When the AI Hallucinates, Who Pays? Mapping LLM Hallucination Types to Legal Liability in EU and Swiss Financial Regulation

Joerg Osterrieder[a,*], Lennart Baals[a]

[a] *University of Applied Sciences of the Grisons (FHGR), Chur, Switzerland*

Abstract

Financial institutions are deploying large language models (LLMs) across compliance-critical functions including client onboarding (KYC/KYB), sanctions screening, beneficial ownership identification, and suspicious activity reporting. These models hallucinate – they produce outputs that are factually wrong, numerically distorted, entirely fabricated, or plausible but unsupported by source documents. In regulated financial services, such errors are not mere inconveniences; they can trigger anti-money laundering violations, sanctions breaches, data protection infringements, and supervisory enforcement actions. Yet current regulatory frameworks – the EU AI Act (Regulation 2024/1689), the Swiss Federal Act on Data Protection (FADP, 2023 revision), FINMA Guidance 08/2024, and the Swiss Anti-Money Laundering Act (AMLA) – address AI-generated inaccuracies only in general terms, without distinguishing hallucination as a distinct failure mode with its own risk profile and liability implications. This paper closes that gap. We develop a domain-specific hallucination taxonomy for financial compliance contexts, distinguishing four types: factual, numerical, fabricated, and unsupported hallucinations. We then construct a mapping framework that connects each hallucination type to specific legal obligations, liability regimes, and enforcement mechanisms across EU and Swiss regulation. Through three detailed case scenarios – a missed beneficial owner due to factual hallucination, a fabricated regulatory exemption enabling non-compliant onboarding, and a

*Corresponding author

Email addresses: joerg.osterrieder@fhgr.ch (Joerg Osterrieder), lennart.baals@fhgr.ch (Lennart Baals)

numerical hallucination causing a failed suspicious transaction report – we demonstrate how hallucination types activate different liability pathways. We conclude with concrete recommendations for compliance AI governance, arguing that regulators should require hallucination-type-specific risk management rather than treating all AI inaccuracies identically.

Keywords: LLM hallucination, financial compliance, EU AI Act, FINMA, liability, KYC, anti-money laundering

---

1. Introduction

1.1. The compliance AI promise

The financial services industry is undergoing a rapid transformation in how it approaches regulatory compliance. Large language models (LLMs) – neural network architectures trained on vast corpora of text data and capable of generating, summarizing, and extracting information from unstructured documents – are being deployed across the compliance value chain. Know Your Customer (KYC) and Know Your Business (KYB) processes, which traditionally required compliance officers to manually extract identity information from passports, corporate registries, and beneficial ownership declarations, are now augmented or partially automated by LLM-based extraction pipelines [13]. Politically exposed person (PEP) screening, sanctions list matching, ultimate beneficial owner (UBO) identification, and suspicious transaction report (STR) filing have all become targets for LLM-assisted automation.

The scale of adoption is substantial. FINMA's April 2025 supervisory survey indicated that approximately half of supervised Swiss financial institutions were either deploying or actively developing AI-based tools for compliance functions.[1] Within the European Union, the European Banking Authority (EBA) has acknowledged a "significant increase" in AI deployment for anti-money laundering (AML) processes since 2023 [16]. Industry-academia collaborations have intensified: Innosuisse-funded research projects are developing compliance copilots that embed LLM extraction capabilities directly

[1] FINMA, *Survey on the Use of Artificial Intelligence by Supervised Institutions*, April 2025.

into client onboarding workflows, treating the AI not as a standalone tool but as an integrated component of the compliance decision pipeline.[2]

The economic incentive is clear. Manual KYC processing costs European banks an estimated EUR 60–80 per client onboarding case, with enhanced due diligence (EDD) cases exceeding EUR 300 [17]. LLM-assisted extraction can reduce processing time by 60–80%, with corresponding cost reductions. For a mid-sized Swiss private bank processing 5,000 new client relationships annually, the potential savings are measured in millions of Swiss francs.

1.2. The hallucination problem

LLMs hallucinate. This is not a bug that will be fixed in a future model release; it is a structural property of autoregressive language models that generate text by predicting the next most probable token given preceding context [1]. The model does not "know" facts – it approximates statistical patterns in training data, and when those patterns are insufficient, ambiguous, or conflicting, the model generates outputs that are linguistically fluent but factually wrong.

In consumer-facing applications – chatbots, content generation, search summarization – hallucinations produce embarrassing or misleading outputs that are quickly correctable and carry limited legal consequence. In regulated financial compliance, the stakes are categorically different. A hallucinated client name triggers an identity verification failure under the Anti-Money Laundering Act (AMLA). A hallucinated transaction amount can suppress a suspicious transaction report that the Money Laundering Reporting Office Switzerland (MROS) should have received. A fabricated regulatory citation can lead a compliance team to approve client onboarding under a due diligence exemption that does not exist.

The critical distinction is this: in compliance AI, errors have *directional regulatory consequences*. An LLM that over-identifies risk (false positives in sanctions screening) creates operational cost but no regulatory violation. An LLM that under-identifies risk (false negatives through hallucination) creates direct regulatory exposure. The directionality of failure determines whether a hallucination is merely costly or potentially illegal.

[2] See, e.g., Innosuisse-funded research projects (2026–2027) investigating AI orchestration for Swiss financial compliance automation.

1.3. The liability gap

Current regulatory frameworks address AI accuracy in general terms but do not distinguish hallucination as a failure mode distinct from other forms of AI error. The EU AI Act (Regulation 2024/1689) requires that high-risk AI systems achieve "an appropriate level of accuracy" (Art. 15) and that deployers implement "human oversight" (Art. 14), but neither provision differentiates between a model that systematically underperforms on a metric and a model that sporadically fabricates outputs with high confidence. The Swiss Federal Act on Data Protection (FADP), in its 2023 revision, establishes rights concerning automated individual decision-making (Art. 21) but does not address the specific risk that the data feeding those decisions may itself be AI-generated and wrong. FINMA Guidance 08/2024 on the use of artificial intelligence adopts a principle-based approach requiring "fit for purpose" validation and model risk management, but treats AI inaccuracy as a unitary phenomenon.

This matters because different hallucination types trigger different legal obligations. A factual hallucination (wrong name) and a fabricated hallucination (invented regulation) expose the deploying institution to entirely different liability regimes, involve different parties in the liability chain, and require different mitigation strategies. Treating all AI inaccuracies identically produces compliance frameworks that are simultaneously over-broad (imposing unnecessary controls on detectable errors) and under-inclusive (missing the specific risks of fabricated or unsupported outputs).

1.4. Paper contribution and structure

This paper makes three contributions. First, we develop a domain-specific hallucination taxonomy for financial compliance contexts, distinguishing four types: factual (F), numerical (N), fabricated (B), and unsupported (U) hallucinations (Section 2). Second, we construct a mapping framework that connects each hallucination type to specific legal obligations, liability regimes, and enforcement mechanisms across EU and Swiss regulation (Section 4). Third, through three detailed case scenarios, we demonstrate how different hallucination types activate different liability pathways, producing materially different legal consequences for deployers, providers, and supervisory authorities (Section 5).

Section 3 surveys the relevant regulatory landscape. Section 6 discusses implications for regulators, financial institutions, and AI providers. Section 7 concludes. Appendix Appendix A provides the complete mapping

matrix, and Appendix Appendix B offers a cross-jurisdictional regulatory comparison.

2. Background: LLM Hallucination in Financial Contexts

2.1. Defining hallucination for regulated environments

The computer science literature defines LLM hallucination broadly. Ji et al. define it as "generated content that is nonsensical or unfaithful to the provided source content" [1]. Huang et al. distinguish intrinsic hallucinations (outputs contradicting the source) from extrinsic hallucinations (outputs unverifiable against the source) [2]. Rawte et al. extend the taxonomy to foundation models more generally, identifying hallucination as a cross-cutting concern affecting text generation, summarization, and information extraction [4]. Weidinger et al. frame hallucination within a broader risk taxonomy for language models, situating it alongside misinformation, discrimination, and privacy risks [3].

These definitions, while technically precise, are insufficient for regulated environments. They focus on the relationship between model output and source text, without considering the *regulatory consequence* of the discrepancy. In a compliance context, a hallucination that substitutes "Zurich" for "Zug" in a client address may have no regulatory consequence (both are valid Swiss cantons). The same factual error in a jurisdiction field – substituting "Switzerland" for "Syria" – could suppress a sanctions match and constitute a violation of SECO sanctions ordinances.[3]

We therefore adopt a compliance-specific definition: *In financial compliance, a hallucination is any AI-generated output that would lead a reasonable compliance officer to make a regulatory decision they would not have made with accurate information.* This definition is outcome-oriented rather than text-oriented. It captures not only clear factual errors but also unsupported inferences and fabricated references that shift the compliance decision boundary, while excluding discrepancies that are factually wrong but regulatorily inert.

[3]State Secretariat for Economic Affairs (SECO), Sanctions Ordinance against Syria, SR 946.231.172.7.

2.2. The four hallucination types in compliance AI

Building on the general taxonomy of Ji et al. [1] and the domain-specific risk categorization of Weidinger et al. [3], we propose four hallucination types relevant to financial compliance. These types are not mutually exclusive – a single model output may contain multiple hallucination types simultaneously – but they are analytically distinct in their evidentiary characteristics, detectability, and regulatory implications.

2.2.1. Type F: Factual hallucination

A factual hallucination occurs when the model outputs an incorrect fact where a correct fact exists in the source material. The defining characteristic is that the error is *verifiable against a ground truth document*. The model had access to the correct information but produced an incorrect extraction or transformation.

Example. A compliance extraction pipeline processes a Swiss commercial register extract for a corporate client. The document states that the company's registered agent is "Hans Müller, Zug." The LLM extracts "Hans Müller, Zurich." The factual error – substituting one Swiss canton for another – may appear minor, but it affects the determination of the competent cantonal commercial register and may interfere with subsequent UBO verification steps that depend on matching the registered address.

In more consequential cases, factual hallucinations involve name substitutions ("Pierre Dubost" extracted as "Pierre Dubois"), nationality errors (Belgian extracted as French), or date transpositions (incorporation date 2019 extracted as 2009) that directly affect PEP screening, sanctions matching, or risk scoring. The distinguishing feature is that the correct answer exists in the source document and the model's error is deterministically verifiable.

2.2.2. Type N: Numerical hallucination

A numerical hallucination occurs when the model outputs an incorrect number, amount, percentage, or quantitative value. While technically a subtype of factual hallucination, numerical errors warrant separate treatment because they interact with regulatory thresholds that create binary legal consequences.

Example. A transaction monitoring system processes a wire transfer instruction. The document states a transfer amount of CHF 150,000. The LLM extracts CHF 15,000. This order-of-magnitude error places the transaction below the CHF 25,000 threshold that triggers enhanced monitoring under

FINMA's AML supervision framework.[4] The consequence is not merely an incorrect data point; it is the suppression of a legally mandated monitoring action.

Numerical hallucinations are particularly dangerous at regulatory decision boundaries. The CHF 100,000 threshold for cash transaction reporting, the 25% beneficial ownership threshold, the EUR 10,000 threshold under the EU's Anti-Money Laundering Regulation – all create cliff effects where a small numerical error has outsized regulatory consequences. Decimal point shifts, digit transpositions, and currency confusion (CHF vs. EUR) are the most common mechanisms.

2.2.3. Type B: Fabricated hallucination

A fabricated hallucination occurs when the model generates content – entities, documents, regulations, citations – that has no basis in either the source documents or external reality. The model does not misread a fact; it invents one.

Example. During a KYB onboarding review, the compliance AI is asked to identify applicable regulatory exemptions. The model states: "Under FINMA Circular 2024/7 on Simplified Due Diligence for Digital Asset Holders, entities with verified blockchain-based identity attestations are exempt from standard documentary evidence requirements for transactions below CHF 5 million." No such FINMA Circular exists. The model has fabricated a regulatory citation, complete with a plausible numbering scheme, a topically relevant subject matter, and specific threshold conditions.

Fabricated hallucinations are qualitatively different from factual errors. They do not involve misreading source material; they involve the model drawing on its training data to generate outputs that are statistically plausible but factually nonexistent. In compliance contexts, fabricated regulatory citations are particularly dangerous because they may be difficult for non-specialist compliance officers to detect – the output *reads* like a real regulation, uses correct formatting and terminology, and addresses a real regulatory question. Detection requires affirmative verification against the actual regulatory corpus, not merely cross-checking against the source document.

[4]FINMA, Circular 2011/1 "Activity of financial intermediaries under AMLA," margin no. 65–72.

2.2.4. Type U: Unsupported hallucination

An unsupported hallucination occurs when the model generates an output that may or may not be true but is not supported by the available source documents. The output is an inference, assumption, or extrapolation that the model presents as established fact.

Example. A KYC extraction pipeline processes onboarding documents for a high-net-worth individual. The model's output includes: "The client's funds originate from inheritance, consistent with the declared family wealth background." No document in the onboarding package mentions inheritance. The model has inferred a plausible source of funds based on contextual cues (age, wealth level, family references) but presents this inference as documentary fact.

Unsupported hallucinations are the most epistemically treacherous type. Unlike factual or numerical hallucinations, which are deterministically wrong, unsupported claims may be coincidentally correct. The client's funds *may* originate from inheritance – the model's inference is plausible. But the compliance decision must rest on documented evidence, not on model inference. The danger is that the unsupported claim fills an evidential gap in the compliance file, creating the appearance of a complete assessment where documentary evidence is actually missing. This violates the fundamental AML principle that compliance judgments must be evidence-based, not inference-based.

2.3. Why the type distinction matters for liability

The four hallucination types create fundamentally different evidentiary and liability situations. Type F hallucinations are detectable through comparison with source documents – a reasonable compliance officer exercising ordinary diligence should catch them, which has implications for the human oversight defense under the EU AI Act. Type N hallucinations are similarly verifiable and become legally decisive at specific regulatory thresholds, creating sharp liability boundaries. Type B hallucinations require affirmative domain expertise to detect, because the fabricated content is designed (by the model's training dynamics, not intentionally) to be plausible; this shifts the liability calculus toward the AI provider, since the deployer's ability to detect the error depends on the model's tendency to fabricate. Type U hallucinations are the most dangerous from a liability perspective: they may be undetectable by any means short of independent investigation, because the claim is plausible and unfalsifiable from the source documents alone.

These differences matter for three liability questions. First, *foreseeability*: was the hallucination type one that the deployer should have anticipated and mitigated? Second, *detectability*: could a human oversight mechanism reasonably have caught the error, and does failure to catch it constitute negligence? Third, *causation*: did the hallucination cause the regulatory violation, or would the violation have occurred regardless? Each hallucination type produces different answers to these questions, and therefore activates different liability pathways – a core claim we develop in the mapping framework of Section 4.

3. Regulatory Landscape

3.1. The EU AI Act

Regulation (EU) 2024/1689, the EU Artificial Intelligence Act, entered into force on 1 August 2024 and establishes the world's first comprehensive regulatory framework for artificial intelligence [6]. The Act adopts a risk-based approach, classifying AI systems into prohibited, high-risk, limited-risk, and minimal-risk categories. AI systems used for creditworthiness assessment and credit scoring are explicitly listed as high-risk in Annex III (point 5(b)), and AI systems used in contexts covered by Union AML legislation fall within the scope of high-risk classification where they serve as safety components or make decisions affecting natural persons.

Several provisions are directly relevant to hallucination liability in financial compliance, though their interaction reveals gaps the drafters appear not to have anticipated.

The Act's accuracy requirements suffer from a critical gap. Art. 15(1) demands that high-risk AI systems "achieve an appropriate level of accuracy, robustness and cybersecurity," and Art. 15(2) requires that accuracy levels "and the relevant accuracy metrics shall be declared in the accompanying instructions of use." Yet nothing in Art. 15 compels accuracy to be reported by hallucination type. A model that hallucinates names at a 2% rate, numbers at a 5% rate, and fabricates citations at a 0.1% rate has a very different risk profile from a model with a uniform 3% inaccuracy rate – but Art. 15 treats both as equivalent. This matters because Art. 9 demands more than Art. 15 delivers. The risk management system required by Art. 9 must identify "known and reasonably foreseeable risks" (Art. 9(2)(a)) and adopt "appropriate and targeted risk management measures" (Art. 9(4)).

The word "targeted" is doing significant work here: a risk management system that treats all AI inaccuracy as a single risk category cannot meet this standard, because it cannot distinguish between error types that require fundamentally different controls. Art. 15 provides the diagnosis ("the system must be accurate") but not the specificity that Art. 9's "targeted" language implies.

The Act's human oversight requirement (Art. 14) interacts with hallucination types in ways the drafters likely did not anticipate. Art. 14(1) requires "effective oversight by natural persons," including the ability to "correctly interpret the high-risk AI system's output" (Art. 14(4)(a)) and to "decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output" (Art. 14(4)(d)). For Type F and Type N hallucinations, this works: a compliance officer cross-references the model's output against source documents and catches the error. For Type B hallucinations, the cognitive task changes qualitatively – detecting a fabricated regulation requires the overseer to recognize that a cited provision does not exist, which demands specialized legal knowledge beyond document cross-referencing. For Type U hallucinations, detection may be functionally impossible, since the unsupported claim is by definition not contradicted by any document in the overseer's possession. This graduated detectability has direct liability implications. Where a hallucination type is detectable through reasonable oversight, failure to detect it may constitute a breach of Art. 14 duties. Where it is not reasonably detectable, the deployer's liability must rest on a different basis – failure to implement adequate technical safeguards under Art. 9, or the provider's failure under Art. 15 to design a sufficiently accurate system.

Art. 26 completes the liability chain by establishing the deployer as the primary responsible party. Deployers must use high-risk AI systems "in accordance with the instructions of use" (Art. 26(1)), ensure that "input data is relevant and sufficiently representative" (Art. 26(4)), and monitor operation for risks (Art. 26(5)). The obligations concerning high-risk AI systems listed in Annex III apply from 2 August 2026.[5] Financial institutions deploying compliance AI tools have a narrow window to align their governance frameworks with these requirements.

[5]EU AI Act, Art. 113(3)(b).

3.2. Swiss Federal Act on Data Protection (FADP)

The revised Swiss Federal Act on Data Protection (FADP, SR 235.1), which entered into force on 1 September 2023, modernized Swiss data protection law to approximate GDPR-level standards while maintaining the Swiss principle-based regulatory tradition [7]. For compliance AI, two provisions are particularly relevant.

The FADP's provision on automated individual decision-making (Art. 21) raises a threshold question: when does LLM-assisted compliance processing become "solely" automated? Art. 21(1) requires that where a decision based "solely on automated processing" has "legal effects" on or "significantly affects" a natural person, the data subject must be informed and given the opportunity to request human review. Compliance decisions – particularly adverse ones such as refusal of onboarding, enhanced due diligence classification, or suspicious activity reporting – plainly meet the "significant effect" threshold. If the compliance officer reviews the LLM's output before making the compliance decision, the process may fall outside Art. 21's scope. But if the LLM's hallucinated output effectively determines the compliance outcome – because the compliance officer relies on it without meaningful independent verification – the process is *de facto* automated regardless of its formal design.

Data protection impact assessments (DPIAs) under Art. 22 FADP are mandatory where processing involves automated decision-making or large-scale processing of sensitive data, both of which are characteristic of compliance AI deployments. A DPIA that does not address hallucination risk – specifically, the risk that the AI produces incorrect personal data that then drives compliance decisions – is incomplete.

3.3. FINMA Guidance 08/2024

The Swiss Financial Market Supervisory Authority (FINMA) issued Guidance 08/2024 on the supervisory expectations for the use of artificial intelligence by supervised institutions [8]. Unlike the prescriptive requirements of the EU AI Act, FINMA's approach is principle-based: it sets expectations without mandating specific technical measures, reflecting the Swiss regulatory tradition of outcomes-based supervision.

The Guidance identifies three risk categories for AI deployment: model risk, operational risk, and conduct risk. For compliance AI, all three are engaged. Model risk arises from the LLM's propensity to hallucinate – producing outputs that deviate from ground truth in ways that affect compli-

ance decisions. Operational risk arises from the integration of LLM outputs into compliance workflows, where a hallucinated output propagates through downstream processes (sanctions screening, risk scoring, reporting) before detection. Conduct risk arises when hallucinated outputs lead to decisions that harm clients (denial of services based on fabricated risk indicators) or third parties (failure to report suspicious activity).

FINMA requires supervised institutions to ensure that AI systems are "fit for purpose" – that is, that the model has been validated for the specific compliance use case in which it is deployed.[6] This validation must be "ongoing and commensurate with the risk posed by the application." For compliance AI, "fit for purpose" validation must address hallucination risk, but the Guidance does not specify whether this means aggregate accuracy testing or hallucination-type-specific evaluation. We argue below that aggregate testing is insufficient.

The Guidance connects to existing FINMA Circulars on operational risk (Circular 2023/1 "Operational Risks and Resilience – Banks") and outsourcing (Circular 2018/3 "Outsourcing – Banks and Insurers"). The outsourcing provisions are relevant because many financial institutions deploy third-party LLMs (from providers such as OpenAI, Anthropic, Google, or Mistral) through API integrations, which may constitute material outsourcing of a regulated function.[7] Under the outsourcing framework, the deploying institution retains full regulatory responsibility for the outsourced function, including liability for errors in the outsourced service – a principle that applies directly to LLM hallucinations regardless of whether the model is operated internally or accessed via a third-party API.

3.4. AML obligations

The Swiss Anti-Money Laundering Act (AMLA, SR 955.0) and its implementing ordinances establish mandatory due diligence obligations for financial intermediaries [9]. Several provisions are directly triggered by compliance AI hallucinations, and they form a causal chain that amplifies the consequences of each error type.

The chain begins with identity verification. Art. 3(1) AMLA requires financial intermediaries to verify the identity of contracting parties based on "a

[6]FINMA Guidance 08/2024, Section 3.2.
[7]FINMA Circular 2018/3, margin no. 6–8.

document of evidentiary value." When an LLM extracts identity information from such documents and hallucinates – producing a wrong name (Type F), wrong nationality (Type F), or wrong date of birth (Type N) – the identity verification is formally deficient. The financial intermediary cannot satisfy Art. 3 with hallucinated identity data, regardless of whether the error was produced by a human or a machine. But the consequences do not stop at a deficient record. Incorrect identity data feeds directly into the risk assessment that determines enhanced due diligence obligations under Art. 6(1) AMLA, which requires that where there are "indications of increased risk," the financial intermediary must "clarify the economic background and purpose of a transaction or business relationship." EDD triggers include PEP status, high-risk jurisdictions, unusual transaction patterns, and complex corporate structures. If an LLM hallucination suppresses an EDD trigger – by extracting a non-PEP name instead of a PEP name (Type F), or by understating a transaction amount below the EDD threshold (Type N) – the financial intermediary fails to apply legally mandated enhanced measures. Wrong identity data under Art. 3 produces wrong risk assessments under Art. 6: the errors compound.

The most consequential provision is AMLA Art. 9, which imposes a reporting duty to MROS when the financial intermediary "know[s] or ha[s] reasonable grounds to suspect" that a transaction involves proceeds of crime (Art. 9(1)). It bears emphasizing that this duty is not contingent on certainty; "reasonable grounds to suspect" is a low threshold. An LLM hallucination that suppresses suspicious indicators – understating amounts (Type N), omitting adverse information (Type U), or providing false reassurance through fabricated regulatory exemptions (Type B) – may prevent the formation of "reasonable grounds to suspect" and thereby suppress a legally mandated report. Under Art. 37 AMLA, failure to report is a criminal offense, punishable by a fine of up to CHF 500,000 for intentional violations and CHF 150,000 for negligent violations.

The FATF Recommendations, particularly Recommendation 10 (Customer Due Diligence) and Recommendation 20 (Reporting of Suspicious Transactions), establish the international standards that AMLA implements [10]. The FATF's 2024 guidance on the use of new technologies in AML/CFT acknowledges that AI tools can enhance compliance effectiveness but cautions that "the use of new technologies does not diminish the obligation to conduct adequate due diligence" [11]. Switzerland's recent enactment of the Federal Act on the Transparency of Legal Entities (LETA), establishing a

beneficial ownership register, creates additional data points against which LLM-extracted UBO information can be verified – and additional liability exposure when hallucinated UBO data diverges from the register.

4. The Mapping Framework

4.1. Framework structure

We construct a three-dimensional mapping framework connecting (i) hallucination types, (ii) compliance use cases, and (iii) regulatory obligations. The framework is designed to answer a specific question: given a particular hallucination type occurring in a particular compliance use case, which regulatory obligations are violated, who bears liability, and what enforcement mechanisms apply?

The three dimensions are:

- **Hallucination type**: Factual (F), Numerical (N), Fabricated (B), Unsupported (U)
- **Compliance use case**: KYC/KYB onboarding, PEP screening, UBO identification, STR filing
- **Regulatory regime**: EU AI Act (Regulation 2024/1689), Swiss FADP (SR 235.1), FINMA supervisory framework (Guidance 08/2024, Circulars), Swiss AML framework (AMLA SR 955.0, AMLO-FINMA)

For each cell in the resulting matrix, we identify: the specific regulatory provision violated, the primary liable party (deployer, provider, or both), the enforcement mechanism (supervisory action, civil liability, criminal sanction), and a severity assessment (low, medium, high, critical). The full matrix is presented in Appendix Appendix A; here we describe the key findings and their analytical logic.

4.2. Mapping matrix: hallucination types to legal obligations

4.2.1. Type F (Factual) mappings

Factual hallucinations are the easy case. Wrong names, dates, nationalities, jurisdictions – these map directly to identity verification obligations under AMLA Art. 3 and accuracy requirements under EU AI Act Art. 15. A factual hallucination in identity data creates a gap between the documentary evidence and the recorded data, which is a formal compliance deficiency regardless of its practical significance. Under FADP Art. 6(5), the data subject's right to data accuracy is also engaged.

Because factual hallucinations are detectable through standard document cross-referencing, the deployer's Art. 14 human oversight obligation provides both a defense (if oversight was effective and the error was caught) and a basis for liability (if it was not). Existing compliance practice already answers the question of who is responsible. The financial institution bears primary responsibility: its obligation to verify AI outputs against source documents is well-established, and the error type does not demand specialized detection methods.

4.2.2. Type N (Numerical) mappings

Numerical hallucinations produce the most legally deterministic outcomes because financial regulation operates through quantitative thresholds. When an LLM extracts CHF 15,000 instead of CHF 150,000, the legal consequence is binary: either the correct amount would have triggered a regulatory obligation that the incorrect amount did not, or it would not have. Unlike factual hallucinations, where a wrong middle name may be inconsequential but a wrong nationality may be critical, numerical errors at regulatory boundaries have unambiguous causation – a key element of any liability claim.

Under AMLA Art. 9, the reporting duty to MROS is triggered by transactions involving suspected proceeds of crime. Transaction amounts are a primary indicator: while AMLA does not establish a mechanical reporting threshold, FINMA supervisory practice treats high-value transactions as requiring heightened attention, and internal policies typically establish amount-based escalation rules. A numerical hallucination that understates a transaction amount suppresses this escalation. Under the EU AI Act, Art. 15 (accuracy) is engaged with particular force precisely because of this threshold determinism. The system either correctly identifies the amount and the obligation is triggered, or it does not and the obligation is missed.

The human oversight defense under Art. 14 is available: a compliance officer comparing the extracted amount against the source document will detect the discrepancy. Yet here lies an unresolved tension in the regulatory framework. In high-volume transaction monitoring environments – precisely the environments where LLM-assisted extraction is most valuable – individual transaction-level review may not be feasible. The economic justification for AI-assisted processing is volume efficiency; the legal requirement for human oversight demands individual accuracy. The deployer bears primary liability for implementing adequate cross-checking, but the practical feasibility of transaction-level review in high-volume environments remains a tension that

neither the EU AI Act nor FINMA guidance has resolved.

4.2.3. Type B (Fabricated) mappings

Fabricated hallucinations – invented entities, documents, or regulations – occupy a unique position in the liability framework because they are the only hallucination type that can activate *provider* liability in addition to deployer liability.

Under AMLA, a fabricated regulatory exemption that leads a financial intermediary to apply simplified due diligence instead of standard or enhanced due diligence constitutes a violation of the applicable due diligence standard. The intermediary cannot argue that it relied on a regulatory exemption that does not exist. However, the mechanism of failure – an AI system that fabricated the exemption – raises the question of whether the AI provider bears secondary responsibility.

Under the EU AI Act, fabricated hallucinations engage Art. 9 (risk management) because fabrication is a foreseeable risk of LLM deployment that must be identified and mitigated. They also engage Art. 14 (human oversight), but with a critical caveat that distinguishes this type from all others. Detecting a fabricated regulation requires the human overseer to *know* that the regulation does not exist. Asking a compliance officer to verify that a regulation exists is verification; asking her to verify that a regulation does not exist is an open-ended search through the entire regulatory corpus. This is a qualitatively different oversight task from cross-referencing an extraction against a source document – it requires affirmative legal knowledge, not just verification skill. The human oversight defense is therefore substantially weaker for Type B hallucinations than for Types F or N.

Crucially, fabricated hallucinations can engage the revised Product Liability Directive (Directive 2024/2853) [14]. If an AI system fabricates a regulatory citation that leads to a compliance failure, the AI system has produced a "defective" output within the meaning of product liability law. The proposed AI Liability Directive (COM(2022) 496) [15] would further facilitate claims against AI providers by establishing a presumption of causality and easing the burden of proof for claimants. Under these instruments, the AI provider – not merely the deployer – faces liability exposure for fabricated outputs.

This dual-track liability structure – deployer for the compliance failure, provider for the defective output – is, we argue, the most significant practical implication of type-specific hallucination analysis. No other hallucination

type routinely splits responsibility across the deployer-provider boundary in this way.

4.2.4. Type U (Unsupported) mappings

Unsupported hallucinations expose what may be the deepest tension in the regulatory framework: the gap between factual accuracy and evidential sufficiency. Consider the claim "the client's funds originate from inheritance." It may be true. But under AMLA Art. 3 and Art. 6, due diligence obligations require the financial intermediary to base compliance judgments on documentary evidence, and this claim is not documented. If the compliance officer accepts the LLM's output as a documented finding, the compliance file contains an undocumented assertion masquerading as an evidentiary finding. The documentary evidence principle is violated even if the assertion happens to be correct.

How should Art. 15 accuracy be measured for claims that are neither verifiably true nor verifiably false? This is the conceptual challenge that unsupported hallucinations pose. The output is not verifiably *wrong* – it is *unverifiable*. Standard accuracy metrics (precision, recall, F1) cannot capture this failure mode, because there is no ground truth label against which to evaluate the claim. Art. 15 accuracy requirements, as currently formulated, are insufficient to address this category and must be supplemented with evidential sufficiency requirements.

The human oversight defense under Art. 14 is at its weakest here. The overseer cannot detect the hallucination by comparing the output against the source document – the claim is not contradicted by any document. Detection would require the overseer to notice the *absence* of supporting evidence, to recognize that the claim, while plausible, is not grounded in any document in the file. This negative verification is cognitively demanding and unlikely to be performed consistently at scale. We accept that the deployer bears primary responsibility for ensuring that compliance decisions rest on documented evidence. But the genuine difficulty of detection should reduce the negligence standard applied – a financial institution that implemented reasonable documentary-verification procedures cannot fairly be held to the same standard for unsupported claims as for factual errors that a simple document comparison would have caught.

4.3. Liability allocation analysis

Table 1 summarizes the liability allocation across hallucination types.

Table 1: Liability allocation by hallucination type

Dimension	**Type F**	**Type N**	**Type B**	**Type U**
Detectability	High	High	Medium	Low
Human oversight	Effective	Effective	Requires expertise	Largely ineffective
Deployer liability	Primary	Primary	Primary	Primary
Provider liability	Low	Low	Significant	Low
Negligence standard	Ordinary care	Ordinary care	Heightened care	Reduced standard
Criminal exposure	Low	Medium[a]	Medium	Low
Key EU AI Act art.	14, 15, 26	9, 15, 26	9, 14, 15	9, 15
Key AMLA art.	3	9, 37	3, 6	3, 6

[a] Criminal exposure under Art. 37 AMLA for failure to report.

The most fundamental finding is also the least surprising: the deployer bears primary liability regardless of hallucination type. This follows from the principle that the entity deploying AI in a regulated function retains full regulatory responsibility for that function's outcomes. Article 26 of the EU AI Act, the AMLA's due diligence obligations, and FINMA's supervisory framework all assign primary responsibility to the deploying institution. The use of AI does not transfer regulatory obligations to the technology provider. This principle is well-established, but its implications for AI governance are underappreciated – it means that the financial institution cannot contractually or operationally shift compliance risk to the AI vendor.

Where the analysis departs from conventional wisdom is in its treatment of provider liability. Only Type B (fabricated) hallucinations routinely activate provider liability, because fabrication represents a product defect – the AI system is generating content that is not merely inaccurate but fictional. Under the revised Product Liability Directive, an AI system that fabricates regulatory citations is "defective" in the product liability sense: it does not provide the safety that a person is entitled to expect. The proposed AI Liability Directive would reinforce this by establishing a rebuttable presumption of causality for AI-caused damage and by providing for disclosure of evidence from the AI provider. For Types F, N, and U, provider liability is typically limited to contractual claims (breach of service level agreements, warranty claims) rather than statutory liability. The type-dependence of provider liability is, as far as we are aware, not recognized in the existing literature on AI compliance risk.

Perhaps the most troubling implication concerns the EU AI Act's human oversight model itself. For Types F and N, human oversight is effective and the liability analysis is straightforward: the deployer should have caught the error. For Type B, human oversight is partially effective but requires specialized knowledge, creating shared liability between deployer and provider. For Type U, human oversight is largely ineffective, creating a liability gap that current regulation does not adequately address. Art. 14 is implicitly designed for detectable errors – for the world of document cross-referencing and data verification. It does not provide a workable framework for hallucination types that evade human detection, and the Act offers no alternative safeguard for cases where the human-in-the-loop cannot meaningfully function as one. This gap requires regulatory attention before the August 2026 compliance deadline.

5. Case Scenarios

We present three scenarios that demonstrate how the mapping framework operates in practice. Each scenario involves a realistic compliance situation, a specific hallucination type, and a detailed analysis of the resulting legal consequences. The scenarios are fictional but constructed from documented compliance failure patterns and regulatory enforcement precedents.

5.1. Scenario A: The Phantom Beneficial Owner

5.1.1. Facts

A Swiss private bank initiates onboarding for a Luxembourg-incorporated holding company. The compliance AI – an LLM-based extraction pipeline – processes the company's beneficial ownership declaration, commercial register extract, and organizational chart. The source documents identify the ultimate beneficial owner as **Pierre Dubost**, a Belgian national and senior official in the European Investment Bank, classified as a politically exposed person (PEP) under the VSB 24[8] and Annex 1 to the FINMA Anti-Money Laundering Ordinance (AMLO-FINMA).

The LLM extracts the name as **Pierre Dubois** – a one-character substitution that changes the surname from an uncommon Belgian name to a common French name. The system's PEP screening module, receiving "Pierre

[8]Agreement on the Swiss banks' code of conduct with regard to the exercise of due diligence (VSB 24), Art. 57–62.

Dubois, French" (the LLM also hallucinated the nationality from Belgian to French), finds no PEP match. The compliance officer reviews the LLM's extraction report, which presents the output confidently and without confidence scores or source document cross-references.

Additionally, the LLM's risk assessment output includes the statement: "No indication of political exposure or public function for any identified beneficial owner." This statement is an unsupported hallucination – the LLM did not check PEP databases; it generated the statement based on the (already hallucinated) extraction result.

The bank onboards the client with standard due diligence instead of the enhanced due diligence required for PEP relationships. Eighteen months later, a correspondent bank's compliance review identifies Dubost as a PEP, triggering a cascade of remediation.

5.1.2. Legal analysis under the mapping framework

This scenario involves two hallucination types operating in combination: **Type F** (factual – wrong name and nationality) and **Type U** (unsupported – the "no political exposure" statement).

AMLA Art. 3 and Art. 6. The factual hallucination caused a failure to correctly identify the beneficial owner (Art. 3) and the unsupported hallucination caused a failure to apply enhanced due diligence for a PEP relationship (Art. 6). The bank violated both provisions. Under Art. 37 AMLA, the negligent failure to apply EDD exposes the bank to administrative fines.

SECO sanctions. If Dubost were listed on a sanctions list (he is not in this scenario, but the analytical framework applies), the factual hallucination would have caused a sanctions screening miss, potentially constituting a violation of the Embargo Act (SR 946.231) – a criminal offense.

EU AI Act Art. 26 and Art. 14. The bank, as deployer, failed in its human oversight obligation. The factual hallucination (wrong name) was detectable: a comparison of the LLM's extraction against the source document would have revealed the discrepancy. The bank's failure to implement systematic source-document cross-checking constitutes a breach of Art. 26(1) (use in accordance with instructions) and Art. 14 (effective human oversight). However, the unsupported hallucination ("no political exposure") was not detectable through document cross-referencing – it required the compliance officer to recognize that the LLM's statement was an inference rather than a documented finding. The dual-type nature of the hallucination creates a *layered liability*: the bank is clearly negligent for missing the Type F

error and has a stronger defense regarding the Type U error.

Framework application. The mapping framework predicts exactly this dual-track outcome: Type F hallucinations in UBO identification trigger AMLA Art. 3 with deployer as primary liable party and ordinary negligence standard. Type U hallucinations in risk assessment trigger AMLA Art. 6 with deployer as primary liable party but a reduced negligence standard due to low detectability. The combination of types produces a more severe regulatory outcome than either type alone.

5.2. Scenario B: The Invented Exemption

5.2.1. Facts

A Swiss-licensed cryptocurrency exchange uses an LLM-based compliance assistant to support its onboarding decisions. During the onboarding of a high-volume institutional client, the compliance officer queries the system about applicable due diligence standards for digital asset holders. The system responds:

> *"Under FINMA Circular 2024/7 on Simplified Due Diligence for Digital Asset Holders, entities that maintain verified blockchain-based identity attestations issued by a FINMA-recognized certification body are exempt from standard documentary evidence requirements for business relationships involving transaction volumes below CHF 5 million annually. The simplified due diligence regime requires only (a) verification of the blockchain identity attestation, (b) confirmation of the entity's registration with a self-regulatory organization (SRO), and (c) annual review of transaction patterns."*

No such FINMA Circular exists. The model fabricated the citation, including its number, title, threshold, and specific requirements. The fabrication is sophisticated: it uses FINMA's actual numbering convention, addresses a real regulatory question (due diligence for digital assets), references real institutional structures (SROs, certification bodies), and includes plausible quantitative thresholds.

The compliance team, consisting of three officers with 4–8 years of experience in traditional banking compliance but limited familiarity with the specific regulatory landscape for digital assets, accepts the LLM's guidance. The exchange onboards the client with simplified due diligence. Six months

later, the client is linked to a ransomware-related money laundering scheme. FINMA opens enforcement proceedings.

5.2.2. Legal analysis under the mapping framework

This scenario involves a pure **Type B** (fabricated) hallucination.

AMLA and FINMA enforcement. The exchange applied a due diligence standard that does not exist in Swiss law. The onboarding was conducted without adequate identity verification (Art. 3 AMLA) and without the enhanced due diligence that the client's risk profile warranted (Art. 6 AMLA). FINMA's enforcement action will focus on organizational deficiencies: the exchange relied on an AI system for regulatory guidance without implementing verification against the actual regulatory corpus.

EU AI Act Art. 9 and Art. 14. The risk management system (Art. 9) failed to identify fabrication as a foreseeable risk of LLM deployment. The human oversight mechanism (Art. 14) failed because detecting the fabrication required the overseers to *know* that FINMA Circular 2024/7 does not exist – a domain knowledge requirement that exceeds routine document verification. The exchange has a partial defense: the hallucination was not detectable through standard compliance procedures but required affirmative regulatory expertise.

Product Liability Directive. This scenario uniquely activates provider liability. The AI system produced a "defective" output – a fabricated regulatory citation that, when relied upon, caused compliance harm. Under the revised Product Liability Directive (Directive 2024/2853), the AI provider may be liable for damage caused by a defective product, and AI system outputs are within the Directive's scope. The proposed AI Liability Directive (COM(2022) 496) would establish a rebuttable presumption that the AI system caused the damage if the provider failed to comply with applicable EU AI Act obligations. For the exchange's civil claims against its AI vendor, this presumption significantly eases the burden of proof.

Framework application. The mapping framework correctly predicts that Type B hallucinations uniquely activate dual-track liability (deployer + provider). The deployer bears primary regulatory liability (AMLA, FINMA) but has a stronger human oversight defense than for Types F or N. The provider faces secondary liability under product liability and contract law. This dual-track structure is specific to fabricated hallucinations and does not apply to the other types.

5.3. Scenario C: The Decimal Point

5.3.1. Facts

A cantonal bank's transaction monitoring system incorporates an LLM-based component for processing international wire transfer instructions. A client initiates a wire transfer of CHF 125,000 to a beneficiary in a jurisdiction classified as elevated risk under the bank's internal risk matrix. The LLM-based extraction module processes the SWIFT message and extracts the amount as CHF 12,500 – a decimal point error that reduces the stated amount by a factor of ten.

The bank's automated monitoring rules are calibrated to flag transactions above CHF 25,000 to elevated-risk jurisdictions for manual review. The hallucinated amount of CHF 12,500 falls below this threshold. The transaction proceeds without manual review. No suspicious transaction report is filed.

Three months later, MROS notifies the bank that the beneficiary is under investigation for money laundering. MROS requests documentation of the bank's due diligence for all transactions with the beneficiary. The bank discovers the extraction error and the resulting monitoring failure.

5.3.2. Legal analysis under the mapping framework

This scenario involves a pure **Type N** (numerical) hallucination.

AMLA Art. 9 and Art. 37. The bank had a duty to report the suspicious transaction to MROS. The numerical hallucination suppressed the internal monitoring trigger that would have led to manual review and, potentially, a suspicious transaction report. Under Art. 37 AMLA, negligent failure to comply with the reporting duty is a criminal offense punishable by a fine of up to CHF 150,000. The question is whether the bank's failure to detect the numerical error constitutes criminal negligence. Given that the error was deterministically verifiable – a comparison of the extracted amount against the SWIFT message would have revealed the discrepancy – the negligence standard is likely met.

EU AI Act Art. 15. The numerical hallucination represents a clear accuracy failure. The system's declared accuracy level (Art. 15(2)) should have included numerical extraction accuracy, and the decimal point error falls below any reasonable accuracy threshold for a financial extraction system. If the provider failed to declare the system's numerical accuracy characteristics, or if the declared accuracy was misleading, the provider may bear liability under Art. 15.

Human oversight. The bank did not implement transaction-level cross-checking of LLM-extracted amounts against source documents. For a system processing thousands of transactions daily, individual cross-checking may be economically infeasible – but this is precisely the tension that the EU AI Act's human oversight requirement (Art. 14) creates. If human oversight is impractical at the required volume, the deployer must either (a) implement automated verification (not LLM-based) as a secondary check, (b) reduce reliance on LLM extraction for threshold-critical fields, or (c) accept the liability risk. The mapping framework flags this as the central governance question for Type N hallucinations: accuracy at regulatory thresholds cannot be left to probabilistic extraction.

Framework application. The mapping framework identifies Type N hallucinations as the most legally clear-cut liability situation. The error is verifiable, the regulatory threshold is deterministic, and the causal chain from hallucination to legal violation is direct. The deployer bears primary liability with limited defenses. Criminal exposure under Art. 37 AMLA distinguishes this scenario from the others and represents the most severe legal consequence of the three case scenarios.

6. Discussion

6.1. Implications for regulators

If our mapping framework demonstrates anything, it is that aggregate accuracy metrics are uninformative for compliance risk management. The EU AI Act's Art. 15 should be interpreted – and where necessary supplemented by delegated acts or harmonized standards – to require *hallucination-type-specific disclosure*. An AI provider's accuracy declaration should distinguish between factual accuracy (correct extraction of documented facts), numerical accuracy (correct extraction of quantities, particularly at regulatory thresholds), fabrication rate (frequency of generated content without source basis), and evidential sufficiency (whether outputs are grounded in source documents or involve model inference). Without this granularity, a deploying institution cannot calibrate its oversight to the actual risk profile of the system it uses.

The gap is equally acute in the Swiss supervisory framework. FINMA should supplement Guidance 08/2024 with specific expectations for hallucination risk management in compliance AI. The current "fit for purpose" requirement is necessary but insufficient: it does not distinguish between extractive AI (which processes existing documents) and generative AI (which

produces new text), even though the hallucination risk profiles of these two modes are fundamentally different. Extractive AI is primarily exposed to Types F and N; generative AI is primarily exposed to Types B and U. A principle-based guideline recognizing this distinction would enable supervised institutions to calibrate their controls appropriately.

These are not uniquely European or Swiss problems. The FATF, the Basel Committee on Banking Supervision, and IOSCO should develop technology-neutral but hallucination-aware guidance for AI use in compliance. The FATF's 2024 technology guidance acknowledges AI risks but does not provide the analytical granularity needed for effective compliance governance. A FATF recommendation or interpretive note specifically addressing AI-generated compliance data would provide a foundation for consistent national implementation.

6.2. Implications for financial institutions

The most immediate practical step for financial institutions deploying compliance AI is *type-specific testing*. Compliance AI systems should be evaluated not merely for aggregate accuracy but for performance on each hallucination type. This requires test sets that separately measure factual extraction accuracy (comparison against ground truth documents), numerical extraction accuracy (with specific attention to regulatory threshold boundaries), fabrication detection (whether the system generates content absent from source documents), and evidential grounding (whether the system distinguishes between documented facts and model inferences). Off-the-shelf LLM benchmarks do not measure these dimensions; financial institutions must develop compliance-domain-specific evaluation protocols.

Testing, however, only identifies the problem. The harder question is how to calibrate human oversight to the detectability gradient our framework describes. For Types F and N, document cross-referencing is effective and should be mandatory for high-risk onboarding cases. For Type B, oversight requires access to the authoritative regulatory corpus (FINMA Circulars, AMLA, implementing ordinances) and the ability to verify that cited provisions actually exist – this demands legal expertise, not merely compliance processing skill. For Type U, the most effective mitigation is not human review but system design: the AI system should be architecturally prevented from generating unsupported claims, for example by requiring explicit citation of source documents for every output element and flagging elements that lack source support.

The dual-track liability structure for Type B hallucinations creates a direct incentive for financial institutions to negotiate contractual provisions that allocate fabrication risk. Service level agreements should include fabrication rate guarantees, indemnification for compliance failures caused by fabricated outputs, and disclosure obligations requiring the provider to report known fabrication vulnerabilities. These contractual protections complement, but do not substitute for, the institution's own governance obligations.

6.3. Implications for AI providers

For AI providers, the most strategically important response is transparent type-specific disclosure. The EU AI Act's Art. 15 accuracy declaration should include type-specific hallucination metrics. Providers who report only aggregate accuracy are not merely failing a regulatory expectation – they are creating litigation risk. If a type-specific hallucination causes compliance harm and the provider's accuracy disclosure did not distinguish between hallucination types, a claim that the disclosure was misleading becomes straightforward. Transparent disclosure of fabrication rates – even if the rates are non-zero – is preferable to aggregate metrics that mask the fabrication risk.

A more structural response is to separate extraction from generation at the architectural level. The compliance AI use case involves two distinct capabilities: extracting information from source documents (a relatively bounded task) and generating assessments, summaries, or recommendations (an unbounded generative task). Fabricated and unsupported hallucinations arise primarily in the generative mode. Providers can reduce liability exposure by architecturally separating the extraction pipeline (which can be validated against ground truth) from any generative components (which carry inherently higher fabrication risk). A system that extracts and presents documented facts without generating interpretive text eliminates Types B and U by design, at the cost of reduced functionality.

The revised Product Liability Directive (Directive 2024/2853) and the proposed AI Liability Directive (COM(2022) 496) are not hypothetical threats – they create a credible pathway for compliance harm claims against AI providers that is already taking shape in EU legislative practice. Fabricated hallucinations are the primary liability vector: a system that invents regulatory citations produces a "defective" output in the product liability sense. Providers should maintain detailed records of hallucination testing, model validation, and known failure modes to support a defense that they exercised

reasonable care in product development. Failure to document these efforts will make the rebuttable presumption of the AI Liability Directive difficult to rebut.

7. Conclusion

This paper has demonstrated that the question "who pays when the AI hallucinates?" does not have a single answer. The legal consequences of an LLM hallucination in financial compliance depend fundamentally on the *type* of hallucination – a distinction that current regulation fails to make.

We developed a domain-specific hallucination taxonomy distinguishing four types: factual (F), numerical (N), fabricated (B), and unsupported (U) hallucinations. Each type has distinct characteristics in terms of detectability, evidentiary status, and the regulatory obligations it triggers. We constructed a mapping framework connecting these hallucination types to specific provisions of the EU AI Act, the Swiss FADP, FINMA supervisory expectations, and Swiss AML obligations. The framework reveals that deployer liability is primary across all types, that provider liability is activated primarily by fabricated hallucinations, that the effectiveness of human oversight degrades from Type F (high) to Type U (low), and that current accuracy requirements (Art. 15 EU AI Act) are insufficient because they do not require type-specific disclosure.

Three case scenarios illustrated the framework's practical application. A factual hallucination causing a missed PEP identification demonstrated dual-type liability compounding. A fabricated regulatory exemption demonstrated the unique provider liability pathway. A numerical hallucination suppressing a suspicious transaction report demonstrated the sharpest criminal liability exposure.

The practical implications are immediate. Financial institutions deploying compliance AI have a narrow window before the EU AI Act's August 2026 deadline for Annex III high-risk systems to implement hallucination-type-specific governance. "Accuracy-agnostic" compliance AI governance – treating all AI errors identically – is insufficient. Regulators should require type-specific accuracy disclosure. Financial institutions should implement type-calibrated human oversight. AI providers should separate extraction from generation and disclose fabrication rates.

Several avenues for further research emerge. Empirical measurement of hallucination rates by type in compliance AI deployments would ground the

theoretical framework developed here. Experimental studies of human oversight effectiveness across hallucination types would test our analytical predictions about graduated detectability. Comparative legal analysis of liability outcomes across jurisdictions – extending beyond the EU and Swiss frameworks examined here to include the UK (FCA), Singapore (MAS), and Hong Kong (HKMA) approaches – would test the framework's generalizability.

When the AI hallucinates, who pays depends not just on who deployed it, but on what kind of hallucination it produced.

Appendix A. Full Mapping Matrix

Table A.2 presents the complete mapping of hallucination types to compliance use cases, regulatory obligations, liable parties, enforcement mechanisms, and severity ratings. Severity is rated as: L (Low) – administrative remediation; M (Medium) – supervisory action and fines; H (High) – significant fines and enforcement proceedings; C (Critical) – criminal exposure or systemic compliance failure.

Table A.2: Complete hallucination type × compliance use case mapping matrix

Type	Use Case	Violated Provision	Liable Party	Enforcement	Sev.
F	KYC/KYB	AMLA Art. 3; AI Act Art. 15, 26; FADP Art. 6(5)	Deployer	FINMA supervisory action; FDPIC order	M
	PEP screening	AMLA Art. 6; AMLO-FINMA Art. 13; AI Act Art. 14	Deployer	FINMA enforcement; EDD remediation	H
	UBO ident.	AMLA Art. 4; LETA; AI Act Art. 15, 26	Deployer	FINMA action; LETA register correction	H
	STR filing	AMLA Art. 9; AI Act Art. 15	Deployer	MROS investigation; Art. 37 fine	H

Type	Use Case	Violated Provision	Liable Party	Enforcement	Sev.
N	KYC/KYB	AMLA Art. 3; AI Act Art. 15	Deployer	FINMA supervisory action	M
	PEP screening	AMLA Art. 6; AI Act Art. 9	Deployer	FINMA enforcement	M
	UBO ident.	AMLA Art. 4 (ownership %); AI Act Art. 15	Deployer	FINMA action; ownership recalculation	H
	STR filing	AMLA Art. 9, 37; AI Act Art. 15; FINMA Circ. 2011/1	Deployer	Criminal prosecution (Art. 37); MROS	C
B	KYC/KYB	AMLA Art. 3, 6; AI Act Art. 9, 14; PLD 2024/2853	Deployer + Provider	FINMA enforcement; product liability claim	H
	PEP screening	AMLA Art. 6; AI Act Art. 9; AILD proposal	Deployer + Provider	FINMA action; civil liability	H
	UBO ident.	AMLA Art. 4; AI Act Art. 9, 14; PLD 2024/2853	Deployer + Provider	FINMA action; product liability	H
	STR filing	AMLA Art. 9; AI Act Art. 9; PLD 2024/2853	Deployer + Provider	Criminal + civil proceedings	C
U	KYC/KYB	AMLA Art. 3; AI Act Art. 9, 15; FADP Art. 21	Deployer	FINMA action; FDPIC order	M
	PEP screening	AMLA Art. 6; AI Act Art. 9	Deployer	FINMA enforcement	H
	UBO ident.	AMLA Art. 4, 6; AI Act Art. 9, 15	Deployer	FINMA action; EDD remediation	H

Type	Use Case	Violated Provision	Liable Party	Enforcement	Sev.
	STR filing	AMLA Art. 9; AI Act Art. 9	Deployer	MROS investigation	M

Appendix B. Regulatory Comparison Table

Table B.3 provides a cross-jurisdictional comparison of AI regulation relevant to compliance hallucination liability. The comparison covers the European Union, Switzerland, the United Kingdom, and Singapore.

Table B.3: Cross-jurisdictional regulatory comparison for AI in financial compliance

Dimension	European Union	Switzerland	United Kingdom	Singapore
Primary legislation	AI Act (Reg. 2024/1689)	FADP (SR 235.1); AMLA (SR 955.0)	No AI-specific legislation (as of 2025)	No AI-specific legislation
Supervisory guidance	EBA Guidelines on AI in AML/CFT (2024)	FINMA Guidance 08/2024	FCA AI Update (March 2025); PRA SS1/23	MAS FEAT Principles (2022); MAS AI Guidelines (2024)
Risk classification	Mandatory high-risk classification (Annex III)	Principle-based ("fit for purpose")	Sector-specific, risk-based	Technology-neutral, outcomes-based
Hallucination-specific provisions	None (Art. 15 accuracy is generic)	None	None	None
Accuracy requirements	Art. 15: "appropriate level of accuracy"; must be declared	FINMA: "fit for purpose" validation	FCA: "appropriate validation"	MAS: FEAT fairness, ethics, accountability, transparency

Dimension	European Union	Switzerland	United Kingdom	Singapore
Human oversight	Art. 14: mandatory for high-risk	FINMA: expected, not prescribed	FCA: "meaningful human involvement"	MAS: human-in-the-loop recommended
Liability model	Deployer primary (Art. 26); provider secondary (Art. 16); PLD + AILD for civil claims	Deployer primary (AMLA/FINMA); provider via contract/tort law (OR Art. 41, 97)	Common law negligence; Consumer Rights Act 2015; no AI-specific liability	Common law; no AI-specific liability framework
Enforcement authority	National competent authorities; market surveillance	FINMA; FDPIC; cantonal prosecutors	FCA; ICO; PRA	MAS
Penalties (max.)	AI Act: EUR 35M or 7% turnover; PLD: uncapped damages	AMLA Art. 37: CHF 500K (intentional), CHF 150K (negligent); FINMA: license withdrawal	FCA: unlimited fines; criminal prosecution for AML failures	MAS: SGD 1M per breach; criminal prosecution
AML framework	AMLR (Reg. 2024/1624); 6AMLD	AMLA (SR 955.0); AMLO-FINMA	POCA 2002; MLR 2017; FCA Handbook	CDSA; MAS Notice PSN02
Timeline	Annex III obligations: Aug. 2026	Guidance 08/2024 effective immediately	FCA consultation ongoing (2025)	Guidelines effective 2024

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (12) (2023) 248:1–248:38. https://doi.org/10.1145/3571730

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, arXiv:2311.05232 (2023).

[3] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, I. Gabriel, Taxonomy of risks posed by language models, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT), ACM, 2022, pp. 214–229.

[4] V. Rawte, A. Sheth, A. Das, A survey of hallucination in "large" foundation models, arXiv:2309.05922 (2023).

[5] S.M.T.I. Tonmoy, S.M.M. Zaman, V. Jain, A. Rani, V.S. Rawber, A. Chadha, A. Das, A comprehensive survey of hallucination mitigation techniques in large language models, arXiv:2401.01313 (2024).

[6] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L 2024/1689.

[7] Federal Act on Data Protection (FADP), SR 235.1, as revised 25 September 2020, in force since 1 September 2023.

[8] Swiss Financial Market Supervisory Authority (FINMA), Guidance 08/2024: Supervisory expectations for the use of artificial intelligence by supervised institutions, 2024.

[9] Federal Act on Combating Money Laundering and Terrorist Financing (Anti-Money Laundering Act, AMLA), SR 955.0.

[10] Financial Action Task Force (FATF), International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation: The FATF Recommendations, updated October 2023.

[11] Financial Action Task Force (FATF), Opportunities and Challenges of New Technologies for AML/CFT, 2024.

[12] P. Hacker, The European AI liability directives – Critique of a half-hearted approach and lessons for the future, Frontiers in Artificial Intelligence 6 (2023) 1217192.

[13] D.W. Arner, J. Barberis, R.P. Buckley, FinTech, RegTech, and the reconceptualization of financial regulation, Northwestern Journal of International Law & Business 37 (3) (2017) 371–413.

[14] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC (revised Product Liability Directive), OJ L 2024/2853.

[15] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final.

[16] European Banking Authority (EBA), Report on the use of machine learning for internal ratings-based models and anti-money laundering, EBA/REP/2024/02.

[17] McKinsey & Company, The future of KYC: How technology is reshaping compliance, 2023.