Journal Title
XX(X):2–37

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A Survey on Neuro-Symbolic Auditing: A Framework for Verification, Traceability, and Correction in High-Stakes AI

Xiaoming Guo[1, 2], Shenglin Li[1, 2], Jiacheng Cao[1, 2] and Jiaqi Gong[1, 2]

Abstract

The rapid deployment of AI in high-stakes domains has outpaced our ability to technically audit model behavior for safety, compliance, and accountability. Existing governance standards largely verify processes rather than validating the correctness of system decisions at inference time, leaving a persistent audit gap for opaque neural models. This survey frames Neuro-Symbolic AI (NSAI) as an enabling approach for inherently auditable systems by combining neural learning with explicit symbolic constraints and reasoning. Following a systematic literature review process, we introduce the VTC auditing framework, Verification, Traceability, and Correction, and use it to construct a comparative taxonomy that maps NSAI architectures to concrete auditing capabilities. We show how symbolic components can (i) formally enforce rules and safety constraints, (ii) generate human-interpretable decision trails suitable for audit logs, and (iii) support targeted correction via constraint updates and runtime mitigation. We synthesize evidence across high-stakes applications (e.g., clinical decision support, financial compliance, and autonomous systems), identify recurring design tradeoffs, and highlight open challenges for auditing LLMs and agentic systems. The resulting taxonomy and design guidance provide a roadmap for selecting NSAI architectures based on domain-specific verification and accountability requirements.

Keywords

Neuro-symbolic AI, AI auditing, high-stakes AI, verification, traceability, accountability, explainable AI

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Introduction

The trajectory of Artificial Intelligence (AI) development between 2020 and 2026 has been defined by a profound paradox: unprecedented advancements in capability alongside escalating failures in reliability. The widespread deployment of Large Language Models (LLMs) and generative architectures has revolutionized computational infrastructure, embedding stochastic decision-making engines into the foundational systems of healthcare, finance, transportation, and legal domains. However, this rapid integration has precipitated an algorithmic trust crisis. As institutions moved from experimental sandboxes to critical deployment, the fundamental inability of purely neural (sub-symbolic) architectures to guarantee safety, factual adherence, and regulatory compliance became starkly apparent Raji et al. (2020); Rudin (2019); Marcus (2020).

The black box nature of deep learning, the opacity of decision pathways distributed across billions of parameters, collides directly with the rigid, deterministic requirements of physical safety and legal accountability, making a transition from theoretical risks to tangible harms Rudin (2019). High-profile incidents, such as the systematic denial of essential care by opaque healthcare algorithms, catastrophic failures in autonomous transportation, and the emergence of real legal liability stemming from AI-generated misinformation, underscore this disconnect. These incidents were not merely technical bugs, but structural inevitabilities of an AI paradigm that prioritized probabilistic correlation over symbolic verification.

In response, oversight frameworks like the NIST AI Risk Management Framework Zurawski and Schopf (2023) and ISO/IEC 42001 have emerged. However, while these standards provide essential procedural auditing at the governance layer, verifying that risk-management processes exist, they fail to reach the inference layer to mathematically validate the correctness of a neural network's decisions, leaving a critical audit gap. Traditional auditing methods, often reduced to post-hoc statistical evaluations or superficial procedural checklists, are insufficient for governing autonomous agents. This gap often results in audit washing, in which superficial checks or process audits that merely confirm the presence of governance policies are substituted for the deeper technical audits required to verify an AI system's underlying reasoning and behavior, as discussed by Raji et al. (2020).

This research serves as a systematic survey of Neuro-Symbolic Auditing, an emerging discipline that seeks to bridge the chasm between the pattern-recognition power of

[1]Department of Computer Science, University of Alabama, Tuscaloosa, Alabama, USA; [2]Alabama Center for the Advancement of Artificial Intelligence, Tuscaloosa, Alabama, USA

Corresponding author:
Jiaqi Gong, Alabama Center for the Advancement of Artificial Intelligence, Department of Computer Science, University of Alabama, Tuscaloosa, Alabama, USA.
Email: jiaqi.gong@ua.edu

Prepared using sagej.cls

neural networks and the logical rigor required for accountability. We argue that effective auditing requires AI architectures that are designed to be inherently governable. This survey defines the key components of this domain:

High-Stakes Artificial Intelligence refers to AI systems deployed in critical infrastructure where algorithmic failure can result in physical injury (e.g., autonomous transport), denial of essential services (e.g., healthcare administration), massive market volatility, or the erosion of judicial integrity.

Auditing denotes the architectural capacity for transparency and interpretability. This necessitates a Glass Box approach, enabling auditors to inspect internal symbolic logic, explain why a specific decision was reached, and verify the reasoning process itself.

Neuro-Symbolic AI (NSAI) represents an architectural approach that synthesizes the learning and perception capabilities of neural networks with the explicit knowledge representation and deductive reasoning of symbolic systems Garcez and Lamb (2023); Kautz (2022). NSAI aims to overcome the opacity and brittleness of deep learning while addressing the scalability and knowledge acquisition bottlenecks of traditional symbolic AI.

This paper categorizes and compares literature based on how NSAI architectures enable a technical Auditing framework across three technical pillars Raji et al. (2020); Yao et al. (2023):

Verification: The ability to formally prove that a system adheres to a specified rule, constraint, or regulation, even when the underlying neural model is probabilistic. This is achieved by integrating symbolic components, such as theorem provers or logic-based constraint mechanisms, to verify the logical validity of neural outputs, thereby ensuring adherence to safety boundaries and mitigating hallucinations risks.

Traceability: The capacity to log the precise, human-intelligible chain of reasoning that leads to a decision. By utilizing symbolic representations, NSAI systems can produce a detailed audit log, transforming the black box into a glass box and allowing auditors to reconstruct exactly which rules or inferences contributed to an outcome Gunning et al. (2019).

Correction: The mechanism to modify a model's behavior or update its knowledge to fix errors in pure deep learning. NSAI enables surgical editing of symbolic constraints, logic-based feedback mechanisms, and runtime mitigation strategies that maintain alignment with domain rules and ensure safe system behavior under failure conditions.

By synthesizing state-of-the-art applications in High-Stakes Domains, including clinical support, financial compliance, and autonomous systems, this survey maps the transition from procedural oversight to technical accountability. We provide a taxonomy that assists researchers in selecting NSAI architectures based on the specific verification requirements of their domain. The remainder of this survey systematically explores the landscape of Neuro-Symbolic Auditing and is organized as follows. Section 2 details the systematic review methodology, documenting our search strategy and selection process in compliance with the PRISMA 2020 statement. Section 3 establishes the foundations of AI auditing within high-stakes domains, defining the Audit Gap and the VTC (Verification, Traceability, and Correction) framework. Section 4 presents a comprehensive taxonomy of the Neuro-Symbolic Architecture Landscape, mapping

specific hybrid designs to their respective auditing capabilities. Section 5 focuses on the frontier of Neuro-Symbolic Auditing for Large Language Models (LLMs) and agentic systems, while Section 6 illustrates these concepts through practical applications in high-stakes domains. Finally, Section 7 identifies open challenges and future research directions, and Section 8 concludes the paper with a summary of the path toward inherently auditable AI.

Systematic Review Methodology

Research Objective and Scope

To ensure a rigorous and reproducible synthesis of the literature, this survey follows the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement Page et al. (2021). To address the escalating reliability crisis in AI deployment, this systematic review focuses on the technical intersection of Neuro-Symbolic AI (NSAI) and Algorithmic Auditing. Unlike traditional oversight methods that rely on external governance policies, our methodology prioritizes the technical auditing capacity of hybrid systems-specifically the ways in which integrating symbolic logic with neural learning creates structural handles for accountability.

We prioritize literature that demonstrates how NSAI architectures provide mathematically grounded solutions for High-Stakes Domains (such as healthcare, finance, and autonomous systems). Consequently, our review is structured to categorize and compare papers according to their ability to meet the VTC (Verification, Traceability, and Correction) framework. We specifically examine how symbolic components enable Verification of safety constraints, provide Traceability through human-intelligible reasoning chains (transitioning from mere Interpretability to evidence-based Accountability), and allow for surgical Correction of model behavior without the opacity of complete retraining.

Search Strategy and Information Sources

We conducted a systematic search across three primary academic databases: IEEE Xplore, ACM Digital Library, and Google Scholar, targeting literature published between 2020 and 2026. This timeframe was selected to capture the current era of AI, which is increasingly defined by the convergence of deep learning and symbolic reasoning. To ensure a balance between high precision and functional recall, the search strategy was organized around three core thematic pillars:

Target Technology: Neuro-symbolic AI, neural-symbolic integration, and hybrid AI architectures.

Auditing Mechanisms: Verification, traceability, interpretability, and formal accountability.

Application Context: High-stakes domains, including healthcare, safety-critical systems, and finance.

While the specific syntax was adapted to the technical requirements of each database (e.g., title-specific filtering or abstract indexing), the underlying keyword logic remained consistent.

Eligibility Criteria

Papers were evaluated based on a strict set of technical requirements. To be included, a study must fulfill the NSAI-Audit coupling requirement: it must propose or evaluate a Neuro-Symbolic architecture specifically to address at least one of the three auditing pillars: Verification, Traceability, or Correction (VTC).

Inclusion Criteria: (1) Peer-reviewed primary research; (2) Integration of neural and symbolic components; (3) Explicit focus on high-stakes application domains (Healthcare, Finance, Autonomous Systems).

Exclusion Criteria: (1) Purely procedural auditing frameworks (e.g., policy checklists) without technical inference-layer mechanisms; (2) Insufficient technical detail for architectural classification.

Selection and Data Extraction

The selection process was systematically executed across three distinct phases, beginning with the identification stage, where initial results from database queries were aggregated, and duplicate records were eliminated. This was followed by the screening phase (first pass), during which titles and abstracts were scrutinized for relevance to technical auditing, leading to the exclusion of generic AI performance papers that failed to address core themes of reliability or transparency. Finally, the eligibility phase (deep dive) involved a rigorous full-text review of the remaining articles against established inclusion and exclusion criteria. This phase also incorporated both backward and forward snowballing (citation chaining) to identify foundational hybrid models that used functional synonyms for auditing, thereby ensuring comprehensive capture of the technical landscape. Table 1 presents the number of records retrieved from each database or search engine for publications between Jan 2020 and Jan 2026.

Table 1. Search Strategy Results: Retrieval, Deduplication, and Eligibility.

Database/Search Engine	Retrieved	Deduplicated	Eligibility
IEEE Xplore	29	29	18
ACM Digital Library	23	23	13
Google Scholar	143	139	119
Total	195	191	**150**

Foundations of AI Auditing in High-Stakes Domains

Effective governance of AI requires a precise definition of the domain scope and the technical mechanisms required to ensure accountability. While auditing is often used

colloquially to refer to any form of model evaluation, in the context of responsible automation, it refers to a rigorous, evidence-based verification process. This section establishes the conceptual foundations of this survey, defining the specific requirements of High-Stakes environments and detailing the VTC framework.

High-Stakes Domains and the Definition of Auditing

We define High-Stakes AI as automated systems deployed in critical infrastructure that result in tangible, irreversible harms rather than merely user dissatisfaction. These domains typically include:

Physical Safety: Autonomous transportation, robotic surgery, and industrial control systems, where failure risks human injury or death.

Essential Services: Healthcare diagnostics, insurance underwriting, and credit scoring, where denial of service impacts fundamental livelihood.

Legal and Epistemic Integrity: Automated evidence processing and generative media, where the erosion of truth can compromise judicial processes Citron and Pasquale (2014).

In these environments, the probabilistic guarantees of deep learning are often insufficient. There is a structural conflict between the opaque nature of neural networks, which distribute reasoning across billions of parameters, and the strict liability requirements of accountability. Consequently, we define auditing not merely as a governance checklist but as a technical capability: the architectural capacity to inspect, verify, and correct a system's internal reasoning. As noted by Raji et al., effective auditing must close the accountability gap by moving beyond superficial audit washing to deep technical audits that validate specific behavioral constraints Raji et al. (2020).

The Normative Goals of Auditing

While the definition of auditing provides the scope, the practice is guided by three normative principles that high-stakes systems must satisfy. These principles form the objective function for the technical pillars discussed later:

- **Fairness:** The requirement that algorithmic decisions do not discriminate against protected groups. In auditing, this translates to the need to prove adherence to equity axioms (e.g., Equalized Odds) despite potentially biased training data Landers and Behrend (2023).
- **Transparency:** The requirement that the system's decision-making process is intelligible to human stakeholders. This demands faithful provenance, ensuring that the explanation provided to the auditor reflects the actual computation performed by the model.
- **Safety:** The requirement that the system operates within defined boundary conditions and does not cause physical or economic harm. This demands guarantees that catastrophic states are unreachable, regardless of input distribution.

The Socio-Technical Audit Lifecycle

Auditing is not a static checkpoint but a continuous lifecycle activity. Drawing on the SMACTR framework (Scoping, Mapping, Artifact Collection, Testing, Reflection) Raji et al. (2020), we view the auditing lifecycle as spanning four distinct stages, each requiring specific technical artifacts:

Data Provenance (Pre-Training): Auditing begins with the substrate of accountability. This stage utilizes standardized documentation frameworks, most notably Datasheets for Datasets Gebru et al. (2021), which mandate the disclosure of a dataset's origins, potential biases, and intended use cases to identify systemic risks before they are encoded into a model.

Model Development (Pre-Deployment): During training, auditors use Model Cards Mitchell et al. (2019), standardized reports that disclose a model's operational constraints and performance benchmarks, to disaggregate performance across demographic groups, testing for fairness criteria such as equalized odds Hardt et al. (2016) and assessing adversarial robustness Kurakin et al. (2017).

Runtime Monitoring (Deployment): Once deployed, systems must be monitored for distribution shift. This requires mechanisms such as interval observers Lan et al. (2024) to maintain bounds on system behavior in real time.

Auditability (Post-Incident): When failures occur, systems must support provenance tracking. Tools like Titian Interlandi et al. (2015) enable auditors to reconstruct the decision chain to identify which input record caused a specific error.

While these frameworks provide the procedural layer of auditing, high-stakes domains require technical guarantees that purely neural architectures struggle to provide.

The Technical Pillars of Auditing: The VTC Framework

To bridge the gap between procedural governance and technical reality, we propose that an auditable AI architecture must support three core technical capabilities: Verification, Traceability, and Correction.

Verification: Proving Correctness Verification addresses the question: *Can we mathematically guarantee that a specific property holds for all possible inputs?*

Unlike empirical testing, which only covers a finite set of samples, formal verification provides a proof of adherence to safety constraints. Significant progress has been made in neural verification. Reluplex Katz et al. (2017) introduced Satisfiability Modulo Theories (SMT) solvers capable of handling the non-convex constraints of ReLU activation functions. This was extended by Marabou Katz et al. (2019), which utilizes symbolic bound tightening, and Beta-CROWN Wang et al. (2021), which leverages efficient bound propagation. By utilizing formal methods, these tools can provide a mathematical guarantee that a safety property, such as an aircraft collision avoidance system always issuing a turn command when an intruder enters a predefined safety radius, will hold true across all possible input scenarios, rather than merely passing a finite set of test cases. However, a fundamental limitation remains: pure neural verification scales poorly. The complexity of verifying deep networks grows exponentially with depth and width,

making it intractable for modern Large Language Models (LLMs) without significant abstraction.

Traceability: The Fidelity Gap Traceability addresses the question: *Can we reconstruct the precise causal chain that led to a decision?*

This requirement for traceability diverges significantly from standard interpretability, which frequently relies on post-hoc approximations that may not reflect the model's actual internal logic. Current post-hoc explanation methods, such as LIME Ribeiro et al. (2016) and SHAP Lundberg and Lee (2017), attempt to explain black-box models by fitting simpler, interpretable models locally around a prediction. While useful for debugging, these methods suffer from a fidelity gap. As Lipton argues, post-hoc explanations are distinct from simulatability, understanding the actual mechanism of the model Lipton (2018). In high-stakes contexts, an approximate explanation that obscures the true decision boundary is legally insufficient. True traceability requires architectural transparency, where the reasoning steps are generated as a byproduct of the computation itself.

Correction: The Ability to Fix Correction addresses the question: *Can we reliably modify system behavior without full retraining?*

In purely neural systems, fixing a specific error (e.g., a bias against a specific demographic or a hallucinated fact) usually requires retraining, which is computationally expensive and risks catastrophic forgetting. Auditable systems require surgical editing capabilities. Techniques such as shielding in reinforcement learning Alshiekh et al. (2018) utilize symbolic logic to override unsafe actions proposed by a neural policy, effectively correcting the system at runtime. Similarly, knowledge-guided frameworks such as CREST Choi et al. (2023) enable targeted updates to decision logic by modifying the underlying knowledge graph, ensuring that regulatory updates can be implemented immediately.

The Neuro-Symbolic Necessity

The VTC framework highlights the limitations of purely neural networks AI. While neural networks excel at perception, they lack the discrete structure required for efficient verification, faithful traceability, and modular correction. Conversely, symbolic systems offer these traits but lack scalability. This creates the architectural imperative for Neuro-Symbolic AI (NSAI): architectures that synthesize the learning capabilities of neural networks with the explicit knowledge representation of symbolic logic. To visualize this integration, Figure 1 presents the conceptual framework of a neuro-symbolic auditing system, illustrating how a symbolic auditor interacts with the machine learning pipeline to enforce verification, traceability, and correction across high-stakes domains. The following section categorizes how diverse NSAI architectures operationalize these auditing goals by synthesizing neural learning with symbolic reasoning.

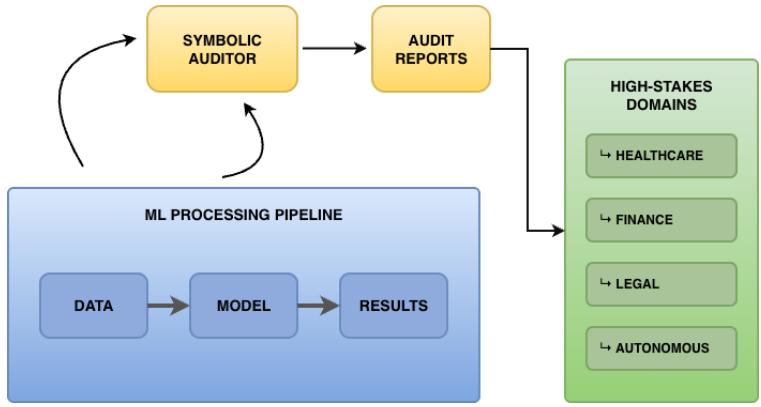

Figure 1. Conceptual framework of a Neuro-Symbolic Auditing System. The diagram illustrates the integration of a Symbolic Auditor component alongside the standard machine learning processing pipeline. Through active feedback loops, the symbolic component monitors the pipeline to generate dynamic, traceable, and verifiable audit reports. This architecture allows auditors to detect errors or irregularities in real-time across the lifecycle, enabling intervention and correction to ensure reliability in high-stakes domains such as healthcare, finance, legal, and autonomous systems.

The Neuro-Symbolic Architecture Landscape

To bridge the gap between probabilistic capability and the VTC auditing requirements defined in the previous section, we propose the integration of Neuro-Symbolic AI (NSAI). As described by Garcez and Lamb, NSAI represents the third wave of AI, aiming to synthesize the robust perception of neural networks with the explicit reasoning of symbolic logic Garcez and Lamb (2020); Kautz (2022). From an auditing perspective, NSAI is not merely a performance booster but a structural necessity. It creates architectural hooks, discrete concepts, logical constraints, or proof trees, that allow auditors to verify and correct system behavior. We organize the landscape into five families, categorizing them not by their learning dynamics.

Neural to Symbolic Pipelines: The Interpretable Interface

The most direct approach to traceability is the sequential decomposition of perception and reasoning. In these architectures, a neural network acts as a perception module, mapping raw inputs (e.g., images) to a symbolic vocabulary, which is then processed by a deterministic symbolic reasoner.

Representative Systems: The Concept Bottleneck Model (CBM) Koh et al. (2020) forces the network to predict human-aligned concepts (e.g., wing color, beak shape) before a linear classifier makes the final prediction. Similarly, the Neuro-Symbolic Concept Learner (NSCL) Mao et al. (2019) functions by mapping visual scenes to object-based latent representations and natural language questions to executable, symbolic programs.

Audit Capabilities:

Verification (Low): While the symbolic reasoning component is verifiable, the initial neural perception layer remains probabilistic and opaque. We cannot mathematically guarantee the model will correctly identify red wing in all lighting conditions or adversarial scenarios.

Traceability (High): The reasoning phase is a Glass Box. Any error can be traced definitively to a specific failure point: either the neural perception (wrong concept detected) or the symbolic rule (wrong logic applied).

Correction (High): CBMs allow for test-time intervention. If a model misclassifies a bird because it detected the wrong wing color, an auditor can manually correct the concept, forcing the model to re-evaluate the decision based on the corrected fact. Recent work Yuksekgonul et al. (2023) extends this to post-hoc settings using multimodal embeddings.

Logic-Informed Neural Networks: The Constraint Engine

Rather than separating components, this family embeds logical constraints directly into the neural loss function. The network remains the primary inference engine, but its decision boundary is shaped by logical axioms.

Representative Systems: Logic Tensor Networks (LTN) Serafini and Garcez (2016); Badreddine et al. (2022) implement Real Logic, where predicates are grounded as differentiable operations. Logical Neural Networks (LNN) Riegel et al. (2020) enforce a one-to-one correspondence between neurons and logical gates, allowing the network to function as a differentiable theorem prover Sen et al. (2022).

Audit Capabilities:

Verification (Medium): These architectures support soft verification. Auditors can define safety axioms and penalize the model for violations during training. Although this approach does not provide the absolute guarantees of formal verification methods, it statistically aligns the model with domain-specific rules.

Traceability (Medium): In LNNs, traceability is higher because weights correspond to logical truth values. In LTNs, reasoning is distributed across fuzzy truth values, making it harder to isolate a discrete reasoning chain compared to sparse symbolic systems.

Correction (Low): Behavior is corrected by modifying the logical constraints in the loss function. However, unlike pipelines or KGs, these changes require retraining or fine-tuning the model to take effect, preventing real-time surgical edits.

Differentiable Logic: The Provenance Engine

These systems embed neural networks into logic programming languages. The symbolic reasoning process itself is differentiable, allowing gradients to flow from the final proof back to the neural perception modules.

Representative Systems: DeepProbLog Manhaeve et al. (2018, 2021) and Scallop Li et al. (2023) treat neural outputs as probabilistic facts. The system constructs a probabilistic proof tree for every query.

Audit Capabilities:

Verification (Medium): The logic program structure is verified, ensuring valid conclusions follow from premises. However, because the facts feeding the logic are probabilistic neural outputs, the system verifies distributions rather than definite outcomes.

Traceability (High): Unlike attention maps, which only show where a model looked, these systems generate the provenance of the decision. The explanation is the valid proof tree used to derive the answer.

Correction (Medium): Rules can be edited directly. Adding a new logical rule immediately updates the system's behavior without requiring full retraining of the neural predicates.

Symbolic-Neural: The Safety Verifier

In this solver-centric approach, a symbolic theorem prover is the primary decision-maker, using a neural network only as a heuristic guide to search the solution space.

Representative Systems: AlphaGeometry Trinh et al. (2024) uses a language model to suggest auxiliary constructions for geometry problems, which a symbolic engine then attempts to prove. Neural Logic Machines Dong et al. (2019) use tensorized logic for inductive reasoning.

Audit Capabilities:

Verification (High): This represents the gold standard for safety-critical auditing. Because the final output is generated by a symbolic solver, it is mathematically correct by definition. The neural network cannot cause a hallucination in the final result, it can only fail to find a solution.

Traceability (High): The output is a formal proof. Every step is explicit and mathematically valid.

Correction (Low): While the symbolic engine is trustworthy, the neural guide is a black box trained on massive datasets. If the system fails to find a solution, correcting the neural heuristic to find the right path is difficult and requires retraining.

Knowledge-Grounded Models: The Fact Checker

This family grounds LLMs in structured Knowledge Graphs (KGs) and uses them as external, editable memory sources.

Representative Systems: GreaseLM Zhang et al. (2022) fuses representations from Transformers and Graph Neural Networks (GNNs) to reason over joint text-structure modalities.

Audit Capabilities:

Verification (Low): While the KG contains true facts, the neural aggregation mechanism (GNN + LLM) is opaque. We cannot guarantee the model will not ignore the KG and hallucinate, requiring external verification mechanisms.

Traceability (High): Reasoning can be traced to specific paths in the Knowledge Graph. Auditors can identify exactly which entities and relations were retrieved to inform the prediction.

Correction (High): These systems address the Knowledge Cutoff problem. If the facts change (e.g., a regulatory update), auditors can edit the Knowledge Graph, and the model's reasoning is corrected instantly. This separation of Reasoning (Neural) and Knowledge (Symbolic) is crucial for long-term maintenance Liu et al. (2024).

Summary: The VTC Capability Map

Table 2 maps these architectures to the VTC framework. The analysis reveals a trade-off: architectures that offer absolute verification such as, Symbolic-Neural, generally struggle with broad scalability, while scalable pipeline models rely on the empirical accuracy of their concept detectors.

Table 2. The NSAI Capability Map: Mapping Architectures to Audit Requirements.

Architecture Family	Verification	Traceability	Correction	Example Models	Primary Utility
Neural-Symbolic	low	High	High	CBM, NSCL	Human-in-the-loop
Logic-Informed	Medium	Medium	low	LTN, LNN	Domain Constraints
Differentiable Logic	Medium	High	Medium	DeepProbLog, Scallop	Reasoning Paths
Symbolic-Neural	High	High	Low	AlphaGeometry, NLM	Safety Certification
KG-Grounded	Low	High	High	GreaseLM	Knowledge Management

The architectures detailed in this section represent intrinsic neuro-symbolic design, systems built from the ground up to be auditable. However, the current industrial landscape is increasingly dominated by pre-trained LLMs, where internal architectural modifications are often impossible. This introduces a scalability gap: formal verification methods that work for bounded neural networks do not scale to billion-parameter generative models. Next section addresses this challenge by examining how neuro-symbolic techniques can be applied extrinsically as wrappers and guardrails to audit the stochastic frontiers of LLMs and Agentic Systems.

Neuro-Symbolic Auditing for Large Language Models and Agentic Systems

LLMs represent both the most capable and the most challenging targets for AI auditing. Their monolithic neural architecture lack the structural affordances necessary for verification, specifically, inspectable intermediate representations and formal specifications, yet their deployment across high-stakes domains (medicine, law, autonomous systems) necessitates rigorous auditing mechanisms. This creates a scalability gap: traditional VTC mechanisms designed for bounded networks are largely inapplicable at the billion-parameter scale.

The Stochastic Audit Challenge. The necessity for neuro-symbolic intervention is not merely due to the scale of these models, but due to two structural failure modes that purely neural auditing fails to resolve:

- **The Detection-Correction Gap**: While detection methods like Semantic Entropy Farquhar et al. (2024) can identify hallucinations post-hoc, they suffer from a critical VTC Asymmetry: they achieve high Verification (detecting errors) but zero

Correction (preventing them). As noted by Ji et al. (2023), distinguishing intrinsic hallucinations from extrinsic ones requires external knowledge bases, suggesting that statistical detection is insufficient without symbolic grounding.
- **The Structural Failure of Safety**: Safety failures in LLMs are often structural, arising from competing objectives (e.g., helpfulness vs. harmlessness) that Reinforcement Learning from Human Feedback (RLHF) cannot fully resolve Wei et al. (2022). Recent benchmarks, such as AgentHarm Andriushchenko et al. (2024), demonstrate that even aligned models remain compliant with malicious requests when tasks are decomposed into seemingly benign steps. This gap between learned safety and guaranteed safety necessitates formal enforcement mechanisms.

The Solution: The Neuro-Symbolic Wrapper. To bridge these gaps, the field has coalesced around the Neural-Symbolic wrapper strategy. Instead of attempting to make the LLM itself symbolic, researchers embed the model within a symbolic control structure. This approach treats the LLM as an untrusted generator, while symbolic components handle verification, constraints, and correction. We organize these auditing mechanisms into three phases based on when the symbolic intervention occurs relative to generation:

- **Phase 1: Pre-Generation Grounding (Traceability)**: Anchoring LLM inputs to symbolic knowledge to ensure source attribution.
- **Phase 2: In-Generation Constraints (Verification)**: Restricting the decoding process to enforce structural validity.
- **Phase 3: Post-Generation Verification (Correction)**: Auditing outputs before execution to enable surgical repair.

The following subsections survey these methods, organizing them by their intervention phase

Pre-Generation Grounding Methods

The most developed neuro-symbolic approach grounds neural generation in external symbolic structures, knowledge graphs, logical rules, and formal constraints, that provide verification handles absent in the language model itself. This subsection surveys methods that operate *before* LLM generation begins.

Retrieval-Augmented Generation Retrieval-Augmented Generation Lewis et al. (2020) established the foundational pattern combining parametric knowledge with non-parametric retrieval. From an auditing perspective, RAG's primary contribution is traceability through source attribution rather than Verification, the LLM can still ignore or contradict retrieved content. Peng et al. (2024) survey GraphRAG approaches using knowledge graph structure for multi-hop reasoning with explicit path provenance. formalizing the workflow into indexing, retrieval, and generation stages while highlighting challenges in subgraph candidate exploration and structure-aware

similarity measurement. These systems provide moderate verification through fact-checking against retrieved content, high traceability through source attribution, and moderate correction since updating knowledge sources affects future generations, though LLM parametric knowledge remains fixed, creating potential conflicts.

KG-Enhanced Reasoning Pan et al. (2024) distinguish three integration paradigms: KG-enhanced LLMs, LLM-augmented KGs, and synergized systems. The recent literature brings transformative advances representing a qualitative shift from fewer errors to zero errors. Graph-Constrained Reasoning (Luo et al. 2024) introduces the KG-Trie pattern: by integrating knowledge graph structure into decoding through trie-based constraints, GCR ensures generated reasoning paths correspond exactly to valid KG paths. The results are striking: zero reasoning hallucination compared to 33% for baselines like Reasoning on Graphs, with zero-shot generalizability to unseen knowledge graphs. Crucially, hallucination is eliminated by construction, not reduced statistically, a paradigm shift from probabilistic improvement to structural guarantee.

Paths-over-Graph (Tan et al. 2025) demonstrates that structured reasoning augmentation can compensate for model capability: PoG with GPT-3.5 surpasses the strong baseline Think-on-Graph (ToG) with GPT-4 by up to 23.9%, challenging the scale solves everything assumption. The approach achieves state-of-the-art results across tested KGQA datasets, outperforming ToG by an average of 18.9% accuracy through three-phase dynamic path exploration. StructGPT Jiang et al. (2023) extends grounding to step-by-step reasoning, linking each step to external structured data (e.g., KGs, tables), achieving significant improvement over CoT baselines through its Iterative Reading-then-Reasoning (IRR) framework. These advances demonstrate that verification through structural constraints can approach guaranteed correctness. Traceability is inherently strong through explicit graph paths. Correction remains moderate: KG updates affect future generations, but parametric knowledge conflicts require architectural intervention.

In-Generation Constraint Methods

While pre-generation grounding provides context, the LLM can still ignore or contradict it. This subsection surveys methods that constrain the decoding process *during* token generation, providing stronger guarantees.

Constrained Decoding Grammar-constrained decoding (Scholak et al. 2021) guarantees structural validity by masking constraint-violating tokens. Geng et al. (2023) extend grammar constraints to broader structural outputs. While this ensures structural validity, verification remains implicit-we know constraints were satisfied but not why particular choices were made. The KG-Trie pattern from Luo et al. (2024) represents the most powerful form: by encoding valid knowledge graph paths as a trie structure, the decoder can only generate tokens that continue valid paths. This provides a structural guarantee-the LLM cannot hallucinate facts that don't exist in the KG because such tokens are masked during decoding.

Logic-Integrated Reasoning Chain-of-Thought prompting (Wei et al. 2022) encourages exposed reasoning steps, but the faithfulness gap is critical: extensive research

demonstrates traces may be unfaithful rationalizations that do not reflect actual computation (Turpin et al. 2023). Logic-LM (Pan et al. 2023) addresses this by delegating verification to symbolic solvers with formal guarantees, achieving 39.2% improvement on logical reasoning benchmarks, symbolic solvers provide independently verifiable proofs.

The recent literature establishes formal foundations. Allen et al. (2025) demonstrate that formal verification is not incompatible with neural uncertainty: Belnap computers use four-valued paraconsistent logic, preserving soundness and completeness while handling LLM-derived inconsistency, we can reason soundly even with contradictory LLM-derived knowledge. Farjami et al. (2026) systematically compare first-order logic, modal logic (KD), and conditional logics, achieving 77.67% explanation success with modal KD at lower computational cost than first-order logic (58.83s vs 103.74s). The CoT Evaluation Survey Lee Hockenmaier (2025) provides comprehensive taxonomy evaluating reasoning across factuality, validity, coherence, and utility.

Post-Generation Verification Methods

Rather than modifying LLM architecture or constraining decoding, external neuro-symbolic systems can audit LLM outputs after generation, the LLM generates, the symbolic system verifies. This approach preserves neural flexibility while adding formal verification.

Fact-Checking with Knowledge Graphs FacTool (Chern et al. 2023) uses appropriate tools (search engines, code execution, calculators) to verify different claim types. The recent literature advances both accuracy and efficiency. Programmatic verification approaches achieve strong results. Pan et al. (2023) achieve high accuracy through explicit reasoning programs using structured verification functions.

Guardrails and Runtime Policy Enforcement NeMo Guardrails (Rebedea et al. 2023) introduced Colang, a DSL for declarative policy specification. The recent literature documents a paradigm shift in guardrails evolving through three stages. Frist, Reactive (post-hoc filtering), NeMo Guardrails Rebedea et al. (2023) operates as an external wrapper, applying declarative policies to filter inputs before generation and verify outputs after generation. Second, Runtime (guaranteed compliance), AgentSpec (Wang et al. 2025) ensures safety through runtime enforcement, intercepting and validating agent actions against formal specifications before execution. Third, Proactive (predictive intervention), Doshi et al. Doshi et al. (2026) introduce proactive guardrails that anticipate hazards via System-Theoretic Process Analysis (STPA) to enforce verifiable safety constraints on agent tool use.

This evolution represents moving from detect and block to prevent by construction to predict and preempt. Wang et al. (2025) provide the first comprehensive runtime enforcement framework, introducing a DSL for customizable safety constraints with triggers, predicates, and multiple enforcement mechanisms. Evaluation demonstrates remarkable effectiveness: over 90% prevention of unsafe execution for code agents, complete elimination of hazardous actions for embodied agents, and 100%

compliance for autonomous vehicle planning, all with millisecond overhead. RoboGuard Ravichandran et al. (2025) specialize formal enforcement for LLM-enabled robots through two-stage guardrails: safety rule contextualization followed by temporal logic control synthesis. Unsafe plan execution drops from 92% to 2.5% under jailbreak attacks while maintaining legitimate request performance. A critical architectural vulnerability complicates the landscape. WildGuard Wang et al. (2024a) addresses adversarial attacks through instruction tuning on the WildGuardMix dataset, achieving GPT-4 level performance in detecting sophisticated jailbreaks and refusals. Table 3 summarizes the landscape of neuro-symbolic auditing for LLMs, mapping specific intervention patterns to their key performance outcomes.

Table 3. Neuro-Symbolic LLM Auditing Methods: A taxonomy of intervention patterns across the Pre-Generation, In-Generation, and Post-Generation phases, highlighting key performance outcomes.

Phase	Pattern	Key Result	Citation
Pre-Generation	RAG	Identified structure-aware retrieval challenges.	Peng et al. (2024)
	KG-Enhanced Reasoning	Shift to zero-error guarantees.	Pan et al. (2024)
		KG-Trie decoding eliminates reasoning hallucination (0% vs. 33% baseline).	Luo et al. (2024)
		Outperforms GPT-4 by 23.9% using GPT-3.5.	Tan et al. (2025)
		Outperforms CoT baselines using Iterative Reading-then-Reasoning.	Jiang et al. (2023)
In-Generation	Constrained Decoding	SOTA text-to-SQL solutions via constrained decoding.	Scholak et al. (2021)
		Grammar-constrained decoding excels at structured tasks.	Geng et al. (2023)
	Logic-Integrated Reasoning	Outperforms standard prompting by 39.2% integrating LLMs with symbolic solvers.	Pan et al. (2023)

Continued on next page...

Table 3 (Continued)

Phase	Pattern	Key Result	Citation
		Achieves sound and complete neurosymbolic reasoning via LLM-grounded interpretations.	Allen et al. (2025)
		Improved verifiable NLI performance and proof efficiency via logic-internal strategies.	Farjami et al. (2026)
Post-Generation	Fact-Checking with KG	Task-agnostic detection of LLM factual errors.	Chern et al. (2023)
		Data-efficient, explanatory verification using programs.	Pan et al. (2023)
	Guardrails and Runtime Policy Enforcement	LLM-independent guardrails for conversational control	Rebedea et al. (2023)
		Over 90% prevention of unsafe execution	Wang et al. (2025)
		Formal safety guarantees via STPA	Doshi et al. (2026)
		Unsafe execution reduced from 92% to 2.5%	Ravichandran et al. (2025)

Auditing Multi-Agent Systems

Tool-Using Agents. ReAct Yao et al. (2023) established the dominant paradigm with Thought-Action-Observation loops creating natural audit points-explicit decision traces that enable verification. Gorilla Patil et al. (2023) addresses API hallucination through retrieval-aware training.

Multi-Agent Coordination. Multi-agent systems create emergent behaviors exceeding individual agent properties, compounding auditing challenges. AutoGen Wu et al. (2023) enables multi-agent conversation with complete interaction logging, MetaGPT Hong et al. (2023) applies Standard Operating Procedures. The compositional verification problem is fundamentally unsolved: properties holding for individual agents may not hold for compositions. Current traceability is limited to logs; causal analysis attributing outcomes across agents remains unsolved. AgentHarm Andriushchenko et al. (2024) demonstrates this concretely: safety training on direct requests fails catastrophically when harmful goals are pursued through multi-step tool-calling

sequences. Wang et al. Wang et al. (2024b) provide a unified framework for autonomous agents comprising profile, memory, planning, and action modules, along with a review of applications in social and natural sciences.

Applications in High-Stakes Domains

The deployment of AI systems in high-stakes domains, healthcare, finance, autonomous systems, and public administration, raises distinct auditing challenges that generic approaches cannot address. A healthcare diagnosis must be justified in clinically meaningful terms; a credit decision must satisfy regulatory explanation requirements; an autonomous action must provide safety guarantees; a government decision must enable democratic contestation. This section examines how neuro-symbolic approaches have been adapted to meet these domain-specific requirements, analyzing the resulting VTC capability profiles and identifying patterns that emerge across domains. The analysis proceeds through four domains before synthesizing cross-cutting themes. For each domain, we examine the primary auditing tension, survey the neuro-symbolic solutions that have emerged, and assess the resulting capabilities and limitations. This comparative structure reveals both domain-specific adaptations and transferable insights.

Healthcare: Balancing Accuracy and Clinical Justification

Healthcare AI faces a distinctive challenge: diagnostic accuracy alone is insufficient for clinical deployment. A neural model achieving high accuracy on chest X-rays may remain unusable if radiologists cannot understand why it flagged a particular image. The auditing requirement extends beyond correctness to encompass justification in medically meaningful terms, a requirement that has driven healthcare toward knowledge graph integration as the dominant neuro-symbolic pattern.

Knowledge Graph Integration for Traceable Reasoning Comprehensive surveys of healthcare knowledge graphs (Li et al. 2023; Al Khatib et al. 2024) establish that medical KGs built on UMLS, SNOMED-CT, and ICD ontologies provide the structured backbone for explainable diagnosis. The healthcare literature demonstrates convergence toward these knowledge graphs as verification anchors, constraining neural outputs to paths within established medical ontologies.

Prenosil et al. (2025) present a compelling demonstration of this approach, integrating GPT-4 with the rule-based expert system Plato-3 for clinical data extraction from PET/CT reports. The hybrid system achieved $F_1 = 1.00$ for study inclusion, compared to $F_1 = 0.63$ for GPT-4 alone, representing not incremental improvement but elimination of extraction errors through symbolic verification. The system intercepted intentionally incorrect anonymizations, demonstrating that symbolic rules can identify errors that neural components miss. This pattern extends across multiple systems with varying architectural choices. DR.KNOWS (Gao et al. 2025) grounds diagnosis prediction in UMLS subgraphs using stack graph isomorphism networks and attention-based path ranking. medIKAL (Jia et al. 2024) introduces entity-weighted KG importance with path reranking for clinical diagnosis on electronic medical records. MedKGI (Wang et al.

2025a) adds information-theoretic question selection for iterative differential diagnosis, achieving 30% dialogue efficiency improvement. The consistent contribution across these systems is the transformation of opaque neural predictions into traceable reasoning paths.

For rare disease diagnosis, where data scarcity poses additional challenges, ontology-grounded approaches have proven particularly valuable. DeepRare (Zhao et al. 2025) employs LLM agents with HPO (Human Phenotype Ontology) integration, enabling traceable reasoning for conditions where training data is limited. A comprehensive survey of neuro-symbolic AI in healthcare (Hossain and Chen 2025) identifies 41 biomedical use cases and documents the LTN-CPI (Logic Tensor Networks for Compound-Protein Interaction) framework achieving state-of-the-art results across multiple benchmarks.

Empirical Validation of Clinician Trust A striking finding replicates across multiple healthcare studies: clinician trust improves from approximately 67% to 94% when neuro-symbolic explanations replace pure neural explanations (Aamir et al. 2025). This improvement appears attributable to two mechanisms. First, knowledge graph paths employ medical terminology (UMLS concepts) that clinicians recognize, unlike attention weights or SHAP values. Second, clinicians can evaluate whether the reasoning path is medically plausible-a verification possibility absent in neural explanations.

Learned Rules Matching Clinical Standards Logical Neural Networks (LNN) for diabetes prediction (Lu et al. 2024) achieve 80.52% accuracy with AUROC 0.8457, outperforming Random Forest baselines (76.95% accuracy, AUROC 0.8342). The learned weights and thresholds within the LNN models provide direct insights into feature contributions, with decision-making processes that align with clinical reasoning, suggesting that neuro-symbolic approaches can bridge the gap between predictive accuracy and clinical interpretability.

Concept Bottleneck Models in Medical Imaging The concept bottleneck model (CBM) paradigm has become a leading approach for interpretable medical imaging, as documented in comprehensive surveys (Zhao et al. 2024). Several architectural innovations have advanced the field. Yan et al. (2023) demonstrate that GPT-4-generated clinical concepts combined with BioViL achieve 19% robustness improvement on medical imaging tasks. Kim et al. (2023) introduce visual concept filtering using activation scores, addressing the faithfulness problem where LLM-generated concepts may not be visually detectable, achieving +16.7% accuracy improvement. The Concept Complement Bottleneck Model (Wang et al. 2024) discovers new concepts beyond predefined sets, contributing 2-5% accuracy gains. MVP-CBM (Wang et al. 2025a) synthesizes these advances, achieving 97.84% BMAC on NCT dataset and 87.83% on ISIC2018 while outperforming black-box ViT-Base. The key innovation, recognizing that different concepts are associated with different network depths through Intra-layer Concept Preference Modeling (ICPM), provides design guidance for inherently interpretable medical imaging systems.

Finance: Navigating the Compliance Gap

The transition from healthcare to finance reveals a shift in the primary auditing driver. Where healthcare justification serves clinical collaboration, financial explanation serves legal obligation with substantial penalties. This regulatory pressure has not, however, produced correspondingly strong verification capabilities-creating what emerges as the most significant gap-to-pressure ratio across domains.

The Dominance of Post-Hoc Explanation Two comprehensive systematic reviews characterize the current state of financial XAI. Golec and AlabdulJalil (2025) analyze 60 papers on LLM-based credit risk using PRISMA methodology, identifying FinBERT and GPT-4 as dominant architectures. Khan et al. (2025) survey 150 papers on model-agnostic XAI in finance, finding SHAP as the most frequently used method. Both reviews reach the same conclusion: causal and counterfactual reasoning is almost entirely absent from the literature. The 5D Framework for evaluating XAI in credit risk (Ye and Chen 2025) provides systematic criteria-Inherent Interpretability, Global Explanations, Local Explanations, Consistency, and Complexity-against which current systems can be assessed. XAI-AutoML combinations (Schmitt 2024) have explored human-AI collaboration in credit decisions, while NLP-based approaches (Tan and Kok 2024) address financial document risk classification. Current systems rely on post-hoc SHAP providing correlation-based explanations (Feature X had high importance), while regulatory requirements implicitly demand causal explanations (You were denied because your debt-to-income ratio exceeded threshold Y). This gap between what systems provide and what regulations require creates compliance risk.

Regulatory Pressure and Technical Capability The regulatory landscape has intensified this gap. Analysis of the SCHUFA CJEU judgment by Muller (2025) clarifies that GDPR Article 22 scope has expanded-automated credit scores constitute decisions when they draw strongly on automated processing. Simultaneously, the EU AI Act classifies credit scoring as high-risk (Annex III point 5(b)), with penalties reaching EUR 20 million or 4% global turnover.

Anti-Money Laundering as Exception Anti-money laundering (AML) represents a notable exception to the broader finance pattern. Deprez et al. (2025) demonstrate that architecture-based continual learning methods maintain detection capability while incorporating new fraud patterns without catastrophic forgetting. Experiments on IBM AML dataset (5M transactions, 0.1% fraud) and Elliptic dataset (203K nodes) validate the approach. Khanvilkar and Kommuru (2025) combine GNNs with regulatory RAG, achieving 98.2% F_1, 97.8% precision, and 97.0% recall with explanations mapping directly to Bank Secrecy Act provisions. Two factors distinguish AML from credit scoring: transaction networks naturally map to GNNs with explicit relationship encoding, and the evolving threat model explicitly requires adaptation to new patterns.

Legal Knowledge Graphs Legal RAG systems have developed sophisticated approaches to citation provenance. Barron et al. (2025) present a system covering 265 constitutional provisions, 28,251 statutes, and 15,799 cases for New Mexico jurisdiction,

combining Vector Store, Knowledge Graph, and Hierarchical NMF topic discovery. Song et al. (2025) demonstrate KG-assisted LLM post-training for enhanced legal reasoning, while ontology-driven approaches (de Martim 2025) address hierarchical and temporal legal norm relationships with deterministic provenance.

Autonomous Systems: Achieving Formal Guarantees

The transition to autonomous systems reveals the most demanding auditing requirements and, correspondingly, the most sophisticated neuro-symbolic solutions. Where healthcare accepts structured justification and finance largely accepts correlation-based explanation, autonomous systems require mathematical guarantees. Statistical confidence (99% safe) is insufficient when one failure in 100 can cause physical harm. Comprehensive surveys of neuro-symbolic reinforcement learning and planning (Acharya et al. 2023) establish the Learning for Reasoning versus Reasoning for Learning taxonomy that characterizes the field.

Formal Methods for Safety Verification SELP (Wu et al. 2025b) combines equivalence voting for natural language to LTL translation with Büchi automata that prune unsafe tokens during generation. Unsafe plans are never generated, a fundamental departure from post-hoc filtering. The system achieves 95.2% safety on drone navigation, 93.6% on tabletop manipulation, 98% LTL translation accuracy, and 34.85% execution time reduction. Aegis (Shi et al. 2024) uses counterexample-guided inductive synthesis (CEGIS) with Bayesian optimization to synthesize programmatic shields, simple linear policies with inequality checks that achieve 100% safety with 2.2× time reduction and 1.5× fewer unnecessary interventions. Hafez et al. (2025) compute over-approximated reachable sets from historical trajectory data without explicit analytical models, achieving zero collisions with formal guarantees on TurtleBot3 and JetRacer platforms (25 Hz computation, 0.1m minimum obstacle distance versus 0.5m baseline). Verification-guided shielding (Corsi et al. 2024) partitions state space into formally verified safe and unsafe regions, while POLAR-Express (Yang et al. 2024) enables runtime verification with controller switching for neural network controlled systems.

From Statistical Confidence to Mathematical Proof These systems represent a paradigm shift from probably safe to provably safe. LogicGuard (Gokhale et al. 2025b) demonstrates correction through this paradigm: a symbolic actor-critic architecture where an LLM critic supervises an LLM actor using LTL constraints. When failures occur, the system synthesizes new temporal logic constraints to prevent recurrence, achieving +25% task completion on Behavior-100 (47%→72%) and reducing failed actions from 23% to 4.5%. Neuro-symbolic approaches to safe autonomous driving (Sharifi et al. 2023) combine deep reinforcement learning with first-order logic for real-time safety verification. LLM-as-PDDL-planner approaches (Silver et al. 2023) train GPT-3 for PDDL-compatible task planning, enabling symbolic verification of generated plans.

Scalability of Formal Methods A key barrier to formal methods has been computational cost. SEVIN (Parameshwaran & Wang 2025) addresses this through VAE-based latent

space encoding, achieving 600× dimensionality reduction and verification in under one second per specification, which is 10× faster than general robustness verification. The approach validates NvidiaNet and ResNet18 controllers.

Hierarchical Architectures NeuroStrata (Zheng et al. 2025) proposes three-level CPS design creating natural auditing boundaries: formally verifiable symbolic specifications at high level, constraint-satisfying neuro-symbolic modules at middle level, and runtime-monitored neural controllers at low level. Imperative Learning (Wang et al. 2024b) combines neural networks with reasoning engines and episodic memory for robot autonomy tasks.

Adversarial Robustness RoboSafe (Wang et al. 2025) demonstrates robustness to jailbreak attacks using executable predicate-based safety logic with backward reflective reasoning (temporal predicates) and forward predictive reasoning (contextual predicates). Results: -36.8% risk occurrence on SafeAgentBench, 92.33% accurate refusal rate, 89% execution success rate on safe tasks. Under adversarial prompting, only 5.22% execution success rate compared to 92% baseline.

Public Sector: The Accountability Paradigm

The transition from autonomous systems to public sector reveals a fundamentally different conception of auditing. Where technical domains prioritize verification, public sector applications prioritize contestability, the ability of affected individuals to challenge decisions. This distinction reflects the constitutional dimension of government algorithms.

Contestability Over Verification Schmude et al. (2025) interview 14 AI regulation experts, distinguishing descriptive explainability (how the system works) from normative explainability (why the decision is appropriate against legal and ethical standards). The key insight is that citizens may not need to verify algorithmic correctness; they need to contest decisions affecting their lives. Neuro-symbolic approaches enable substantive contestability: when decisions derive from symbolic rules, affected individuals can challenge specific rule interpretations. The study identifies three implementation frictions: spirit of law preservation, responsibility assignment across regulators, and interdisciplinary collaboration gaps.

Individual and Collective Contestation Current frameworks emphasize individual rights (explain my decision). However, algorithmic harms are often systemic, affecting demographic groups rather than individuals. Schmude et al. (2025) identify collective contestation mechanisms as underdeveloped-a gap with implications for addressing patterns of algorithmic discrimination.

Cross-Domain Synthesis

The preceding domain analyses reveal patterns that transcend individual applications. Table 4 summarizes the VTC profiles that emerge from each domain's adaptation of neuro-symbolic approaches.

Table 4. Neuro-Symbolic Integration Patterns Across High-Stakes Domains: State-of-the-Art

Domain	Pattern	Mechanism	Citations
Healthcare	LLM + Rule-Based Expert System	Neural LLM extracts information; symbolic expert system (e.g., Prolog) validates outputs against medical rules	Prenosil et al. (2025)
	KG-Grounded Diagnosis	LLM predictions constrained to paths within medical ontologies (UMLS, SNOMED-CT)	Gao et al. (2025); Jiang et al. (2025); Zuo et al. (2024); Jiang et al. (2024)
	Medical KG + LLM	Entity-weighted path reranking; info-theoretic question selection; HPO for rare diseases	Jia et al. (2024); Wang et al. (2025a); Zhao et al. (2025)
	Logical Neural Net	First-order logic with learnable thresholds matching clinical standards (HbA1c 6.5%)	Lu et al. (2024)
	Concept Bottleneck	Multi-level image-concept alignment; energy-based CBM; incremental concept discovery	Wang et al. (2025a); Bie et al. (2024); Xu et al. (2024); Shang et al. (2024)
	GNN Clinical Risk	Graph attention networks on EHR; collaborative filtering	Aamir et al. (2025)
Finance & Credit Risk	Post-hoc Feature Attribution	SHAP/LIME applied after model decision to explain feature importance	Khan et al. (2025); Mohsin and Nasim (2025)
	Credit Risk LLM	FinBERT/GPT-4 with explanations; counterfactual with causal discovery	Golec and AlabdulJalil (2025); Takahashi et al. (2024); Shreya and Pathak (2025)
	GNN Anti-Money Laundering	Continual graph learning; regulatory RAG with BSA citations	Khanvilkar and Kommuru (2025); Deprez et al. (2025); Cheng et al. (2024)
	Fraud Detection	CNN+GNN+LSTM ensemble with attention; 100+ GNN studies	Chagahi et al. (2024); Cheng et al. (2024)

Continued on next page...

Table 4 – Continued from previous page

Domain	Pattern	Mechanism	Citations
	Regulatory Compliance	GDPR Art.22 + AI Act Annex III; SCHUFA judgment	Muller (2025)
Legal & Compliance	Legal KG + RAG	Vector store + KG + hierarchical NMF; ontology-driven temporal norms	Barron et al. (2025); Song et al. (2025); de Martim (2025)
	Legal RAG Benchmark	6,858 QA pairs; hallucination assessment (17-33%); CBR+LLM; adaptive parameter tuning	Pipitone and Houir Alami (2024); Magesh et al. (2024); Wiratunga et al. (2024); Kalra et al. (2024)
Autonomous & CPS	LTL-Constrained Decoding	Büchi automata pruning; LTL safety module; NL-to-TL translation; Z3 SMT verification	Wu et al. (2025b); Yang et al. (2024); Chen et al. (2023); Hao et al. (2025)
	Runtime Shields	CEGIS + Bayesian optimization; Deep Kernel Learning; parametric safety specs; conformal prediction	Shi et al. (2024); Reed and Lahijanian (2024); Corsi et al. (2025); Scarbro et al. (2025)
	Reachability Analysis	Taylor model propagation; constraint-aware refinement; VAE latent encoding for vision controllers	Hafez et al. (2025); Parameshwaran & Wang (2025); Wang et al. (2024); Rober and How (2024)
	Safe RL / Neural CBF	Graph neural control barrier functions; safety attention mechanism; DRL+FOL zero unsafe actions	Zhang et al. (2024); Sharifi et al. (2023)
	LTL Actor-Critic	LLM critic supervises actor with temporal logic constraints; constraint synthesis on failure	Gokhale et al. (2025b); Wang et al. (2024b)
	Hierarchical NS-CPS	Three-level architecture with formal verification boundaries; STL specifications	Zheng et al. (2025)

Continued on next page...

Table 4 – Continued from previous page

Domain	Pattern	Mechanism	Citations
	Adversarial Robustness	Executable predicate-based safety logic; backward/forward reasoning; 5.22% adversarial success	Wang et al. (2025)
Public Sector	Contestability	Normative vs descriptive explainability; individual and collective action channels	Schmude et al. (2025)

Emergent Patterns Three cross-domain patterns emerge from this analysis:

- **Pattern 1: Domains achieve VTC capabilities aligned with their primary auditing need.** Healthcare prioritizes traceability through KG paths (Gao et al. 2025; Jia et al. 2024; Wang et al. 2025a); finance prioritizes traceability through feature attribution (Khan et al. 2025; Golec and AlabdulJalil 2025); autonomous systems prioritize verification through formal methods (Wu et al. 2025b; Shi et al. 2024; Hafez et al. 2025); public sector prioritizes correction through contestability (Schmude et al. 2025). This alignment reflects rational allocation of research effort. However, only autonomous systems achieve high capability across all three dimensions.
- **Pattern 2: A shift from post-hoc to generation-time validation is underway, but unevenly distributed.** Healthcare demonstrates this shift through hybrid expert systems (Prenosil et al. 2025) achieving F_1 improvement from 0.63 to 1.00. Autonomous systems demonstrate it through LTL-constrained decoding (Wu et al. 2025b) that prevents unsafe generation. Finance remains largely at the post-hoc SHAP stage (Khan et al. 2025) as a potential vulnerability given regulatory pressure from GDPR and AI Act (Muller 2025).
- **Pattern 3: Formal methods demonstrate feasibility but require domain-specific formalization.** Autonomous systems prove that High VTC is achievable: SELP achieves 95.2% safety (Wu et al. 2025b); Aegis achieves 100% safety (Shi et al. 2024); reachability analysis achieves zero collisions (Hafez et al. 2025). SEVIN demonstrates scalability with sub-second verification (Parameshwaran & Wang 2025). The question for other domains is whether auditing requirements can be formalized sufficiently to apply these techniques.

Research Opportunities The gaps identified suggest several research directions:

- **Causal reasoning for finance** represents the largest gap-to-pressure ratio. The absence of causal methods documented in systematic reviews of 60 papers (Golec and AlabdulJalil 2025) and 150 papers (Khan et al. 2025), despite strong regulatory pressure (Muller 2025), suggests significant opportunity for counterfactual credit scoring. Real-time medical verification could transfer autonomous systems techniques. SEVIN achieves sub-second verification through latent space encoding

(Parameshwaran & Wang 2025); similar approaches might enable formally verified diagnostic support, addressing healthcare's current gap in verification capability (Prenosil et al. 2025).

- **Cross-domain knowledge graph integration** addresses the reality that healthcare decisions involve legal, financial, and social dimensions. Current systems operate within single domains (Gao et al. 2025; Barron et al. 2025); federated approaches remain unexplored.
- **Collective contestation mechanisms** address the gap between individual explanation and systemic accountability (Schmude et al. 2025). Patterns of algorithmic discrimination affecting demographic groups require mechanisms beyond individual-level XAI.
- **Continual learning beyond AML** could extend PackNet-style approaches (Deprez et al. 2025) to evolving clinical guidelines, legal precedents, and regulatory requirements across domains.

High-Stake Application Summary

This section's examination of high-stakes domain applications reveals that neuro-symbolic approaches have been adapted to meet diverse auditing requirements, with varying degrees of success. Healthcare has achieved high traceability through knowledge graph integration (Gao et al. 2025; Jia et al. 2024; Prenosil et al. 2025); finance faces a significant gap between regulatory pressure (Muller 2025) and technical capability (Golec and AlabdulJalil 2025; Khan et al. 2025); autonomous systems demonstrate that formal verification is achievable at scale (Wu et al. 2025b; Shi et al. 2024; Parameshwaran & Wang 2025).

The cross-domain comparison yields three findings. First, domains rationally optimize for their primary auditing need, but this specialization creates gaps that cross-domain learning might address. Second, the autonomous systems literature provides existence proofs for formal verification that healthcare and finance have not yet exploited. Third, finance faces the largest discrepancy between regulatory requirements and technical capabilities, representing both significant risk and significant opportunity.

Open Challenges and Future Directions

While Neuro-Symbolic AI provides the architectural handles for verification, traceability, and correction, significant barriers remain to deploying these systems at the scale required for high-stakes infrastructure. This section analyzes the open challenges that impede the VTC framework and outlines a research roadmap for trustworthy automation.

The Scalability-Verification Paradox

The central promise of NSAI, the formal verification of neural behavior, confronts a fundamental obstacle in the computational complexity of symbolic reasoning. A persistent barrier exists in the joint training bottleneck, where neural networks require accurate symbolic rules for supervision while symbolic learners simultaneously require

accurate neural predictions to identify patterns. This chicken-and-egg dilemma forces a trade-off between expressiveness and efficiency. For instance, frameworks like DeepProbLog offer high expressiveness for auditing but face exponential growth in complexity as systems scale. This creates a distinct verification gap in current auditing capabilities: while tools exist to formally verify small, safety-critical controllers, they lack the unified framework necessary to apply formal verification to billion-parameter models without aggressive abstraction. Future research must prioritize solver-free approaches that can approximate constraint satisfaction efficiently without sacrificing the rigorous mathematical guarantees required for safety certification.

The Dynamic Knowledge Bottleneck and Transparency Gap

The correction pillar of auditing relies on the assumption that model errors can be fixed by editing a Knowledge Graph (KG), yet this assumes the KG itself remains a static, accurate reflection of ground truth. In dynamic, high-stakes environments, traditional KGs risk symbolic concept drift, in which an ontology becomes outdated, leading to stale explanations that amount to effective hallucinations. To mitigate this correction lag, future auditing tools must transition from static knowledge verification to dynamic infrastructure auditing that automatically flags outdated regulations within the inference chain. Furthermore, A critical disconnect persists between research outputs and audit requirements, as current benchmarks primarily measure outcomes, such as accuracy, rather than the underlying reasoning process. This leads to the ”Faithfulness Problem,” in which systems may take shortcuts to achieve correct answers via spurious correlations rather than the intended symbolic path. An auditor cannot trust a traceability log if the model’s actual computation deviates from the provided logic, resulting in a significant traceability gap. Because formal reasoning traces currently appear in only a small fraction of the literature, there is an urgent need for standardized metrics for Process Fidelity. Future benchmarks must evaluate not only whether an explanation is faithful to the system’s behavior but also whether the system maintains logical stability across similar inputs to prevent procedural audit through inconsistent rationalizations.

The Generative Future of Neuro-Symbolic Auditing

The next generation of high-stakes auditing will likely invert the current paradigm by utilizing LLMs as the operational engine for oversight rather than viewing them solely as sources of risk. This generative future is defined by three critical roles that directly address the bottlenecks of the VTC framework: automating compliance, scaling verification, and democratizing traceability. First, LLMs will function as regulatory-to-symbolic compilers, ingesting massive volumes of natural language statutes, such as the EU AI Act or complex clinical guidelines, and automatically translating them into machine-verifiable axioms. This transformation automates the correction layer, enabling high-stakes systems to update their safety boundaries in real time as laws evolve. Second, to overcome the limitations of static benchmarks, generative adversarial auditing will deploy LLMs to proactively hunt for failure modes by generating thousands of diverse, high-stakes synthetic scenarios that the neuro-symbolic core must then formally verify.

Finally, LLMs will serve as the narrative interface for the glass box, ingesting rigorous symbolic proof trees and synthesizing them into human-readable explanations. By decoupling the reasoning process (handled by the symbolic solver) from the explanation generation (handled by the LLM), this approach ensures that decision traces are both factually grounded in verifiable logic and intelligible to non-technical stakeholders, such as judges and medical professionals.

Discussion and Conclusion

The trajectory of Artificial Intelligence from 2020 to 2026 has brought the field to a critical inflection point. While the generative revolution has delivered unprecedented capabilities in language, vision, and planning, it has simultaneously exposed a dangerous audit gap in high-stakes infrastructure. As this survey has detailed, the widespread deployment of probabilistic, black-box architectures in domains such as healthcare, finance, and autonomous systems has outpaced our ability to govern them. Current oversight mechanisms, often limited to procedural checklists and post hoc statistical evaluations, have led to a culture of audit washing, where compliance is performative rather than mathematical.

This research argues that the solution to this reliability crisis is not merely regulatory, but architectural. We have defined the VTC framework as the technical standard for responsible automation. High-stakes systems must be able to formally prove adherence to safety constraints (Verification), reconstruct faithful decision trails for accountability (Traceability), and accept surgical repairs without catastrophic forgetting (Correction). Our analysis demonstrates that purely neural architectures are structurally ill-equipped to meet these demands.

NSAI emerges not simply as a means to improve performance, but as the essential infrastructure for auditability. By synthesizing the perceptual adaptability of neural networks with the rigorous structure of symbolic logic, NSAI provides the necessary handles for governance. We have demonstrated how Neural-Symbolic Pipelines and KG-Grounded models transform the opaque Black Box into a transparent Glass Box, ensuring that high-stakes decisions in healthcare and finance are traceable to specific, validatable concepts. Beyond transparency, we highlighted how Symbolic-Neural systems enable formal verification, allowing autonomous systems to operate within guaranteed safety envelopes that purely stochastic models cannot provide. Ultimately, while we acknowledge trade-offs between scalability and logical completeness, we identify integrating symbolic constraints as a viable path to eliminate hallucinations and enforce regulatory alignment by construction.

Looking forward, the landscape of auditing is evolving from a passive evaluation task to an active, generative process. As outlined in our research roadmap, the next generation of auditing will likely leverage LLMs as regulatory-to-symbolic compilers, bridging the chasm between natural-language laws and machine-executable code. This points toward a future where auditing is continuous, automated, and deeply integrated into the model's inference cycle. Ultimately, trust in high-stakes automation cannot be claimed but must be proven. The synthesis of neural learning and symbolic reasoning

provides the robust pathway to convert theoretical principles of fairness, transparency, and safety into technically verifiable realities (VTC). For AI to safely govern critical infrastructure, it must first be inherently governable.

Acknowledgements

The authors would like to thank the reviewers for their constructive feedback. This work was supported by the National Science Foundation.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This material is based upon work supported by the U.S. National Science Foundation (NSF) under Award Abstract #2333836, Proto-OKN Theme 1: Creating a Cross-Domain Knowledge Graph to Integrate Health and Justice for Rural Resilience. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Raji ID, Smart A, White RN, et al. (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44.

Landers RN and Behrend TS (2023) Auditing the AI auditors: A framework for evaluating fairness and bias in high-stakes AI predictive models. *American Psychologist* 78(1): 36–49.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5): 206–215.

Marcus G (2020) The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Zurawski J and Schopf J (2023) National Institute of Standards and Technology Requirements (Analysis Report).

Garcez AD and Lamb LC (2023) Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review* 56(11): 12387–12406.

Kautz H (2022) The third AI summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine* 43(1): 93–104.

Yao Y, Wang P, Tian B, et al. (2023) Editing large language models: Problems, methods, and opportunities. In: *Proceedings of EMNLP 2023*, pp. 10222–10240.

Gunning D, Stefik M, Choi J, et al. (2019) XAI-Explainable artificial intelligence. *Science Robotics* 4(37): eaay7120.

Page MJ, McKenzie JE, Bossuyt PM, et al. (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*

Citron DK and Pasquale F (2014) The scored society: Due process for automated predictions. *Washington Law Review* 89: 1–33.

Gebru T, Morgenstern J, Vecchione B, et al. (2021) Datasheets for datasets. *Communications of the ACM* 64(12): 86–92.

Mitchell M, Wu S, Zaldivar A, et al. (2019) Model cards for model reporting. In: *Proceedings of FAT* 2019*, pp. 220–229.

Hardt M, Price E, and Srebro N (2016) Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems*, vol. 29.

Kurakin A, Goodfellow I, and Bengio S (2017) Adversarial examples in the physical world. In: *ICLR Workshop*.

Lan Z, Chen Y, and Wang X (2024) Runtime monitoring for neural network systems. *ACM Computing Surveys* 56(3): 1–35.

Interlandi M, Shah K, Tetali SD, et al. (2015) Titian: Data provenance support in Spark. *Proceedings of the VLDB Endowment* 9(3): 216–227.

Katz G, Barrett C, Dill DL, et al. (2017) Reluplex: An efficient SMT solver for verifying deep neural networks. In: *CAV 2017*, pp. 97–117.

Katz G, Huang DA, Ibeling D, et al. (2019) The Marabou framework for verification and analysis of deep neural networks. In: *CAV 2019*, pp. 443–452.

Wang S, Zhang H, Xu K, et al. (2021) Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In: *NeurIPS 2021*.

Ribeiro MT, Singh S, and Guestrin C (2016) Why should I trust you?: Explaining the predictions of any classifier. In: *KDD 2016*, pp. 1135–1144.

Lundberg SM and Lee SI (2017) A unified approach to interpreting model predictions. In: *NeurIPS 2017*, pp. 4765–4774.

Lipton ZC (2018) The mythos of model interpretability. *Queue* 16(3): 31–57.

Alshiekh M, Bloem R, Ehlers R, et al. (2018) Safe reinforcement learning via shielding. In: *AAAI 2018*, pp. 2669–2678.

Choi J, Kim S, and Park H (2023) CREST: Knowledge-guided clinical reasoning. In: *Proceedings of ACL 2023*.

Garcez AD and Lamb LC (2020) Neurosymbolic AI: The 3rd wave. *arXiv preprint arXiv:2012.05876*.

Koh PW, Nguyen T, Tang YS, et al. (2020) Concept bottleneck models. In: *ICML 2020*, pp. 5338–5348.

Mao J, Gan C, Kohli P, et al. (2019) The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In: *ICLR 2019*.

Yuksekgonul M, Wang M, and Zou J (2023) Post-hoc concept bottleneck models. In: *ICLR 2023*.

Serafini L and Garcez AD (2016) Logic tensor networks: Deep learning and logical reasoning from first principles to machines. *arXiv preprint arXiv:1606.04422*.

Badreddine S, Garcez AD, Serafini L, et al. (2022) Logic tensor networks. *Artificial Intelligence* 303: 103649.

Riegel R, Gray A, Luus F, et al. (2020) Logical neural networks. *arXiv preprint arXiv:2006.13155*.

Sen P, Riegel R, and Gray A (2022) Neuro-symbolic reasoning with logical neural networks. In: *AAAI 2022 Tutorial*.

Manhaeve R, Dumancic S, Kimmig A, et al. (2018) DeepProbLog: Neural probabilistic logic programming. In: *NeurIPS 2018*, pp. 3749–3759.

Manhaeve R, Dumancic S, Kimmig A, et al. (2021) Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence* 298: 103504.

Li Z, Huang J, and Naik M (2023) Scallop: A language for neurosymbolic programming. *Proceedings of the ACM on Programming Languages* 7(PLDI): 1–25.

Trinh TH, Wu Y, Le QV, et al. (2024) Solving olympiad geometry without human demonstrations. *Nature* 625(7995): 476–482.

Dong H, Mao J, Lin T, et al. (2019) Neural logic machines. In: *ICLR 2019*.

Zhang X, Bosselut A, Yasunaga M, et al. (2022) GreaseLM: Graph reasoning enhanced language models for question answering. In: *ICLR 2022*.

Liu X, Chen Y, and Wang Z (2024) Neural-symbolic integration for knowledge-grounded reasoning. *AI Open* 5: 1–15.

Wagner B and Garcez AD (2021) Neural-symbolic integration for fairness in AI. In: *NeSy 2021*.

Dong Y, Liu N, Jalaian B, et al. (2023) Fairness in graph neural networks: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Andriushchenko M, Souly A, et al. (2024) AgentHarm: A benchmark for measuring harmfulness of LLM agents. *arXiv preprint* arXiv:2410.09024.

Bai Y, Kadavath S, Kundu S, et al. (2022) Constitutional AI: Harmlessness from AI feedback. *arXiv preprint* arXiv:2212.08073.

Ravichandran Z, Robey A, Kumar V, et al. (2025) Safety guardrails for LLM-enabled robots. *arXiv preprint* arXiv:2503.07885.

Chern S, Chern E, Neubig G, et al. (2023) FacTool: Factuality detection in generative AI– a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint* arXiv:2307.13528.

Lee J, Hockenmaier J. (2025) Evaluating step-by-step reasoning traces: A survey. *arXiv preprint* arXiv:2502.12289.

Allen B P, Chhikara P, Ferguson T M, et al. (2025) Sound and complete neurosymbolic reasoning with LLM-grounded interpretations. *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, pp. 392–419.

Doshi A, Hong Y, Xu C, et al. (2026) Towards verifiably safe tool use for LLM agents. *arXiv preprint* arXiv:2601.08012.

Farquhar S, Kossen J, Kuhn L, et al. (2024) Detecting hallucinations in large language models using semantic entropy. *Nature* 630: 625–630.

Geng S, Josifoski M, Perez E, et al. (2023) Grammar-constrained decoding for structured NLP tasks without finetuning. In: *Proceedings of EMNLP 2023*. ACL.

Hong S, Zhuge M, Chen J, et al. (2023) MetaGPT: Meta programming for a multi-agent collaborative framework. *arXiv preprint* arXiv:2308.00352.

Hu Y, Lei Z, Zhang Z, et al. (2025) GRAG: Graph retrieval-augmented generation. *arXiv preprint* arXiv:2405.16506.

Huang L, Yu W, Ma W, et al. (2023) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint* arXiv:2311.05232.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, et al., Survey of hallucination in natural language generation, ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

Inan H, Upasani K, Chi J, et al. (2023) Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint* arXiv:2312.06674.

Jiang J, Zhou K, Dong Z, et al. (2023) StructGPT: A general framework for large language model to reason over structured data. In: *Proceedings of EMNLP 2023*. ACL.

Pan L, Wu X, Lu X, et al. (2023) Fact-checking complex claims with program-guided reasoning. *arXiv preprint* arXiv:2305.12744.

Lewis P, Perez E, Piktus A, et al. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proceedings of NeurIPS 2020*.

Lin S, Hilton J, Evans O (2022) TruthfulQA: Measuring how models mimic human falsehoods. In: *Proceedings of ACL 2022*. ACL.

S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3580–3599, 2024.

Luo L, Zhao Z, Haffari G, et al. (2024) Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. *Proceedings of the 42nd International Conference on Machine Learning*, pp. 41540–41565.

Manakul P, Liusie A, Gales MJF (2023) SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In: *Proceedings of EMNLP 2023*. ACL.

Min S, Krishna K, Lyu X, et al. (2023) FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: *Proceedings of EMNLP 2023*. ACL.

Mu T, Helyar J, Heidecke J, et al. (2024) Rule-based rewards for language model safety. *arXiv preprint* arXiv:2411.01111.

Pan L, Albalak A, Wang X, et al. (2023) Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In: *Findings of EMNLP 2023*. ACL.

Patil S, Zhang T, Wang X, et al. (2023) Gorilla: Large language model connected with massive APIs. *arXiv preprint* arXiv:2305.15334.

Peng B, Zhu Y, Liu Y, et al. (2024) Graph retrieval-augmented generation: A survey. *arXiv preprint* arXiv:2408.08921.

Rebedea T, Dinu R, Sreedhar M, et al. (2023) NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In: *EMNLP 2023 System Demonstrations*. ACL.

Scholak T, Schucher N, Bahdanau D (2021) PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In: *Proceedings of EMNLP 2021*. ACL.

Tan X, Wang X, Liu Q, et al. (2025) Paths-over-Graph: Knowledge graph empowered large language model reasoning. *Proceedings of the ACM on Web Conference 2025*, pp. 3505–3522.

Tonmoy SM, Zaman SM, Jain V, et al. (2024) A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint* arXiv:2401.01313.

Turpin M, Michael J, Perez E, et al. (2023) Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In: *Proceedings of NeurIPS 2023*.

Farjami A, Redondi L, Valentino M (2026) Logic-parametric neuro-symbolic NLI: Controlling logical formalisms for verifiable LLM reasoning. *arXiv preprint* arXiv:2601.05705.

Wang H, Chen Y, Liu Z, et al. (2024) WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint* arXiv:2406.18495.

Wang L, Ma C, Feng X, et al. (2024) A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18(6): 186345.

Wei J, Wang X, Schuurmans D, et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. In: *Proceedings of NeurIPS 2022*.

Wei A, Haghtalab N, Steinhardt J (2024) Jailbroken: How does LLM safety training fail? In: *Proceedings of NeurIPS 2023*.

Wu Q, Bansal G, Zhang J, et al. (2023) AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework. *arXiv preprint* arXiv:2308.08155.

Wang H, Poskitt C M, Sun J (2025) AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. *arXiv preprint* arXiv:2503.18666.

Yao S, Zhao J, Yu D, et al. (2023) ReAct: Synergizing reasoning and acting in language models. In: *Proceedings of ICLR 2023*.

Zeng W, Liu Y, Mullins R, Peran L, Fernandez J, Harkous H, Narasimhan K, Proud D, Kumar P, Radharapu B, et al. (2024) ShieldGemma: Generative AI content moderation based on Gemma. *arXiv preprint* arXiv:2407.21772.

Aamir M, Khan A, Ahmad R, et al. (2025) Enhancing chronic disease management: hybrid graph networks and explainable AI for intelligent diagnosis. *Scientific Reports* 15: 17593. DOI: 10.1038/s41598-025-34065-5.

Al Khatib AS, Neupane S, Manchukonda HK, Golilarz NA, Mittal S, Amirlatifi A, Rahimi S (2024) Patient-centric knowledge graphs: A survey of current methods, challenges, and applications. *Frontiers in Artificial Intelligence* 7: 1388479. arXiv:2402.12608.

Wang Q, Sheng R, Li Y, Qu H, Sun Y, Zhu M (2025) MedKGI: Iterative differential diagnosis with medical knowledge graphs and information-guided inquiring. *arXiv preprint arXiv:2512.24181*.

Zhao W, Wu C, Fan Y, Zhang X, Qiu P, Sun Y, Zhou X, Wang Y, Sun X, Zhang Y, Yu Y, Sun K, Xie W (2025) An agentic system for rare disease diagnosis with traceable reasoning. *arXiv preprint arXiv:2506.20430*.

Gao Y, Li R, Croxford E, Caskey J, Patterson BW, Churpek M, Miller T, Dligach D, Afshar M (2025) Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI* 4(1): e58670. DOI: 10.2196/58670. arXiv:2308.14321.

Jiang P, Xiao C, Cross A, Sun J (2024) GraphCare: Enhancing healthcare predictions with personalized knowledge graphs. In: *Proceedings of ICLR 2024*. arXiv:2305.12788.

Hossain D, Chen JY (2025) A study on neuro-symbolic artificial intelligence: Healthcare perspectives. *arXiv preprint arXiv:2503.18213*.

Jiang P, Xiao C, Jiang M, Bhatia P, Kass-Hout T, Sun J, Han J (2025) KARE: Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In: *Proceedings of ICLR 2025*. arXiv:2410.04585.

Zuo K, Jiang Y, Mo F, Lio P (2024) KG4Diagnosis: A hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis. *arXiv preprint arXiv:2412.16833*.

Cui H, Lu J, Xu R, Wang S, Ma W, Yu Y, Yu S, Kan X, Ling C, Zhao L, Qin ZS, Ho JC, Fu T, Ma J, Huai M, Wang F, Yang C (2023) A review on knowledge graphs for healthcare: Resources, applications, and promises. *arXiv preprint arXiv:2306.04802*.

Lu Q, Li R, Sagheb E, Wen A, Wang J, Wang L, Fan JW, Liu H (2024) Explainable diagnosis prediction through neuro-symbolic integration. In: *Proceedings of AMIA Informatics Summit 2025*. arXiv:2410.01855.

Prenosil GA, Weitzel TK, Bello SC, Mingels C, et al. (2025) Neuro-symbolic AI for auditable cognitive information extraction from medical reports. *Communications Medicine* 5: 309. DOI: 10.1038/s43856-025-01194-x.

Jia M, Duan J, Song Y, Wang J (2024) medIKAL: Integrating knowledge graphs as assistants of LLMs for enhanced clinical diagnosis on EMRs. In: *Proceedings of COLING 2024*. arXiv:2406.14326.

Xu X, Qin Y, Mi L, Wang H, Li X (2024) Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In: *Proceedings of ICLR 2024*. arXiv:2401.14142.

Kim I, Kim J, Choi J, Kim HJ (2023) Concept bottleneck with visual concept filtering for explainable medical image classification. In: *MedAGI Workshop, MICCAI 2023*. arXiv:2308.11920.

Bie Y, Luo L, Chen H (2024) MICA: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. In: *Proceedings of AAAI 2024*, 38: 837–845. arXiv:2401.08527.

Shang C, Zhou S, Zhang H, Ni X, Yang Y, Wang Y (2024) Incremental residual concept bottleneck models. In: *Proceedings of CVPR 2024*. arXiv:2404.08978.

Wang H, Hou J, Chen H (2024) Concept complement bottleneck model for interpretable medical image diagnosis. *arXiv preprint arXiv:2410.15446*.

Wang C, Zhang K, Liu Y, He Z, Tao X, Zhou SK (2025) MVP-CBM: Multi-layer visual preference-enhanced concept bottleneck model for explainable medical image classification. In: *Proceedings of IJCAI 2025*, 529–537. arXiv:2506.12568.

Yan A, Wang Y, Zhong Y, He Z, Karypis P, Wang Z, Dong C, Gentili A, Hsu CN, Shang J, McAuley JJ (2023) Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*.

Hou J, Liu S, Bie Y, Wang H, Tan A, Luo L, Chen H (2024) Self-explainable AI for medical image analysis: A survey and new outlooks. *arXiv preprint arXiv:2410.02331*.

Ye R, Chen J (2025) Unlocking the black box: A five-dimensional framework for evaluating explainable AI in credit risk. *arXiv preprint arXiv:2511.04980*.

Bank for International Settlements (2025) Managing explanations: How regulators can address AI explainability. *FSI Occasional Papers*.

Takahashi D, Shimizu S, Tanaka T (2024) Counterfactual explanations of black-box machine learning models using causal discovery with applications to credit rating. In: *Proceedings of IJCNN 2024*. arXiv:2402.02678.

Deprez B, Wei W, Verbeke W, Baesens B, Mets K, Verdonck T (2025) Advances in continual graph learning for anti-money laundering systems: A comprehensive review. *WIREs Computational Statistics*. arXiv:2503.24259.

Cheng D, Zou Y, Xiang S, Jiang C (2024) Graph neural networks for financial fraud detection: A review. *Frontiers of Computer Science*. arXiv:2411.05815.

Golec M, AlabdulJalil M (2025) Interpretable LLMs for credit risk: A systematic review and taxonomy. *Expert Systems with Applications*. arXiv:2506.04290.

Khan S, Ahmad A, Naseem U, et al. (2025) Model-agnostic explainable AI methods in finance: A systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review* 58: 215. DOI: 10.1007/s10462-025-11215-9.

Khanvilkar K, Kommuru N (2025) Regulatory graphs and generative AI for real-time transaction monitoring. *arXiv preprint arXiv:2506.01093*.

Mohsin MT, Nasim NB (2025) Explaining the unexplainable: A systematic review of explainable AI in finance. *International Journal of Science and Research Archive* 16(3): 476–497. arXiv:2503.05966.

Muller M (2025) AI-based credit scoring at the intersection of GDPR and AI Act: Lessons from the SCHUFA judgment. *ERA Forum* 26: 647–662. DOI: 10.1007/s12027-025-00865-5.

Marc Schmitt (2024) Explainable automated machine learning for credit decisions. *arXiv preprint arXiv:2402.03806*.

Shreya, Pathak H (2025) Explainable artificial intelligence credit risk assessment using machine learning. *arXiv preprint arXiv:2506.19383*.

Tan XW, Kok S (2024) Explainable risk classification in financial reports. *arXiv preprint arXiv:2405.01881*.

Chagahi MH, Delfan N, Dashtaki SM, Moshiri B, Piran MJ (2024) Explainable AI for fraud detection: An attention-based ensemble of CNNs, GNNs, and a confidence-driven gating mechanism. *arXiv preprint arXiv:2410.09069*.

Barron R, Eren ME, Serafimova OM, Matuszek C, Alexandrov B (2025) Bridging legal knowledge and AI: Retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization. In: *Proceedings of ICAIL 2025*. arXiv:2502.20364.

Wiratunga N, Abeyratne R, Jayawardena L, Martin K, Massie S, Nkisi-Orji I, Weerasinghe R, Liret A, Fleisch B (2024) CBR-RAG: Case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In: *Proceedings of ICCBR 2024*. LNCS 14775. arXiv:2404.04302.

Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE (2024) Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies* (2025). DOI: 10.1111/jels.12413. arXiv:2405.20362.

Kalra R, Wu Z, Gulley A, Hilliard A, Guan X, Koshiyama A, Treleaven P (2024) HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications. In: *CustomNLP4U Workshop, EMNLP 2024*, 237–256. arXiv:2409.09046.

Pipitone N, Houir Alami G (2024) LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Song D, Bonifazi G, Schilder F, Schwarz JR (2025) Knowledge graph-assisted LLM post-training for enhanced legal reasoning. *arXiv preprint arXiv:2601.13806*.

de Martim H (2025) An ontology-driven graph RAG for legal norms: A structural, temporal, and deterministic approach. *arXiv preprint arXiv:2505.00039*.

Acharya K, Raza W, Dourado Jr CMM, Velasquez A, Song H (2023) Neurosymbolic reinforcement learning and planning: A survey. *Transactions on Machine Learning Research*. arXiv:2309.01038.

Rober N, How JP (2024) Constraint-aware refinement for safety verification of neural feedback loops. *IEEE Control Systems Letters*. arXiv:2410.00145.

Wang C, Ji K, Geng J, Ren Z, Fu T, Yang F, Guo Y, He H, Chen X, Zhan Z, Du Q, Su S, Li B, Qiu Y, Du Y, Li Q, Yang Y, Lin X, Zhao Z (2024) Imperative learning: A self-supervised neuro-symbolic learning framework for robot autonomy. *International Journal of Robotics Research* (2025). arXiv:2406.16087.

Wu Y, Xiong Z, Hu Y, Iyengar SS, Jiang N, Bera A, Tan L, Jagannathan S (2025) SELP: Generating safe and efficient task plans for robot agents with large language models. In: *Proceedings of ICRA 2025*. IEEE. arXiv:2409.19471.

Scarbro W, Imrie C, Yaman SG, et al. (2025) Conformal safety shielding for imperfect-perception agents. In: *Proceedings of RV 2025*. LNCS 16087. arXiv:2506.17275.

Sharifi I, Yildirim M, Fallah S (2023) Towards safe autonomous driving policies using a neuro-symbolic deep reinforcement learning approach. *Transportation Research Record* (2026). arXiv:2307.01316.

Corsi D, Mallik K, Rodriguez A, Sanchez C (2025) Efficient dynamic shielding for parametric safety specifications. In: *Proceedings of ATVA 2025*, 157–179. arXiv:2505.22104.

Zhang S, So O, Garg K, Fan C (2024) GCBF+: A neural graph control barrier function framework for distributed safe multi-agent control. *IEEE Transactions on Robotics* 41: 1533–1552. arXiv:2401.14554.

Reed R, Lahijanian M (2024) Learning-based shielding for safe autonomy under unknown dynamics. *arXiv preprint arXiv:2410.07359*.

Hao Y, Chen Y, Zhang Y, Fan C (2025) Large language models can solve real-world planning rigorously with formal verification tools. In: *Proceedings of NAACL 2025*, 3434–3483. arXiv:2404.11891.

Parameshwaran A and Wang Y (2025) Scalable and interpretable verification of image-based neural network controllers for Autonomous Vehicles. In: *Proceedings of ICCPS 2025*. ACM. arXiv:2501.14009.

Chen Y, Gandhi R, Zhang Y, Fan C (2023) NL2TL: Transforming natural languages to temporal logics using large language models. In: *Proceedings of EMNLP 2023*, 15880–15903. arXiv:2305.07766.

Colelough BC, Regli W (2025) Neuro-symbolic AI in 2024: A systematic review. *arXiv preprint arXiv:2501.05435*.

Wang Y, Zhou W, Fan J, et al. (2024) POLAR-Express: Efficient and precise formal reachability analysis of neural-network controlled systems. *IEEE TCAD* 43(3): 994–1007. arXiv:2304.01218.

Yang F, Zhan SS, Wang Y, Huang C, Zhu Q (2024) Case study: Runtime safety verification of neural network controlled system. *arXiv preprint arXiv:2408.08592*.

Yang Z, Raman SS, Shah A, Tellex S (2024) Plug in the safety chip: Enforcing constraints for LLM-driven robot agents. In: *Proceedings of ICRA 2024*, 14435–14442. arXiv:2309.09919.

Shi Y, Wang Z, Chen X, et al. (2024) Aegis: Synthesizing efficient and permissive programmatic runtime shields for neural policies. *arXiv preprint arXiv:2410.05641*.

Silver T, et al. (2023) A framework for neurosymbolic robot action planning using large language models. *arXiv preprint arXiv:2303.00438*.

Corsi D, Amir G, Rodriguez A, Sanchez C, Katz G, Fox R (2024) Verification-guided shielding for deep reinforcement learning. In: *Proceedings of RLC 2024*. arXiv:2406.06507.

Gokhale A, Srivastava V, Bullo F (2025) LogicGuard: Improving embodied LLM agents through temporal logic based critics. *arXiv preprint*.

Wang L, Ying Z, Yang X, et al. (2025) RoboSafe: Safeguarding embodied agents via executable safety logic. *arXiv:2512.21220v2*.

Hafez A, Akhormeh A, Hegazy A, et al. (2025) Safe LLM-controlled robots with formal guarantees via reachability analysis. *arXiv preprint arXiv:2503.03911*.

Zheng Y, Wang X, Liu H, et al. (2025) NeuroStrata: Harnessing neurosymbolic paradigms for improved design, testability, and verifiability of autonomous CPS. *arXiv preprint arXiv:2502.12267*.

Schmude T, Yurrita M, Alfrink K, et al. (2025) Two means to an end goal: Connecting explainability and contestability in the regulation of public sector AI. *arXiv preprint arXiv:2504.18236*.