

Evaluating the Reliability of LLM-Based Document Intelligence for KYC/AML: An Evaluation Framework for an AI-Native Compliance Platform

Joerg Osterrieder^{a,*}, Lennart Baals^a, Illia Kosterin^a, Didier Baertschiger^b, Nathan Douet^b, Vincent Pignon^b

^aUniversity of Applied Sciences of the Grisons (FHGR), Chur, Switzerland

^bWeCanGroup SA, Geneva, Switzerland

Abstract

Large language models are being deployed into client onboarding and anti-money-laundering workflows at regulated institutions, private banking among them, where they extract structured facts from identity papers, ownership charts and source-of-wealth narratives. Whether such a system is accurate is measurable and well studied. Whether it is *reliable enough to be trusted inside a regulated process* is a different question, and it is the one an institution has to answer. **This is a position paper: it proposes an evaluation framework and reports no empirical results.**

Aggregate extraction accuracy is the wrong primary criterion for a deployment decision. It averages over error types whose consequences differ by orders of magnitude, and it is silent on the property that threshold-routed human review depends on, namely whether confidence scores are calibrated and sharp. We develop a framework around four dimensions that practice conflates: extraction accuracy, confidence calibration, hallucination as a distinct failure class, and auditability as a property of the retained record, specifying for each what to measure, on what sampling design and against what acceptance criteria. We give particular attention to a stratum routine monitoring cannot observe, the cases the system did *not* escalate, and relate the framework to conformal prediction, whose guarantees are marginal where ours must hold per stratum.

We then map the dimensions onto the duties they serve under Swiss supervisory, data-protection and anti-money-laundering instruments, and show that the EU AI Act's high-risk regime does not reach this class of system on the present text of Annex III. That absence widens the case for a voluntary framework: the instruments that do bind say what must be managed without saying what must be measured. The framework is developed against a deployed platform, described qualitatively; the study it specifies is planned.

Keywords: LLM reliability, confidence calibration, KYC, anti-money laundering, document intelligence, selective prediction, conformal prediction, EU AI Act, FINMA, RegTech

1. Introduction and related work

1.1. The setting

Opening a private-banking relationship generates documents: identity papers, proofs of address, corporate ownership charts, beneficial-ownership declarations, tax forms and narrative accounts of the origin of wealth, in several languages and several formats, many of them scans of originals. Somebody must read each one, extract the facts that matter, check those facts against each other, screen the names involved against sanctions and politically-exposed-person lists, and record the result in a form that will still be intelligible to a supervisor years later.

That work is expensive, and it is the obvious target for large language models (LLMs), which are good at reading heterogeneous documents and emitting structured fields.

A number of vendors and institutions have deployed such systems. The supervisory response has followed: FINMA issued dedicated guidance on the governance and risk management of artificial intelligence by supervised institutions (FINMA, 2024), and a regulator does not write guidance for a technology its supervised population is not using.

What is in doubt is how an institution establishes that one of these systems is good enough to rely on. This is not a rhetorical question. Compliance is a domain where the cost of error is sharply asymmetric, since a missed field costs an email while a missed beneficial owner can cost a regulatory finding, and where the obligation is continuous rather than discharged at account opening. An evaluation regime designed for a benchmark leaderboard does not transfer without modification.

1.2. The gap this paper addresses

The literature offers a great deal on the components and comparatively little on the composition. There are surveys of hallucination in language models (Huang et al., 2024; Ji et al., 2023), a mature line on the calibration of neural

*Corresponding author

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

network confidence (Desai & Durrett, 2020; Guo, Pleiss, Sun, & Weinberger, 2017; Naeini, Cooper, & Hauskrecht, 2015), foundational work on selective prediction and abstention (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017), and work on measuring whether a generated statement is attributable to its source (Honovich et al., 2022; Maynez, Narayan, Bohnet, & McDonald, 2020; Rashkin et al., 2023). Financial-domain benchmarks exist (Shah et al., 2022; Xie et al., 2024), and industry bodies have begun to argue that model-level benchmarks miss what matters in financial contexts: FINOS holds that traditional model-level benchmarks “fall short of capturing what matters most in financial contexts” and that, because real deployments combine AI and non-AI components, “testing an LLM alone is insufficient” (Borella, 2025). We agree, and Section 3.6 gives that argument a concrete measurement consequence.

There is also no shortage of AI governance frameworks. The NIST AI Risk Management Framework (2023) organises risk into govern, map, measure and manage functions; ISO/IEC 42001 (2023) specifies a management system for AI; and the documentation tradition of model cards (Mitchell et al., 2019) and datasheets (Gebru et al., 2021) standardises what a system discloses about itself. These are complementary to what follows and we do not propose to replace them. The distinction is one of altitude and of specificity: they say that an organisation must measure, and establish who owns the measuring and how it is recorded; they do not say *which* quantities decide whether an LLM extraction pipeline is fit for a KYC file, nor in what relation those quantities stand to one another. NIST’s “measure” function is a slot, and this paper is a candidate filling of it for one domain.

What is missing, then, is a framework at that lower altitude: one that tells a deploying institution which measurements it must make *before* it puts such a system into a regulated process, in what relationship those measurements stand, and what regulatory duty each one discharges.

Our reading of the gap’s practical consequence is that evaluation defaults to whatever is easiest to report, and that this is aggregate extraction accuracy, the one number least informative about deployment risk. We offer that as an observation rather than a finding, and Section 6.3 sets out the content analysis of vendor documentation that would put it on an evidential footing.

Adjacent to this, document information extraction has its own evaluation literature and its own benchmarks for forms and receipts (Jaume, Ekenel, & Thiran, 2019; Park et al., 2019). Those measure extraction quality on document images and are a necessary input here. They do not address calibration, abstention, or the auditability of the resulting record, which are the three things that decide whether an extraction pipeline can be relied on inside a regulated decision.

1.3. Contributions

1. An evaluation framework for LLM-based document intelligence in regulated onboarding, organised around four dimensions that practice conflates: extraction accuracy, confidence calibration, hallucination as a distinct failure class, and auditability of the retained record (Section 3).
2. **The requirement to audit the cases the system did *not* escalate.** Errors in the escalated stratum are visible because somebody looks at them; errors in the auto-approved stratum are invisible for the same reason, and that is the stratum where an institution has chosen to accept risk without review. Operational monitoring reports improving quality as the threshold is lowered, because the cases it can see are the ones it still reviews. Every incentive points away from this measurement and it is the one that establishes whether the threshold is right (Section 4.2).
3. A concrete evaluation protocol built around that requirement, covering metrics, stratified sampling design, and acceptance criteria stated in advance (Section 4).
4. The argument that a deployment decision rests on the risk-coverage curve rather than on aggregate accuracy, so a system with higher accuracy but uninformative confidence can be the worse choice. We transfer this rather than derive it: as a general result it belongs to the selective-prediction literature (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Two things are specific to this domain. The first is the practice the result displaces, which is the reporting of a single aggregate accuracy figure. The second is the corollary that calibration must hold *per stratum*, because a globally calibrated system can be badly miscalibrated on scanned documents alone, and one global threshold then under-escalates exactly the population needing more review (Section 3.2).
5. A mapping from the four dimensions onto the duties they serve under FINMA Guidance 08/2024, the revised FADP and Swiss AMLA, with the argument that several duties which read as documentation requirements are in substance evaluation requirements. The mapping includes a negative result that we have not seen stated: on the present text of Annex III, the EU AI Act’s high-risk regime does not reach KYC and anti-money-laundering document intelligence, so the obligation most often assumed to govern these systems does not (Section 5).
6. A research agenda, stating the problems the framework surfaces without solving, from deriving acceptance criteria to measuring cross-document coherence, together with the paper’s own evidential position and the provenance discipline under which it was written (Section 6).

1.4. What this paper is not

It is a position paper. It proposes what should be measured and does not report measurements. The framework is developed against a deployed platform which is described qualitatively in Section 2, and the empirical study specified in Section 4 is planned rather than complete. We state this here, in the abstract, and again in Section 6.4, because the failure mode for a paper of this kind is to be read as reporting the evaluation it in fact only designs.

It is also not a benchmark proposal. A shared community benchmark for compliance AI, meaning task definitions, public datasets and baselines that other groups report against, is a different contribution with a different unit of analysis, and it is the subject of separate work. The distinction is that this paper asks whether *one* deployed system is fit for use in a regulated process, and a benchmark asks what the field as a whole should measure against.

2. System overview: an AI-native compliance platform

The framework in this paper is developed against a deployed system rather than a hypothetical one. This section describes that system at the level of architecture and design commitment. It reports no performance figures; see Section 6.4 for why.

2.1. Design posture

The platform under study is operated by a Swiss RegTech firm and used by private banks and independent asset managers. Its distinguishing design choice is that the compliance workflow was built around the language model rather than the model being added to an existing workflow. Two consequences matter for evaluation.

First, the platform is *model-agnostic*: the LLM is a replaceable component rather than the architecture. This is commercially sensible and it has an awkward evaluation implication, developed in Section 3.4: any reliability claim is a claim about a configuration at a point in time, and the configuration is expected to change.

Second, the design concedes at the outset that the model will sometimes be wrong. Rather than targeting an accuracy threshold, the workflow attaches a confidence score to each extracted field and routes cases to human review according to configurable thresholds. The system’s objective is not to be right; it is to be right where it is confident and to escalate where it is not.

We regard that as the right *design pattern*, and the endorsement is of the pattern rather than of any implementation of it. Three of the present authors are employed by the operator, so it is worth being exact about what is being praised and what follows from it: conceding fallibility and routing on confidence is the correct response to an asymmetric error cost, and it also relocates the evaluation burden onto a quantity the operator has not yet published. A threshold-routing architecture is only as good as

the confidence scores it routes on. If those scores are uninformative, the routing is arbitrary and the human review it feeds is ceremonial. Accuracy does not tell you whether that is the case. Calibration does.

2.2. Pipeline stages

The operator states that the platform is hosted in Switzerland and that it holds ISO 27001 certification. For the institutions concerned, data residency is a procurement requirement rather than a preference, which is why hosting is treated as a design principle. Neither property bears on reliability, and we mention them because they are frequently offered as though they did. An information-security certification is a statement about controls within a defined scope, not about whether a model’s confidence scores mean anything, and the scope boundary is worth reading before the certificate is relied upon.

2.3. Why the coherence stage deserves separate attention

Of the automated checks, coherence is the one with the most evaluation content and the least literature. Completeness is close to a schema check. Plausibility is a per-field validity question. Coherence is cross-document: the nationality on the passport, the address on the utility bill, the shareholding in the corporate chart and the narrative in the source-of-wealth statement must tell one story, and a system that extracts each document correctly in isolation can still fail to notice that they do not.

This matters for framework design because cross-document consistency is a property of a *file*, not of an extraction, and it therefore requires a unit of evaluation larger than the field. We return to this in Section 3.5.

2.4. Deployment settings, and what each demands of an evaluation

The five settings below are composites, assembled from the platform described above and from the arrangements a regulated onboarding process presents. No institution is identified and none of the five carries a measurement. Each is stated in terms of the duty that governs it, so that an institution outside Swiss private banking meets its own case in one of them. Their purpose is methodological: the four dimensions of Section 3 are not equally stressed in every setting, and an evaluation designed against one of them can be silent on the failure mode of another.

Setting A: onboarding through an intermediated relationship. An independent asset manager assembles a client file and submits it to a custodian bank, which must satisfy its own identification duties on documents it did not collect. Documents originate with three parties, arrive asynchronously, and are superseded before the file closes. The unit of evaluation is the client file (Section 3.5), so the burden falls on the coherence stage of Section 2.3. The setting also makes the propagation problem of Section 3.6 concrete: the institution carrying the regulatory duty must

Stage	Function
Document intake	Heterogeneous formats and languages, including photographed and scanned originals of varying quality
Extraction	Structured fields recovered from unstructured documents, with a confidence score attached per extraction
Automated checks	Completeness (is anything missing), plausibility (is the value credible), coherence (does it agree with the other documents in the file)
Screening	Sanctions and politically-exposed-person matching against third-party data providers
Risk scoring	A risk classification for the relationship, derived from the extracted and screened facts
Human-in-the-loop	Routing to a compliance officer according to configurable confidence thresholds
Record	Retention of the material needed to reconstruct the decision

Table 1: Pipeline stages. Evaluation attaches at every stage, not only at extraction; error introduced early propagates and is attributed late (Section 3.6).

Setting	What it presents	Where the evaluation load falls
A. Intermediated onboarding	A file assembled by one firm and relied on by another, from three sources, arriving asynchronously	Coherence, on the file as the unit (Sections 2.3 and 3.5); attribution across an organisational boundary (Section 3.6)
B. Layered ownership	Corporate charts, trust instruments and register extracts under several jurisdictions and conventions	Unsupported and fabricated output on screening fields (Section 3.3, criterion 3 of Section 4.4); the stratum taxonomy of open problem 2
C. Heterogeneous document quality	Photographed originals, languages beyond the institution’s working ones, scripts other than Latin	Per-stratum calibration (Sections 3.2 and 3.7) and thin-cell over-sampling (Section 4.1)
D. Continuous review	Documents arriving out of sequence against a record years old, under a configuration since replaced	Auditability and record completeness (Section 3.4); the re-measurement trigger (Section 5.6, open problem 7)
E. Institution-configured thresholds	Thresholds set per client segment and document type by the deploying institution	The non-escalated audit, once per configuration (Section 4.2); review capacity, criterion 5 of Section 4.4

Table 2: Five deployment settings and the dimension each one loads. The settings are composites and carry no measurements. The third column is a judgement about where an evaluation designed against a different setting would be silent.

attribute an end-to-end error to a stage it does not itself operate.

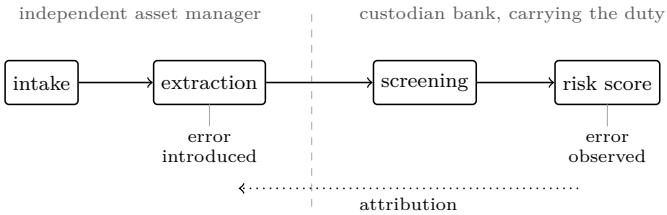


Figure 1: Setting A, schematic. The stages are those of Table 1. An extraction error enters the file on one side of the boundary and is observed as a risk classification on the other, so the institution that must answer for the classification has to attribute it to a stage it does not operate. Nothing in the picture is measured.

Setting B: layered ownership. Reaching the beneficial owner means traversing a corporate chart, trust instruments and register extracts issued under several jurisdictions and conventions. This is the stratum in which a fabricated or unsupported output is at once most consequential and least detectable by inspection, since an invented intermediate entity is exactly as plausible as a real one. It is the setting that motivates criterion 3 of Section 4.4, an

unsupported-output rate held at or near zero on the fields that feed screening, and it is the hard case for the stratum taxonomy of open problem 2, where a trust with layered ownership and a single natural person are not comparable units.

Setting C: heterogeneous document quality. Intake carries photographed originals, documents in languages other than the institution’s working languages, and scripts other than Latin. These are the thin cells a proportional sample cannot speak to and pooled reporting conceals, which is why the calibration requirement of Section 3.2 attaches per stratum, with the consequence Section 3.7 draws, and why Section 4.1 over-samples them deliberately.

Setting D: continuous review. Where an institution replaces periodic review campaigns with continuous re-verification, documents arrive out of sequence and are assessed against a record established years earlier under a configuration since replaced. The load falls on auditability (Section 3.4), and the retention requirement is set by the review that will read the record years later. Which observable should fire the re-measurement that Section 5.6 argues for is open problem 7.

Setting E: institution-configured thresholds.. Escalation thresholds are set by the deploying institution, and one institution may hold different thresholds for different client segments and document types. A reliability finding is therefore a statement about a configuration (Section 5.6), so the non-escalated audit of Section 4.2 runs once per configuration. This is also the setting in which criterion 5 of Section 4.4, the review-capacity constraint, binds in practice, and in which its tension with criterion 1 is decided.

None of the five is answered by an aggregate accuracy figure. What they fix is the population over which the protocol of Section 4 has to be executed for its result to carry.

3. The framework: four dimensions of reliability

We propose four dimensions. The claim is not that each is individually novel, since calibration and attribution both have substantial literatures. It is that the four are jointly necessary, that reporting practice as we have encountered it foregrounds the first alone, and that the relationships between them determine what a deployment decision can rest on.

3.1. Dimension 1: extraction accuracy, necessary and insufficient

Extraction accuracy answers: does the extracted value match the document? It should be measured, and measured better than it usually is.

Metrics.. Field-level precision, recall and F_1 , reported *per field type* rather than pooled. Exact match and normalised match reported separately, since a date rendered in a different format is not the same class of event as a wrong date. For fields with structure, such as names, addresses and ownership percentages, a distance-based partial-credit measure alongside exact match, because a one-character difference in a name and a different name entirely are not the same error for screening purposes.

Stratification.. Aggregate accuracy over a mixed corpus is close to uninterpretable. At minimum, results should be stratified by document modality (born-digital versus scanned versus photographed), by language, by document type, and by structural complexity of the underlying entity. The reason is not statistical fastidiousness: it is that these strata correspond to client populations, and an accuracy difference between them is a difference in the friction that different clients experience (Section 3.7).

Why it is insufficient.. An aggregate accuracy figure averages over error types whose costs differ by orders of magnitude. A misread postcode and a missed beneficial owner are one event each. More fundamentally, accuracy is a property of the *output* and says nothing about whether the system can identify its own likely failures, which in a threshold-routing architecture is the property the design actually depends on.

3.2. Dimension 2: confidence calibration, and the curve it licenses

A confidence score is calibrated if, among the extractions to which the system assigns confidence p , the proportion that are correct is approximately p . Formally, for a predictor producing $(\hat{y}, \hat{p})$ pairs, calibration requires

$$\mathbb{P}(\hat{y} = y \mid \hat{p} = p) = p \quad \text{for all } p \in [0, 1].$$

Metrics.. Expected calibration error (ECE) and maximum calibration error, computed over binned confidences (Guo et al., 2017; Naeini et al., 2015); reliability diagrams, which should be inspected and not only summarised, since ECE can be low while the curve is badly shaped in the region where the operating threshold sits; and the Brier score (Brier, 1950) as a proper scoring rule combining calibration and sharpness.

The operational form: risk-coverage.. For a deployment decision, the informative object is the risk-coverage curve familiar from selective prediction (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017): for each escalation threshold τ , the proportion of cases handled automatically (coverage) plotted against the error rate among them (risk). This curve, not an accuracy number, is what a threshold decision is actually made against, and it can be computed from the same labelled sample.

Relation to distribution-free methods.. The formal machinery for turning a score into a guarantee is conformal prediction (Shafer & Vovk, 2008), which converts any scoring function into prediction sets that contain the truth at a chosen rate, with finite-sample validity and no distributional assumption beyond exchangeability. Where a threshold is the thing an institution has to defend to a supervisor, a coverage guarantee is a stronger object than a calibration statistic, and an evaluation of the kind proposed here should report one wherever it can be computed.

Exchangeability is also where this domain pushes back, and the way it does so is specific enough to be worth stating. First, conformal validity is *marginal*: it holds on average over the calibration distribution. A system can be conformally valid at a stated rate overall while systematically under-covering photographed documents in a minority language, because those cells are a small share of the average, which is exactly the failure Section 3.7 is about. Second, onboarding populations shift continuously as clients from new jurisdictions arrive and document formats change; the weighted extension for covariate shift (Tibshirani, Foygel Barber, Candès, & Ramdas, 2019) requires the shift to be known or estimable, which for a live client book it generally is not.

Conformal prediction and the framework proposed here therefore compose rather than compete, and the composition has a shape worth naming: conformal methods supply the guarantee, stratification decides where the guarantee has to hold, and the non-escalated audit of Section 4.2

is what checks that it does. The same reading applies to learning to defer (Mozannar & Sontag, 2020), which trains the routing rule jointly with the predictor instead of thresholding a score after the fact. That is the better design. It is not the design deployed here, nor in the compliance products we are aware of, and a framework for evaluating what institutions have already bought has to evaluate threshold routing while noting that a jointly trained deferral rule would change the measurement question rather than retire it.

Why the curve, and not accuracy, decides a deployment.. Consider two systems, one more accurate in aggregate but with confidence scores uncorrelated with correctness, the other less accurate but with informative confidence. On the first, no threshold is available that isolates the likely errors: the institution must review everything or nothing, and the human-in-the-loop stage delivers cost without safety. On the second, a threshold exists that concentrates the errors into the escalated stratum. The less accurate system is the better deployment.

The argument is easier to see drawn. Figure 2 is schematic and the quantities in its caption are **invented for illustration**: they are not measurements of any system, and no claim in this paper rests on them.

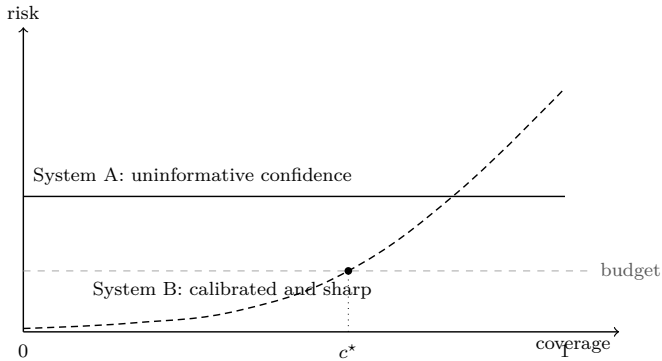


Figure 2: Risk against coverage for two systems, schematic. System A is the more accurate model, with a lower base rate, but its confidence carries no information about correctness, so any threshold partitions its errors in the same proportion as its successes and its curve is flat. No operating point holds an error budget below its base rate, so the only way to meet the budget is to review everything. System B is less accurate overall and wrong *predictably*, so its curve rises only as coverage grows and it meets the same budget at c^* . On invented figures: A at 96 percent aggregate accuracy, B at 92 percent, a budget of one error in a hundred, and c^* at about three fifths coverage.

System A is the better model and the worse deployment. Because its confidence carries no information about correctness, any threshold partitions its errors in the same proportion as its successes, so the only way to hold an error budget below its base rate is to review everything, and reviewing everything is the state the system was bought to escape. System B is wrong more often in total and wrong

predictably, so most of its output can be trusted and the reviewer’s time goes where the doubt is.

Calibration is necessary and not sufficient: the other half is sharpness.. Precision matters here, because the shorthand “calibration beats accuracy” is not quite the claim and a reader who knows the area will notice. A system that assigns confidence $p = 0.92$ to every extraction is *perfectly calibrated*, in the sense that it is right 92 percent of the time and says so, and it supports no useful threshold at all, because every case looks identical. In deployment terms it is System A.

The two properties do different work:

- **Calibration** makes a stated error budget mean what it says. It is what licenses the *threshold*: without it, “review everything below 0.9” does not correspond to any particular error rate.
- **Sharpness**, meaning a confidence distribution that is spread out rather than concentrated, is what makes the coverage at that budget *large*. It is what licenses the *automation*.

The object that carries both at once is the risk–coverage curve introduced above, which is why we treat it as the operational form rather than as a presentational convenience. The precise version of this paper’s claim is therefore that **the risk–coverage curve dominates aggregate accuracy as a deployment criterion**, with calibration the property that makes the curve transferable to a threshold decision and sharpness the property that makes the resulting automation worth having. A proper scoring rule such as the Brier score (Brier, 1950) decomposes into a calibration term and a resolution term, plus an irreducible uncertainty term that depends only on the base rate. That is why Section 4 asks for it alongside ECE rather than instead of it: the Brier score is sensitive to both properties at once, where ECE is sensitive to one.

We should also be precise about the strength of the claim, because it is easy to overstate. Formally it is an existence result and follows from the selective-prediction literature (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017): there exist pairs of systems where the less accurate calibrated one is preferable. What makes it a *practical* claim rather than a mathematical curiosity is the empirical premise that real deployed systems occupy the region where the gap is wide. That is exactly what Section 4 sets out to measure, and what Section 4.5 would accept as refutation. We assert the general form and offer the domain-specific form as the hypothesis under test.

This inversion is the most consequential thing this dimension has to say, because it contradicts the metric that reporting practice defaults to. It also has a shelf life. Pre-trained transformers are known to be miscalibrated in ways that shift under distribution change (Desai & Durrett, 2020), so a system calibrated on this quarter’s document mix is not thereby calibrated on the next one’s, and Section 5.6 draws the governance consequence.

A caveat we want to record.. Self-reported confidence from a generative model is not automatically a probability. Whether LLMs’ expressed confidences are informative is contested and task-dependent (Kadavath et al., 2022). An evaluation must therefore establish that the score being routed on is calibrated *for this system on this document population*, rather than assuming a property demonstrated elsewhere.

3.3. Dimension 3: hallucination as a distinct failure class

An extraction error returns the wrong value from the document. A hallucination returns a value that is not in the document at all: a beneficial owner who does not appear in the chart, a contractual clause that was never written. Treating the second as a subtype of the first is a category error with operational consequences, because the ordinary human check, whether the value looks reasonable, is precisely the check a hallucination passes.

Taxonomy.. We adopt the four-type distinction developed for financial compliance in prior work from this project (Osterrieder & Baals, 2026): *factual* (a wrong value where a right one exists), *numerical* (a distorted quantity), *fabricated* (an entity or document that does not exist), and *unsupported* (a plausible inference not grounded in the source). The types differ in detectability and in which regulatory obligation they threaten, which is why a single hallucination rate is not a useful quantity.

Metrics.. Groundedness or attribution: every extracted value should be traceable to a specific span in a specific source document, and the rate of outputs lacking such a span is measurable directly (Rashkin et al., 2023). Faithfulness measures developed for summarisation transfer with modification (Honovich et al., 2022; Maynez et al., 2020). The unsupported-output rate must be reported separately from the error rate, since an unsupported output can be correct by coincidence and is still a control failure: the system had no basis for it.

A design implication.. Span-level attribution is not only a measurement device. A pipeline that requires every field to carry a source span makes fabrication detectable at inference time rather than only at audit, and it supplies exactly the material that Section 3.4 requires the record to retain. Measurability and auditability point at the same architectural choice, which is a good sign for both.

3.4. Dimension 4: auditability as a property of the record

A supervisor asking in 2029 why a relationship was classified low-risk in 2026 needs an answer in 2029. The requirement therefore lands on the retained record rather than on the model, which by then will have been replaced, possibly more than once.

What the record must support.. Reconstruction of the decision: the input document as received, the extracted values, the confidence score for each, the threshold in force at that time, the model and prompt configuration identifier, whether a human reviewed the case, and what they changed. Anything less does not permit reconstruction, and a record that does not permit reconstruction does not discharge the obligation however voluminous it is.

The measurable properties.. Record completeness, the proportion of decisions for which every element above is present and resolvable, is directly auditable. We expect it to be the element most often assumed rather than checked, and Section 4 treats that as a prediction to test rather than as an observation. Reproducibility under component replacement is the harder one: if the model was a third-party API version since retired, the 2026 decision may not be reproducible at all, in which case the institution should establish deliberately what its record then evidences: that a decision was made on a stated basis, rather than that it would be made again.

Interaction with the model-agnostic design.. The platform’s replaceable-model posture (Section 2.1) makes a configuration identifier in the record not a nicety but the thing that makes any historical reliability claim meaningful. It also implies that reliability evidence has a shelf life, and that re-evaluation on component change is part of the control, not a project.

3.5. Units of evaluation

The four dimensions do not share a natural unit, and conflating units is a common source of misleading reporting.

3.6. Error propagation across stages

An extraction error does not stay where it happened. A misread nationality flows into sanctions screening and into risk scoring, so a data-quality problem surfaces as a compliance outcome several stages downstream, where nobody attributes it to extraction. The weakest stage therefore bounds the accuracy of everything after it, and stage-local evaluation systematically underestimates end-to-end risk.

The framework consequence is that at least one end-to-end measurement is required, and it cannot be inferred from per-stage figures: the rate at which a final risk classification is wrong, and the attribution of those cases to the stage that caused them.

3.7. Data quality, strata and their distributional consequences

Compliance corpora are heterogeneous in ways that correlate with client populations: document modality, language, and the structural conventions of the jurisdiction of incorporation. If extraction is weaker on photographed documents, or on one language, the operational effect is

Unit	Applies to	Why this unit
Field	accuracy, calibration	The extraction is the atomic prediction and the thing a confidence score attaches to
Document	hallucination, attribution	Groundedness is a relation between an output and a source, so the source is part of the unit
File (client)	coherence, risk score	Cross-document consistency is undefined below the file, and the risk classification is per relationship
Decision	auditability	The reconstructable object is the decision, including the human action taken on it

Table 3: Evaluation units. A pooled “accuracy” figure that mixes fields across files of differing complexity is not wrong so much as uninterpretable.

that clients presenting those documents are escalated more often and experience more friction.

We raise this as an evaluation requirement rather than an ethical aside. Nobody designs that outcome; it emerges from a per-stratum accuracy difference that pooled reporting hides. Measuring per stratum is the only way it becomes visible, which is the same argument as Section 3.1 arriving from a different direction.

4. Planned evaluation and error analysis

This section specifies the study the framework implies. **It is a protocol, not a report.** No results are presented here, and none exist at the time of writing. We state the design in advance partly because that is good practice and partly because a pre-registered protocol is harder to reshape around a result one would prefer.

4.1. Corpus and sampling design

The intended evidence base is an annotated corpus of real onboarding documents, drawn from the deployed platform’s document population and annotated to a written guideline with measured inter-annotator agreement. The target size recorded in the project’s own deliverable plan is on the order of 300 documents; that figure is a plan and is not measured evidence, and no claim in this paper rests on it.

Sampling must be stratified rather than random, on the axes of Section 3.1, and deliberately over-sampled in the thin strata. A proportional sample of a real document population yields too few photographed documents in a minority language to say anything about them, which is precisely the cell where the risk concentrates.

4.2. The stratum routine monitoring cannot see

This is the methodological point we would most want a reader to take away.

In a threshold-routing deployment, errors in the *escalated* pile are visible by construction: a human looks at each one, and disagreements are recorded. Errors in the *auto-approved* pile are invisible by construction, because nothing looks at them. Since the auto-approved pile is where an institution has chosen to accept risk without review, it is also where an under-set threshold does its damage. Operational monitoring will report improving quality

as the threshold is lowered, because the cases it can see are the ones it still reviews.

Any credible evaluation must therefore include a deliberate audit of a sample of non-escalated cases, drawn independently of the confidence score, and re-adjudicated blind. This is the measurement that establishes whether the escalation threshold is correctly placed. It is also the measurement most easily omitted, because it costs reviewer time on cases the system said were fine and it is the one nobody’s dashboard demands.

4.3. Measurement plan

4.4. Acceptance criteria, and stating them in advance

A framework that specifies measurements without acceptance criteria leaves the deployment decision exactly where it was. We propose that an institution fix, before measuring:

1. A maximum tolerable error rate *in the non-escalated stratum*, per field type, derived from the consequence of that field being wrong, and not a single global number, because the consequences are not comparable.
2. A maximum calibration error in the confidence region where the operating threshold sits. Calibration far from the threshold is of academic interest only.
3. A maximum unsupported-output rate, set at or near zero for fields that feed screening, since an ungrounded name is a control failure irrespective of whether it happens to be right.
4. A minimum record-completeness rate, which we would argue should be one: a decision that cannot be reconstructed has not been recorded.
5. A review-capacity constraint, since a threshold that escalates more cases than the institution can adjudicate is not a control but a backlog.

Criteria 1 and 5 interact and will not both be satisfiable at every threshold. That tension is the real deployment decision, and making it explicit is more useful than any single figure of merit.

Dimension	Unit	Planned measurements
Accuracy	field	Per-field-type precision, recall, F_1 ; exact and normalised match reported separately; partial credit for structured fields; all stratified
Calibration	field	ECE and maximum calibration error; reliability diagrams inspected, not only summarised; Brier score; risk–coverage curve; calibration recomputed <i>per stratum</i>
Hallucination	document	Rate of outputs with no source span; unsupported-output rate reported separately from error rate; incidence by the four-type taxonomy
Coherence	file	Rate of undetected cross-document inconsistency, measured against adjudicated files
Auditability	decision	Record completeness against the element list of Section 3.4; reproducibility under a simulated component change
End-to-end	file	Rate of incorrect final risk classification, with attribution to the causing stage
Threshold placement	field	Error rate in an independently drawn, blind-adjudicated sample of <i>non-escalated</i> cases

Table 4: Planned measurements. Every row is a commitment, not a result.

4.5. What would falsify the framework’s central claim

Our central claim is that the risk–coverage curve dominates aggregate accuracy as a deployment criterion in this setting. It would be substantially weakened if, on the planned corpus, the risk–coverage curves of a higher-accuracy configuration with uninformative confidence and a lower-accuracy configuration with informative confidence turned out to be close. That is the case in which the curve buys no usable threshold advantage over the accuracy figure, and choosing on accuracy alone would cost an institution nothing. We regard it as unlikely but not impossible, and we would report it.

We state the falsification condition because a framework paper that cannot be wrong is not offering much.

5. Regulatory and governance implications

The four dimensions are not only good practice. Each serves identifiable duties, and this section makes two arguments about them.

The first is negative, and it comes first because it removes the instrument a reader will reach for: on the present text, the EU AI Act’s high-risk regime does not reach the system under study. The second is that among the instruments that do apply, several requirements which read as documentation duties are in substance evaluation duties, because they cannot be satisfied by a document describing a system, only by measurements of it.

5.1. The EU AI Act, and why its high-risk regime does not settle the question

It is tempting to reach first for Regulation (EU) 2024/1689 (EU AI Act, 2024), and the reach does not succeed. Annex III, which fixes the high-risk list, covers the evaluation of the creditworthiness of natural persons and the pricing of life and health insurance, and it excludes from the former systems used to detect financial fraud. Anti-money-laundering work, client identification, sanctions screening and beneficial-ownership determination do not appear in Annex III at all. The obligations usually cited in this setting are scoped to match: Article 15(1)

opens “High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy”, and Article 15(3) requires that “the levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use”. Articles 9, 11 and 14 carry the same qualifier.

A KYC extraction pipeline that routes uncertain cases to a compliance officer is therefore, on the present text, unlikely to be a high-risk AI system, and the duties in Articles 9 to 15 are unlikely to attach to it. The omission looks deliberate rather than accidental, because the drafters carved fraud detection out of the creditworthiness entry explicitly. Financial-crime applications were evidently considered and left off the list.

Four routes remain by which the regime can still bite, and an institution should check each of them rather than assume the answer. The pipeline may feed a creditworthiness decision, which is listed. Article 6(3) allows a system listed in Annex III to fall outside the high-risk class only where it does not pose a significant risk of harm, and Article 6(4) then requires a provider who reaches that conclusion to “document its assessment before that system is placed on the market or put into service”, so the conclusion is itself a compliance artifact rather than a default. Article 7, headed *Amendments to Annex III*, empowers the Commission to adopt delegated acts adding or modifying high-risk use cases, so today’s absence is not permanent. And an institution, or its procurement function, may adopt the high-risk controls contractually, which is common where a vendor sells across jurisdictions.

The more useful observation is what follows from the absence. The most obvious regulatory hook does not apply, and the duties that do apply to a Swiss institution, set out in Sections 5.2 to 5.4, say what must be managed without saying what must be measured. No instrument tells this institution that calibration has to be reported *per stratum*, or that the non-escalated stratum has to be audited. That gap is the reason a voluntary framework is worth specifying, and it is wider than it would be if Article 15(3) applied. Where the Act does bind, it at least forces a provider to choose an accuracy metric and publish

it, and the choice is substantive: an aggregate figure that averages over strata and error types is declarable and close to uninformative. Read as a model of good practice rather than as a binding duty, Article 15(3) is still the most useful sentence in the Act for present purposes, and Article 14 on human oversight is still where calibration would enter, because oversight that cannot identify which cases warrant attention is not effective oversight.

5.2. FINMA Guidance 08/2024

FINMA’s guidance on governance and risk management for artificial intelligence (FINMA, 2024), published in December 2024, identifies model risk with attention to robustness, correctness, explainability and bias, data risk including data quality, and increasing third-party dependencies.

Its status needs stating precisely, because the argument of this section depends on being exact about which instruments bind. A FINMA Guidance is not a Circular or an Ordinance. It communicates supervisory expectations and observations, which in practice shape examination questions rather than creating free-standing duties. That makes it weaker than a rule and stronger than advice: an institution that cannot answer the questions it implies has a supervisory problem whether or not it has a legal one.

For a Swiss institution deploying a third-party platform the consequence for evaluation is direct. The statutory duties are the institution’s own, and they sit under AMLA and the FADP rather than under the vendor’s contract, so a vendor’s performance figures cannot discharge them. The measurements of Section 4 therefore have to be reproducible on the institution’s own document population. That is a stronger requirement than accepting a vendor benchmark, and it is why the framework specifies stratification by the institution’s own strata.

5.3. Data protection

The revised Swiss Federal Act on Data Protection (2023) imposes accuracy duties on personal data processing and provides rights in relation to automated individual decisions. Onboarding data is personal data of an unusually sensitive kind, and an extraction error is frequently an accuracy failure in the data-protection sense as well as a compliance one. The attribution requirement of Section 3.3 does double duty here: a value traceable to a source span is one whose accuracy can be contested and corrected, which is what the data-subject rights presuppose.

5.4. Anti-money-laundering obligations

Swiss AMLA (1997) and its implementing ordinances, and the FATF standards behind them (2012), require identification of the contracting party and the beneficial owner,

ongoing monitoring of the business relationship, and retention of documentation sufficient to reconstruct transactions and decisions. Two points connect to the framework. First, the retention requirement is the legal basis for Section 3.4: reconstruction is the standard, and reconstruction is a property of the record. Second, beneficial-ownership identification is exactly where a fabricated hallucination is both most consequential and least detectable by plausibility inspection, which is the argument for reporting the unsupported-output rate on screening-relevant fields separately and holding it near zero.

5.5. Mapping

5.6. Governance consequence: reliability evidence expires

Because the platform treats the model as replaceable, and because calibration is known to shift under distribution change (Desai & Durrett, 2020), a reliability finding is a statement about a configuration and a document population at a time. The governance implication is that re-evaluation must be triggered by component change and by drift in the document population, not scheduled annually because that is the audit cycle. An institution that evaluated once, at procurement, and treats the result as a standing property of the system has misunderstood what was measured.

6. Outlook: from framework to evidence

A framework paper earns its place by what it makes possible next. This section sets out the measurement programme the four dimensions imply, the directions in which the framework extends beyond the system it was built against, the problems it brings into view for others to work on, and the scope of what it establishes.

6.1. The immediate programme

Section 4 is a protocol, written to be executed as it stands. Its first instantiation will report per-stratum risk-coverage curves, a record-completeness figure, an unsupported-output rate broken down by hallucination type, and the error rate in an independently drawn sample of non-escalated cases.

Both outcomes are informative, which is the property a protocol should have before it is run. If the risk-coverage curves separate as we expect, the central claim holds in the domain it was written for, and a deploying institution gains a defensible basis for its threshold. If they run close together, the claim resolves to the existence result it inherits from selective prediction, and aggregate accuracy turns out to serve as an adequate proxy in this setting. Section 4.5 states that condition, and we will report whichever way it falls.

Publishing the measurement design ahead of the measurements is deliberate. It allows the second half of a research programme to be checked against the first, and it commits us to a specification we cannot quietly adjust once results are in hand.

Dimension	Duties and supervisory expectations engaged	Why measurement, not description, is required
Accuracy, stratified	FADP accuracy duty. AI Act Art. 15(3) only where the system is high-risk, which on the present Annex III it is not	A declared metric must be a measured metric, and an unstratified one conceals the cells that matter
Calibration	FINMA 08/2024 model risk. AI Act Art. 14 by analogy only	Oversight that cannot select the cases needing attention is not effective, and effectiveness is measurable only via calibration and risk-coverage
Hallucination and attribution	AMLA beneficial-ownership identification. FADP accuracy and contestability	A plausible ungrounded value passes human inspection, and only an attribution measurement surfaces it
Auditability	AMLA and ordinance retention. FINMA 08/2024. AI Act Art. 11 and Annex IV by analogy only	Reconstruction is the standard, and record completeness is an empirical property of the stored record
End-to-end and propagation	FINMA 08/2024 model risk	Stage-local figures systematically understate end-to-end risk

Table 5: Evaluation dimensions mapped to the duties they serve. AI Act references are marked *by analogy only* because, as Section 5.1 sets out, the high-risk regime does not reach this class of system on the present text of Annex III. They are retained because they describe good practice precisely, not because they bind.

6.2. Directions in which the framework extends

Three directions carry the framework past the system it was developed against. Each is a research programme rather than an application exercise.

Comparability across institutions.. Two vendors’ accuracy figures are not comparable today, because neither declares the strata over which the average was taken. The same four dimensions, measured to a common specification across different document populations, would yield the first reliability evidence in this domain that supports comparison at all. Reaching that point turns on agreeing a common set of strata, which is open problem 2 below.

A shared benchmark as the successor artifact.. Section 1.4 distinguishes this paper from a benchmark proposal, and the distinction doubles as a sequence: a benchmark standardises a measurement, so it presupposes that the measurement has been specified. The natural successor is a public task suite with annotated data and baselines. The binding constraint there is data rather than method, since the documents that matter are the ones that cannot be published, and the research question is which properties a synthetic or consented corpus would have to preserve for the task to remain the same task.

Adjacent regulated document work.. Insurance underwriting files, credit dossiers and documentary trade finance share the structure this framework assumes: heterogeneous documents, asymmetric error costs, threshold-routed human review, and a record that must survive an audit years later. The dimensions should transfer directly. The acceptance criteria will need rederiving, because the cost asymmetry differs, and testing that boundary is the cheapest available way to establish whether the framework is domain-specific or merely domain-motivated.

6.3. Open problems the framework brings into view

Specifying what to measure brings eight problems into focus that the framework itself leaves open. They are ordered from the cheapest to settle to the least technical, and each is stated as work that can be picked up rather than as a gap in this paper.

1. **Characterising evaluation practice.** The framework is motivated by our reading that institutions and vendors report an aggregate accuracy figure and little else. Establishing how far that holds is a content analysis of vendor documentation and procurement templates, and it is the one study in this programme that requires none of the data described in Section 4.1. It is also the cheapest study that would materially raise the standing of this paper’s motivation.
2. **Agreeing a common set of strata.** Comparability across institutions rests on shared stratum definitions. Document modality and language are straightforward to agree; the structural complexity of the underlying entity is the hard case, since a trust with layered ownership and a single natural person are not comparable units. Producing a defensible taxonomy is a contribution in itself, and it is a precondition for the cross-institution comparison of Section 6.2.
3. **Measuring cross-document coherence.** Section 2.3 identifies the failure and Section 3.5 identifies its unit, the client file. Supplying the metric is the open step. Adjudicating whether a set of documents tells one story is a harder annotation task than field-level correctness, and it is the measurement most likely to require a purpose-built protocol rather than an adaptation of an existing one.

4. **A defensible unsupported-output rate.** We argue that ungrounded outputs should be counted separately from wrong ones, and that the rate belongs at or near zero for fields feeding screening. Turning that direction into a criterion means establishing what value is defensible and how it should depend on the downstream control, which is an empirical question about how often an ungrounded value survives the controls placed after it.
5. **Deriving acceptance criteria from expected loss.** Section 4.4 argues that thresholds belong fixed before measuring and that they follow from the consequence of a given field being wrong. The open step is the derivation itself, from consequence to number, and it calls for a loss function over compliance outcomes. Compliance outcomes are not currently priced, so constructing that function is a research contribution rather than a calculation.
6. **Resolving the error budget against review capacity.** Criteria 1 and 5 of Section 4.4 will not both hold at every threshold, and we present that tension as the substance of the deployment decision. Resolving it is a constrained optimisation, and the work it needs is the objective function, which for this domain remains to be written down.
7. **Choosing a re-measurement trigger.** Section 5.6 argues that reliability evidence expires and that re-evaluation on component change belongs inside the control rather than alongside it. Which observable should fire that trigger is open: a model version changes, a document population shifts, a language is added. Deciding which of those warrants the full protocol and which warrants a spot check is a governance question with an empirical core, and answering it would make re-evaluation schedulable rather than discretionary.
8. **What a non-reproducible record evidences.** Section 3.4 observes that a decision taken on a since-retired model version may not be reproducible, and argues that an institution should settle deliberately what its record then establishes. Whether a record of a decision’s stated basis satisfies a supervisor who asked about the decision is a question for supervisory practice as much as for system design, and it is the item on this list least likely to be settled by measurement alone.

6.4. Limitations

Scope of the framework as established.. The framework was developed against a single deployed platform, and the regulatory analysis is specific to Swiss and EU instruments. The dimensions are stated so that they should generalise; establishing that they do, in particular for an architecture that reaches its human reviewer by some route other than a confidence threshold, is the extension test described in Section 6.2.

The prevalence premise.. Section 3.2 separates the general result, which is inherited from selective prediction, from the domain claim that deployed systems occupy the region where the risk–coverage advantage is wide. That domain claim is the hypothesis the protocol in Section 4 is built to test, and Section 4.5 states what would count against it.

Evidential status of this paper.. The contribution here is a measurement design. The measurements are the programme’s next output, and Section 6.1 sets out what they will report and what each result would mean for the central claim.

Operator-reported efficiency figures are out of scope.. The operator reports workload and unit-cost improvements from client deployments. These are internal estimates rather than controlled measurements, and they answer a different question, since a system can reduce cost per file while raising the rate of a rare and expensive error. Bringing efficiency into scope would require a throughput study run alongside the reliability protocol, with its own design and its own controls, and we treat that as separate work rather than as a hedge to attach here.

Provenance discipline.. Every quantity considered for this paper was checked against a ledger recording, for each figure, its source, its measurement date where one exists, and whether it is measured, projected, unconfirmed or illustrative. Only figures carrying the status *measured* are eligible to appear, and a release check enforces that at the point a draft is prepared. We describe the mechanism because a paper that merely happens to omit unmeasured figures is indistinguishable from a lucky one, and because the discipline is reusable by anyone writing alongside a live deployment.

7. Conclusion

The question institutions deploying LLM document intelligence into onboarding have to answer is not “is it accurate” but “is it good enough to put into a regulated process”. Those are different questions, and the second is not answered by a better version of the first.

We have argued that answering it requires four measurements rather than one. Extraction accuracy is necessary and, on its own, close to uninformative about deployment risk. Confidence calibration, together with sharpness, is what licenses a threshold-routing, human-in-the-loop architecture, which is the shape the platform we study adopts. A system with higher accuracy but uninformative confidence can therefore be the worse choice. Hallucination is a distinct failure class rather than a subtype of error, and it demands attribution measurement because it is the failure that survives human plausibility inspection. Auditability is a property of the retained record rather than of the model, and it is therefore measurable now, on the record, rather than argued about later.

Three of those four points imply the same architectural commitment, that every extracted value carry a source span. We take that as evidence that the framework is describing one coherent requirement rather than four preferences.

The methodological recommendation we would press hardest is narrower and, we think, the most neglected: audit the cases the system did not escalate. Errors in the escalated stratum are visible because somebody looks at them. Errors in the auto-approved stratum are invisible for the same reason, and that is the stratum where an institution has chosen to accept risk without review. Every incentive in an operational setting points away from that measurement, and it is the one that establishes whether the threshold is right.

Finally, none of the four dimensions is a property of the model alone. They are properties of the model, plus the workflow, plus the record. A firm that improves its model and changes nothing else may not have improved its compliance position at all, and on the argument above will have no way of knowing.

The empirical instantiation of this framework is planned, and Section 6 sets out both that programme and the problems the framework raises without settling. We have tried to specify the measurements precisely enough that they can be held to, including the parts that would embarrass us.

Acknowledgements

Innovation project supported by Innosuisse. This work was carried out within the Innosuisse innovation project “A specialized and secure AI orchestrator”, a collaboration between the University of Applied Sciences of the Grisons (FHGR) and WeCanGroup SA.

The authors thank the project’s industry partner for access to the deployed system described in Section 2.

CRedit author contribution statement

Joerg Osterrieder: conceptualisation, methodology, regulatory analysis, writing of the original draft, supervision, funding acquisition. **Lennart Baals:** methodology, investigation, writing, review and editing. **Illia Kosterin:** validation, review and editing. **Didier Baertschiger, Nathan Douet, Vincent Pignon:** description of the deployed platform in Section 2, domain expertise on compliance workflow and document populations, review and editing.

The evaluation framework of Sections 3 and 4, and the regulatory analysis of Section 5, were specified by the FHGR authors. We state that division explicitly because three of the authors are employed by the operator of the platform under study, and the criteria by which a system is judged should not be set by the party supplying it.

Declaration of competing interest

Three of the authors are employed by WeCanGroup SA, which operates the platform discussed in Section 2, and the work was carried out under a funded collaboration with that company. Two consequences are worth stating rather than leaving to inference. No performance figure supplied by the operator is reported anywhere in this paper, for the reasons given in Section 6.4. And the framework in Sections 3 and 4 sets out measurements the operator has not yet published, so it is a standard the platform is held to rather than a description of results it has achieved.

Note on the legal analysis

Section 5 reads primary instruments and states what they say. It is analysis of what those texts require and it is not legal advice; it has not been reviewed by qualified counsel, and no author holds a legal qualification. Where a reading is contestable, and the scope of Annex III in Section 5.1 is the clearest instance, we quote the provision so that a reader can check the reading against the text rather than against our summary of it.

Data availability

No new data are reported. The protocol in Section 4 specifies an annotated corpus of real onboarding documents which has not been collected. That corpus will contain personal data of natural persons under the revised Swiss Federal Act on Data Protection and client-confidential material under the project’s research agreement, so it is not expected to become publicly available. Section 6.2 discusses what a releasable substitute would have to preserve, and treats the question as open.

Code availability

No code is reported. The framework specifies measurements rather than an implementation of them.

Ethics and broader impact

The paper reports no experiments involving human participants and processes no personal data.

The framework does carry an equity consequence that we would rather state than leave implicit. Section 3.7 observes that extraction quality varies with document modality and language, and that both correlate with a client’s jurisdiction of origin. A system that performs worse on photographed documents, or in a less-resourced language, escalates those files more often and so imposes more friction on the clients who present them. Nobody designs that outcome; it emerges from a per-stratum difference that pooled reporting conceals. We therefore treat stratified measurement as an evaluation requirement rather than

as an optional fairness audit, which is why stratification appears in every row of Table 4.

The framework’s intended effect is to make it harder to deploy a system on the strength of an aggregate accuracy figure. If it has an unintended effect, we expect it to be that a well-resourced institution can satisfy the framework and a smaller one cannot, since the non-escalated audit of Section 4.2 costs reviewer time that a small compliance function may not have. Reducing that cost, most plausibly by establishing how small an audit sample suffices, is work the framework invites rather than supplies, and we record the burden as a cost of the proposal.

Funding

This work was carried out within the Innosuisse innovation project “A specialized and secure AI orchestrator”. The funder had no role in the design of the framework, in the analysis, or in the decision to submit. Under article 6.2 of the project’s research agreement the industry partner receives any publication draft for review before submission, and that provision applies to this paper.

Declaration of generative AI use in the writing process

No generative AI tool is an author of this work. Attribution of authorship carries accountability for the content, and that accountability rests with the named authors alone.

Generative AI tools were used for language editing and to support bibliographic verification. The authors reviewed and edited the content as required. The framework, the evaluation protocol and all conclusions are the authors’ own, and the authors take full responsibility for the content of the publication. The verification records, covering reference resolution and a recorded locator for every citation, accompany the paper in its repository.

References

- AMLA. (1997). *Federal Act of 10 October 1997 on Combating Money Laundering and Terrorist Financing (Anti-Money Laundering Act, AMLA)*. Swiss Confederation, SR 955.0, as amended. Retrieved from https://www.fedlex.admin.ch/eli/cc/1998/892_892_892/en
- Borella, L. (2025, September). *Building a shared benchmarking framework for AI in financial services*. FINOS (Fintech Open Source Foundation). Retrieved from <https://www.finos.org/blog/building-a-shared-benchmarking-framework-for-ai-in-financial-services> (Published 26 September 2025)
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
- Desai, S., & Durrett, G. (2020). Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 295–302). doi: 10.18653/v1/2020.emnlp-main.21
- El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11, 1605–1641. doi: 10.5555/1756006.1859904
- EU AI Act. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. (European Parliament and Council)
- FADP. (2023). *Federal Act on Data Protection (FADP)*. Swiss Confederation, SR 235.1. (Revised version in force from 1 September 2023)
- FATF. (2012). *International Standards on Combating Money Laundering and the Financing of Terrorism and Proliferation: The FATF Recommendations*. Financial Action Task Force. Paris. (Adopted 2012, as amended)
- FINMA. (2024). *Guidance 08/2024: Governance and risk management when using artificial intelligence*. Swiss Financial Market Supervisory Authority (FINMA). Bern.
- Geburu, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. doi: 10.1145/3458723
- Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 4878–4887).
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330).
- Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansky, D., Cohen, V., ... Matias, Y. (2022). TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 3905–3920). doi: 10.18653/v1/2022.naacl-main.287
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55. doi: 10.1145/3703155
- ISO/IEC. (2023). *ISO/IEC 42001:2023, Artificial intelligence management system*. International Organization for Standardization / International Electrotech-

- nical Commission. Geneva.
- Jaume, G., Ekenel, H. K., & Thiran, J.-P. (2019). FUNSD: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)* (pp. 1–6). doi: 10.1109/ICDARW.2019.10029
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. doi: 10.1145/3571730
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., ... Kaplan, J. (2022). *Language models (mostly) know what they know*. arXiv preprint arXiv:2207.05221. doi: 10.48550/arXiv.2207.05221
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 1906–1919). doi: 10.18653/v1/2020.acl-main.173
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)* (pp. 220–229). doi: 10.1145/3287560.3287596
- Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (ICML)* (Vol. 119, pp. 7076–7087).
- Naeni, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 2901–2907). doi: 10.1609/aaai.v29i1.9602
- NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, U.S. Department of Commerce.
- Osterrieder, J., & Baals, L. (2026). *When the AI hallucinates, who pays? mapping LLM hallucination types to legal liability in EU and Swiss financial regulation*. FHGR working paper.
- Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., & Lee, H. (2019). CORD: A consolidated receipt dataset for post-OCR parsing. In *Workshop on Document Intelligence at NeurIPS*.
- Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., ... Reitter, D. (2023). Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4), 777–840. doi: 10.1162/coli_a_00486
- Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421.
- Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., ... Yang, D. (2022). When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2322–2335). doi: 10.18653/v1/2022.emnlp-main.148
- Tibshirani, R. J., Foygel Barber, R., Candès, E. J., & Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS)* (pp. 2526–2536).
- Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., ... Huang, J. (2024). FinBen: A holistic financial benchmark for large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)* (pp. 95716–95743). doi: 10.52202/079017-3033