Risk Management: Evidence Grounding, Knowledge Verification, Source Quality, Claim Verification, Cross-Checking, Uncertainty Communication, Qualified Responses, Human Review, Abstention, Refusal, Deferral, Decision Boundaries, Failure Containment, Monitoring, and Governance Boundaries

Safety requirement	Core safety question	Required evidence, practice, or trigger	Safety function	Potential output	Failure or residual risk	Human review, monitoring, or governance need	Frame for evaluation
Evidence grounding	What supports the claim?	Claim-to-evidence mapping	Reduces unsupported output	Grounded or withheld claim	Weak evidence may appear sufficient	Claim-level evaluation	Gro
Knowledge grounding	Was the appropriate knowledge domain consulted?	Relevant and current knowledge resources	Constrains open generation	Domain-linked answer	Knowledge base may be incomplete	Domain stewardship	Gro
Source quality	Is the source suitable and dependable?	Authority, currency, method, and independence review	Filters weak support	Quality-qualified source set	Prestigious sources may still be inapplicable	Source governance	Evid
Source relevance	Does the evidence address the exact claim?	Contextual and semantic matching	Prevents irrelevant support	Relevance-qualified evidence	Keyword similarity can mislead	Relevance checking	Evid
Claim verification	Does evidence support the generated assertion?	Claim extraction and evidence comparison	Detects conflict	Supported, disputed, or unresolved claim	Verification errors remain possible	Independent assessment	Ve
Cross-checking	Do independent evidence paths converge?	Comparison across sufficiently independent sources	Reduces single-source dependence	Corroborated or conflicting result	Sources may share one original error	Independence assessment	Ve
Consistency checking	Is the output internally and externally coherent?	Contradiction and entailment review	Detects logical conflict	Consistent or unresolved output	Consistency does not prove truth	Method evaluation	Ve
Uncertainty communication	Are knowledge limits understandable?	Context-specific confidence and evidence language	Reduces false certainty	Qualified response	Users may misinterpret qualifiers	Human-centered testing	Que

Qualified response	Can useful information be given within limits?	Scope, assumptions, uncertainty, and evidence stated	Avoids false completeness	Conditional or partial answer	Qualifications may be overlooked	Reliance monitoring	Qualified response
Human review	Is contextual judgment required?	Competent reviewer with evidence and authority	Adds accountable assessment	Approve, revise, reject, or escalate	Automation bias and overload	Review-quality assessment	Review quality
Abstention	Is answering less justified than withholding?	Insufficient evidence, competence, or permitted scope	Prevents forced guessing	Explicit non-answer	Excessive or uneven abstention	Coverage and outcome monitoring	Abstention
Refusal	Is the requested response prohibited or unsafe?	Defined safety, authorization, or scope trigger	Blocks impermissible output	Refusal with limited explanation	False refusal or missed risk	Refusal audit	Bounded refusal
Deferral	Should another agent decide?	Specialist knowledge or authority required	Transfers judgment	Referral or pending review	Reviewer may be unavailable	Escalation governance	Bounded deferral
Decision boundary	What may the system answer or influence?	Use-case, evidence, uncertainty, and consequence analysis	Limits unsupported use	Proceed, qualify, defer, or refuse	Boundary may be bypassed	Boundary review	Certain consequences
Output constraint	Should content or actionability be restricted?	Elevated uncertainty or consequence	Limits downstream exposure	Extractive or source-bounded output	Constraints may be circumvented	Constraint testing	Containment
Failure containment	How can suspected error be prevented from spreading?	Error signal or incident trigger	Reduces propagation	Warning, suppression, correction, or rollback	Error may already have influenced users	Incident response	Containment
Monitoring	How will recurring or emerging failures be detected?	Complaints, corrections, source changes, and performance signals	Extends safeguards over time	Alert or investigation	Silent failures and alert fatigue	Lifecycle ownership	Monitoring

Governance boundaries	Do safeguards establish safety or compliance?	Explicit residual-risk and scope statements	Prevents overclaiming	Non-certified risk assessment	Users may infer approval	Documentation and oversight	Mand boun
-----------------------	---	---	-----------------------	-------------------------------	--------------------------	-----------------------------	-----------

Proposed Hallucination and Epistemic Risk Framework

The proposed framework begins with use-case and risk-context definition. Human-centered AI emphasizes that reliable automation should be combined with meaningful human control rather than treated as an autonomous substitute for accountable judgment [27]. The first stage therefore identifies the intended task, likely users, evidence environment, possible decision consequences, and conditions under which the system should not operate. It distinguishes informational support from consequential decision authority and establishes that framework completion does not imply operational safety or deployment readiness.

The second stage classifies the claim type and its evidence needs, followed by source and knowledge-domain assessment. Hybrid intelligence suggests that human and machine capabilities may be combined where their strengths are complementary and their limitations are explicitly managed [28]. A system may assist with retrieval, comparison, and drafting, while people may contribute domain knowledge, contextual interpretation, and accountable judgment. This complementarity should be designed rather than assumed, because human involvement can introduce delay, inconsistency, or bias and model output can shape the reviewer’s interpretation before independent assessment occurs.

The third stage defines grounding requirements, retrieves or reviews evidence where supported, and verifies material claims. The fourth stage estimates and communicates uncertainty, designs a qualified response, and determines whether meaningful human review is required. Automation and augmentation should not be treated as mutually exclusive choices because each can create dependencies that alter the risks of the other [29]. The framework therefore asks which activities may be automated, which should augment human analysis, which require independent review, and which should remain outside the system’s permitted role.

The fifth stage defines abstention, refusal, deferral, escalation, and decision-boundary triggers. The sixth establishes failure-containment, traceability, monitoring, incident detection, incident response, and feedback updating. Broad AI-governance principles frequently converge around transparency, responsibility, fairness, and human oversight, but their interpretation and implementation vary considerably [30]. Accordingly, the framework does not prescribe universal thresholds or institutional procedures. Each control requires context-specific validation, ownership, documentation, and reassessment when models, sources, users, or operating conditions change.

The final stage is a lifecycle governance review that preserves explicit sufficiency boundaries. Principles and conceptual frameworks cannot independently guarantee ethical AI because implementation depends on institutions, professional responsibilities, procedures, incentives, and mechanisms for accountability [31]. The proposed framework consequently retains no-grounding-sufficiency, no-verification-guarantee, no-uncertainty-guarantee, no-abstention-sufficiency, no-human-review-sufficiency, no-decision-boundary-sufficiency, no-hallucination-elimination, no-safety-guarantee, and no-deployment-readiness boundaries. Table 3 presents the proposed hallucination and epistemic risk framework linking grounding, qualification, verification, abstention, human review, failure containment, monitoring, and lifecycle governance.

Table 3. Proposed Hallucination and Epistemic Risk Framework for Generative AI: Use-Case Definition, Risk Context, Claim Classification, Evidence Need, Source Assessment, Grounding Requirements, Verification, Uncertainty Communication, Qualified Response Design, Human Review, Abstention, Refusal, Deferral, Decision Boundaries, Failure Containment, Auditability, Monitoring, Incident Response, and Deployment Boundaries

Framework stage	Core purpose	Required input	Safety function	Potential output	Validation or governance need	Failure risk	Feed mech
Use-case and risk-context definition	Establish intended role and consequence	Task, users, domain, stakes	Prevents context-free use	Use-case profile	Stakeholder and domain review	Context misclassification	Review mat cha
Claim and evidence-need classification	Match support requirements to claim type	Proposed claims and consequence level	Identifies material assertions	Evidence profile	Classification validation	Weak requirements	Verific feed
Source and knowledge-domain assessment	Identify suitable knowledge resources	Available domains and sources	Filters unsuitable evidence	Approved or qualified source domain	Domain stewardship	Outdated or irrelevant domain	Sou perform rev
Grounding requirement definition	Specify required claim support	Evidence profile	Makes grounding proportionate	Grounding specification	Context-specific testing	Requirements too weak or rigid	Incic driv rev
Evidence retrieval or source review	Gather relevant support	Source domain and query	Supplies external evidence	Evidence set	Retrieval evaluation	Missing or manipulated evidence	Retri error k
Claim verification and cross-checking	Test claims against evidence	Claims and evidence set	Detects unsupported assertions	Supported, disputed, or unresolved status	Independent assessment	False verification	Com lat corre
Uncertainty estimation and communication	Identify and express residual limits	Evidence gaps and model signals	Reduces false confidence	Qualified uncertainty statement	Calibration and comprehension testing	Miscalibration or misunderstanding	Us resp evalu
Qualified response design	Provide bounded useful information	Verified claims, scope, assumptions	Prevents false completeness	Conditional or partial answer	Human-centered assessment	Qualifier is overlooked	Relia feed
Human review and escalation	Add contextual judgment and authority	Trigger, evidence package, reviewer role	Supports accountable intervention	Revision, rejection, or escalation	Reviewer competence and workload review	Automation bias	Rev outc moni
Abstention and refusal triggers	Define when not to answer	Evidence, uncertainty, scope, consequence	Prevents forced completion	Abstention, refusal, or evidence request	Coverage and consistency evaluation	Excessive or uneven refusal	Defe case r

Decision-boundary definition	State permitted support role	Use-case and control capabilities	Separates assistance from decision authority	Boundary specification	Governance approval	Ambiguous or bypassed boundary	Violate analysis
Failure-containment design	Limit propagation of error	Distribution paths and incident scenarios	Reduces downstream exposure	Warning, correction, suppression, or rollback	Scenario testing	Delayed detection	Post-incident analysis
Auditability and traceability	Reconstruct evidence and actions	Logs, provenance, versions, reviews	Supports accountability	Auditable record	Access and retention governance	Missing or misleading records	Audit findings
Monitoring and incident detection	Detect drift and recurring failures	Reports, corrections, source and model changes	Extends control over time	Alert or incident record	Validated signals and ownership	Silent failure or alert fatigue	Significant refinement
Incident response and feedback updating	Correct failure and reduce recurrence	Confirmed incident and causal analysis	Supports learning	Correction, suspension, or redesign	Change control	Corrective action creates new risk	Closed learning
Lifecycle governance review	Reassess after material change	Updated model, data, source, or context	Prevents obsolete controls	Revised framework profile	Scheduled and event-driven review	Static governance	Versioned decisions
Sufficiency and deployment boundaries	Prevent framework overclaiming	Results from all stages	Preserves residual-risk awareness	Conceptual risk profile	Independent domain validation	Framework mistaken for certification	Evidence accumulation

Figure 2 presents the proposed hallucination and epistemic risk framework for generative AI, integrating error identification, grounding, verification, uncertainty communication, human review, abstention, failure containment, monitoring, and governance updating.

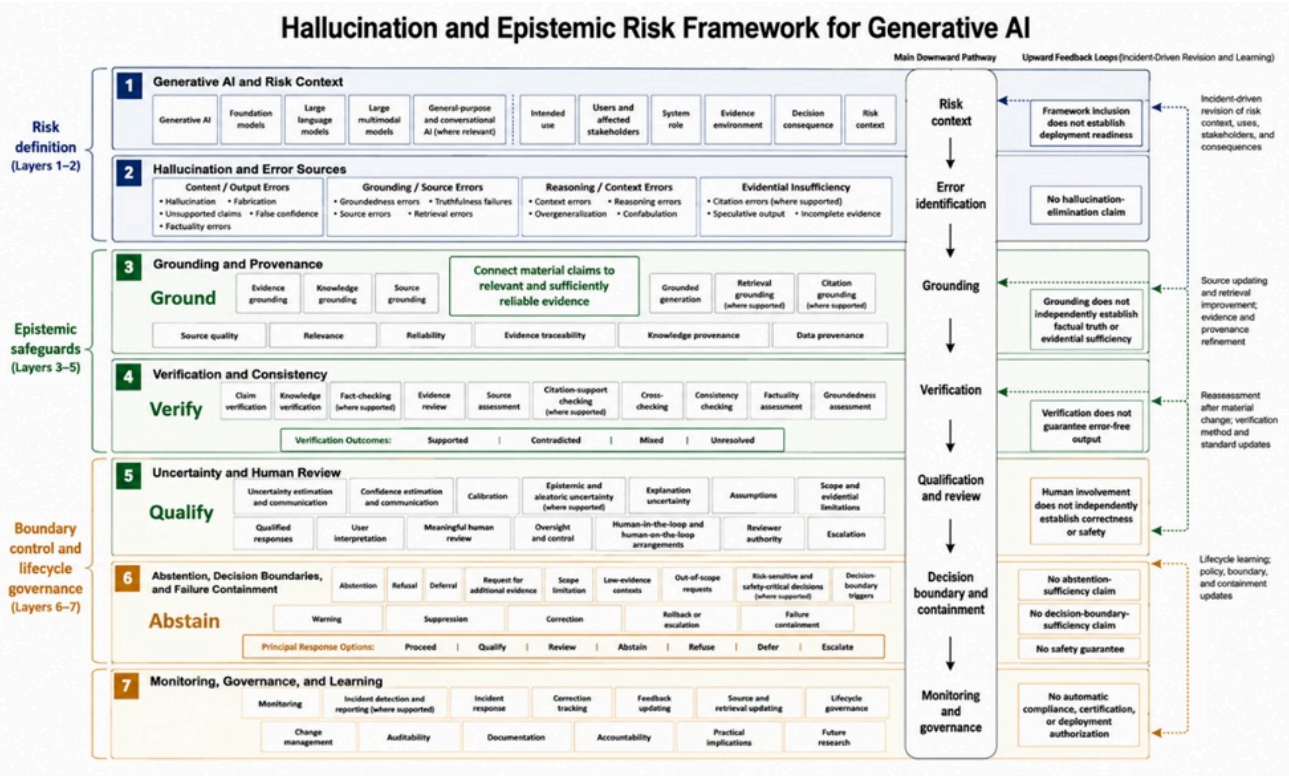


Figure 2. Hallucination and Epistemic Risk Framework for Generative AI

Alt-text description.

A conceptual generative AI safety framework showing hallucination risk, error sources, grounding, verification, uncertainty communication, human review, abstention, decision boundaries, failure containment, monitoring, incident response, and governance updating.

Practical Implications and Future Research

Future hallucination-risk assessment should distinguish factuality, groundedness, citation support, reasoning validity, confidence expression, and decision consequence rather than reporting one undifferentiated error category. The conceptual distinctions developed in hallucination research can support claim-level evaluation, but measures should be tested across domains, languages, interaction formats, and changing model versions [5]. Foundation-model factuality research should also examine temporal knowledge, rare facts, conflicting evidence, and the incentives created by evaluation procedures. Multimodal research is needed because cross-modal inconsistencies may not be captured by text-centered detection methods.

Grounding research should evaluate the complete evidence pathway rather than whether retrieval was merely used. Relevant questions include whether suitable evidence existed, whether the retrieval process found it, whether sources were reliable and independent, whether the system represented them faithfully, and whether cited evidence entailed the generated claim. Retrieval-supported generation remains exposed to missing evidence, irrelevant passages, source manipulation, and defective context construction [13]. Verification workflow studies should consequently evaluate claim extraction, evidence retrieval, source appraisal, cross-checking, judgment, and unresolved-case handling as separate but interacting stages.

Human-centered research should investigate how users interpret qualified answers, source limitations, uncertainty language, abstention, and escalation. Linguistic calibration may improve confidence communication, but its effectiveness depends on context and user interpretation [18]. Human-review research should measure reviewer expertise, workload, independence, evidence access, disagreement handling, and the ability to override outputs. Abstention and refusal evaluation should consider not only error reduction but also coverage, fairness, usefulness, delay, and whether deferred users can access an appropriate alternative process.

Lifecycle research should examine how incident monitoring, correction, source updating, model change, and feedback affect subsequent risk. Assurance cannot end with an initial evaluation because model behavior, retrieval resources, interfaces, and use contexts can change [25]. These implications are conceptual research directions rather than legal advice, compliance guidance, procurement or audit instructions, model- or platform-selection advice, safety certification, clinical, financial, educational, or legal decision guidance, or deployment instruction. Operational requirements must be developed and validated for each domain, institution, user population, and consequence structure.

Conclusion

This article proposes a conceptual generative AI safety framework organized around grounding, qualification, verification, and abstention, supported by human review, decision boundaries, failure containment, monitoring, and lifecycle governance. Its central premise is that fluent, complete, confident, or useful output is not necessarily evidence-grounded, appropriately uncertain, verifiable, or safe to rely upon. Epistemic safety therefore requires attention to what a system claims, what evidence supports the claim, how limitations are communicated, when independent review is meaningful, when answering should be constrained, and how failures are detected and contained. The framework can help structure research and domain-specific assessment, but it does not eliminate hallucination or independently establish factuality, calibrated trust, ethical sufficiency, legal compliance, operational safety, certification, or deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Jul 2025

Revised: 12 Sep 2025

Accepted: 06 Nov 2025

Published online: 15 March 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Bus Inf Syst Eng*. 2024;66(1):111-26.
<https://doi.org/10.1007/s12599-023-00834-7>.
- Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach*. 2020;30(4):681-94.
<https://doi.org/10.1007/s11023-020-09548-1>.
- Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: A review. *ACM Comput Surv*. 2023;55(2):1-38.
<https://doi.org/10.1145/3491209>.
- Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. *Electron Mark*. 2021;31(2):447-64.
<https://doi.org/10.1007/s12525-020-00441-4>.
- Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55(12):1-38.
<https://doi.org/10.1145/3571730>.
- Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst*. 2025;43(2):1-55.
<https://doi.org/10.1145/3703155>.
- Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. *Nature*. 2024;630(8017):625-30.
<https://doi.org/10.1038/s41586-024-07421-0>.
- Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. *Inf Fusion*. 2023;99:101861.
<https://doi.org/10.1016/j.inffus.2023.101861>.
- van Deemter K. The pitfalls of defining hallucination. *Comput Linguist*. 2024;50(2):807-16.
https://doi.org/10.1162/coli_a_00509.
- Kalai AT, Nachum O, Vempala SS, Zhang E. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*. 2026;653(8116):1047-51.
<https://doi.org/10.1038/s41586-026-10549-w>.
- Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. *Nature*. 2024;631(8022):755-9.
<https://doi.org/10.1038/s41586-024-07566-y>.
- Miao MM, Kearns M. Hallucination, monofacts, and miscalibration: An empirical investigation. *Proc Natl Acad Sci U S A*. 2026;123(8):e2533582123.
<https://doi.org/10.1073/pnas.2533582123>.
- Wu S, Xiong Y, Cui Y, Wu H, Chen C, Yuan Y, et al. Retrieval-augmented generation for natural language processing: A survey. *Artif Intell Rev*. 2026;59:192.

<https://doi.org/10.1007/s10462-026-11605-7>.

Augenstein I, Baldwin T, Cha M, Chakraborty T, Ciampaglia GL, Corney D, et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat Mach Intell*. 2024;6(8):852-63.

<https://doi.org/10.1038/s42256-024-00881-z>.

Zeng X, Abumansour AS, Zubiaga A. Automated fact-checking: A survey. *Lang Linguist Compass*. 2021;15(10):e12438.

<https://doi.org/10.1111/lnc3.12438>.

Guo Z, Schlichtkrull M, Vlachos A. A survey on automated fact-checking. *Trans Assoc Comput Linguist*. 2022;10:178-206.

https://doi.org/10.1162/tac1_a_00454.

Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach Learn*. 2021;110(3):457-506.

<https://doi.org/10.1007/s10994-021-05946-3>.

Mielke SJ, Szlam A, Dinan E, Boureau YL. Reducing conversational agents' overconfidence through linguistic calibration. *Trans Assoc Comput Linguist*. 2022;10:857-72.

https://doi.org/10.1162/tac1_a_00494.

Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digit Med*. 2021;4(1):4.

<https://doi.org/10.1038/s41746-020-00367-3>.

Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: A state of the art. *Artif Intell Rev*. 2023;56(4):3005-54.

<https://doi.org/10.1007/s10462-022-10246-w>.

Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. *Acad Manag Ann*. 2020;14(2):627-60.

<https://doi.org/10.5465/annals.2018.0057>.

Hendrickx K, Perini L, Van der Plas D, Meert W, Davis J. Machine learning with a reject option: A survey. *Mach Learn*. 2024;113(5):3073-110.

<https://doi.org/10.1007/s10994-024-06534-x>.

Schuster D. Abstaining machine learning: Philosophical considerations. *AI Soc*. 2025;40(6):4213-33.

<https://doi.org/10.1007/s00146-025-02188-y>.

Burton S, Habli I, Lawton T, McDermid J, Morgan P, Porter Z. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artif Intell*. 2020;279:103201.

<https://doi.org/10.1016/j.artint.2019.103201>.

Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Comput Surv*. 2021;54(5):1-39.

<https://doi.org/10.1145/3453444>.

Fügner A, Grahl J, Gupta A, Ketter W. Cognitive challenges in human-artificial intelligence collaboration: Investigating the path toward productive delegation. *Inf Syst Res*. 2022;33(2):678-96.

<https://doi.org/10.1287/isre.2021.1079>.

Shneiderman B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int J Hum Comput Interact*. 2020;36(6):495-504.

<https://doi.org/10.1080/10447318.2020.1741118>.

Dellermann D, Ebel P, Söllner M, Leimeister JM. Hybrid intelligence. *Bus Inf Syst Eng*. 2019;61(5):637-43.

<https://doi.org/10.1007/s12599-019-00595-2>.

Raisch S, Krakowski S. Artificial intelligence and management: The automation-augmentation paradox. *Acad Manage Rev*. 2021;46(1):192-210.

<https://doi.org/10.5465/amr.2018.0072>.

Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389-99.

<https://doi.org/10.1038/s42256-019-0088-2>.

Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. 2019;1(11):501-7.

<https://doi.org/10.1038/s42256-019-0114-4>.