

An NDMO-Inspired, SAMA-Aware, Human-Gated Classifier for Saudi Banking Metadata: Auditable, Impact-Driven, Sovereign-by-Design

Rashad Kamsheh
Independent Researcher
Sweden

rashad.kamsheh@outlook.com

Abstract—Data classification is a foundational governance and cybersecurity control in Saudi banking, yet it remains largely manual and rarely leaves an audit trail. Commercial banks answer to the Saudi Central Bank (SAMA), whose Cyber Security Framework requires an information-asset classification scheme but does not prescribe one; the National Data Management Office (NDMO) four-level scheme, designed for public entities, is a well-specified template a bank may adopt, and the Personal Data Protection Law (PDPL) applies directly. We present a design-science artifact for this setting: a policy-grounded *deterministic* classification and decision-support engine, not a learned classifier, that proposes a security level only where a policy-cited rule supports one, records the impact reasoning behind every proposal, declines to label fields it does not recognise, and routes every declined or high-stakes field to a human reviewer. PDPL obligations are reported on a separate axis that annotates a field but never changes its level. The evaluation separates classification quality from workflow safety. On 60 columns of external metadata from the open-source core-banking system Apache Fineract, every level the engine proposes agrees with the reference, but it proposes one for only 23.3% of fields, and 81.7% still require human review: coverage, not precision, is the binding limit. On a 200-column synthetic corpus built from the rule set’s own vocabulary, agreement is 99.4% and no under-classification ever passes unreviewed — a rule-conformance result, not independent validation, since all security-level reference labels were assigned by the authors. Two blind annotators ($\kappa \geq 0.87$) independently evaluated the PDPL axis: precise where it fires, but its Credit Data limb misses all 18 positive cases. These results validate the design and its evaluation protocol; they do not yet establish operational readiness for banking deployment.

Index Terms—data classification, data sovereignty, PDPL, NDMO, SAMA, RegTech, design science research, human-in-the-loop, selective classification.

I. INTRODUCTION

Saudi Arabia is investing at national scale in an AI-driven digital economy, which creates a tension between using data for value and keeping it protected and sovereign. Vision 2030 reframes data as sovereign national infrastructure, and the instruments that enforce this view (the SDAIA/NDMO Data Classification Policy [1], the PDPL [2], and the SAMA Cyber Security Framework [3]) all depend in practice on a common prerequisite: knowing what each data field is. The level drives the security controls (encryption, access, residency), while

PDPL duties attach to the statutory category, purpose, and lawful basis, assessed separately; both assessments begin at the same place.

For a bank the governing authority is worth stating exactly. The NDMO policy binds *public* entities; a commercial bank is supervised by SAMA, whose framework requires the institution to operate an information-asset classification scheme but does not mandate NDMO’s particular one [3]. The PDPL applies directly to any controller of personal data [2]. Our artifact is therefore deliberately **NDMO-inspired**: it adopts NDMO’s well-specified four-level, four-dimension impact model as a template a SAMA-regulated bank may lawfully choose. It is equally **SAMA-aware**, situating that scheme inside SAMA’s governance requirement, while treating PDPL obligations as directly binding. We do not claim NDMO jurisdiction over banks; we claim NDMO’s scheme is a defensible, auditable choice under SAMA. The NCA’s Essential Cybersecurity Controls apply where the institution falls within their scope [4].

Classification done by hand is slow, varies between analysts, and seldom records *why* a field received its label: when a regulator asks why a field was marked Secret, the answer is too often institutional memory rather than a citation to policy. No published measurement exists of how often Saudi banks misclassify metadata or of the analyst effort involved — itself a symptom of the missing audit trail — so the problem is motivated from the obligations the instruments impose, and the operational base rate is left to the deployment study Section V specifies. The instruments are in force; defensible labels for every field remain the gap.

We contribute an impact-driven classification agent that treats the adopted scheme not as a keyword dictionary but as a step-by-step impact assessment, attaching an auditable justification to every label. The positioning matters for an AI venue: this is a policy-grounded *deterministic* classification and decision-support artifact with selective abstention and human gating, not a learned classifier. (“Agent” denotes the orchestration-and-escalation layer; the level-deriving logic admits no model, and the optional AI detector, off in every experiment reported here, proposes only a category for an

unmatched field, never a level, and is always gated.) Research question: how can a data-classification artifact be designed so that each decision is at once automated and defensible under the Saudi regulatory stack a bank actually answers to?

Contributions. **(C1)** a conservative adaptation of the NDMO four-level, four-dimension impact scheme in computable form, with PDPL obligations emitted as a *separate* flag axis derived independently of the level rule — the impact representation contributes traceability and replayability, not predictive power (Section IV-B); **(C2)** a banking rule set of sixteen policy-cited sensitivity rules encoding the NDMO-style rollout, a protective interim default, and a gated national-security path, with the rule inventory and the complete pattern set shipped in the artifact package (RULES.md, PATTERNS.md); **(C3)** an abstention-aware evaluation that separates classification quality from workflow safety, with ablations that isolate what each design element contributes and an adversarial probe that locates where silent error actually enters; **(C4)** the three-alignment test (regulatory, operational, infrastructural) as a reusable evaluation lens; and **(C5)** a sovereignty-by-design deployment model that keeps classified content inside the institution’s boundary.

II. BACKGROUND AND RELATED WORK

A. The Saudi regulatory stack a bank answers to

The NDMO policy defines four levels (Public, Restricted, Secret, Top Secret) and assigns a level by assessing the impact of unauthorised disclosure across four dimensions: national interest, entity activities, individuals, and environment [1]. Two further principles matter computationally: a dataset inherits the highest level among its fields (Principle 4), and data not classified at creation or receipt is treated as Restricted until a Business Data Executive completes a formal impact assessment (§1.4). The PDPL defines Personal Data broadly, defines Credit Data separately, and enumerates Sensitive Data as seven limbs: racial or ethnic origin; religious, philosophical (intellectual) or political beliefs; security and criminal data; biometric data used for identification; genetic data; health data; and data revealing that one or both parents are unknown [2]. Cross-border transfer is governed separately [5]. These are distinct axes: a field’s security level and its statutory data category answer different questions, and an artifact that collapses them cannot cite either correctly.

B. Sovereignty, governance, and classification

Digital sovereignty is a contested construct spanning territorial control, self-determination, and infrastructural independence [6]–[10]; cross-border transfer rules are where it becomes operational for GCC controllers [11]. Data governance frameworks position classification as a control input [12], and classification systems are never neutral: what a scheme makes visible is a political choice with downstream consequences [13]. Deployed on-premise, our artifact operationalises this literature at the field level as an *input* to sovereignty controls, not a complete control: residency,

TABLE I
THE CLAIMED COMBINATION AGAINST NEIGHBOURING STRANDS (✓ PRESENT, ~ PARTIAL, – ABSENT). **IMPACT** = IMPACT-DERIVED LEVEL WITH REGULATORY GROUNDING; **AXIS** = STATUTORY CATEGORIES KEPT SEPARATE FROM THE LEVEL; **CITE** = A POLICY CITATION ON EVERY DECISION; **GATE** = EXPLICIT ABSTENTION WITH HUMAN GATING; **SAUDI** = NDMO/SAMA/PDPL BANKING SCOPE. NONE COMBINES THE FIVE.

Strand	Impact	Axis	Cite	Gate	Saudi
PII / privacy benchmarks [16]–[18]	–	~	–	–	–
Sensitive-data grading [19], [20]	~	–	–	–	–
RegTech gen. / retrieval [21]–[23]	–	–	✓	–	–
Policy-aware governance [15]	–	–	~	~	–
Saudi PDPL analyses [24], [25]	–	~	~	–	~
Selective classification [14]	–	–	–	~	–
This artifact	✓	✓	✓	✓	✓

transfer blocking, and access enforcement must consume the classification before sovereignty becomes operational.

C. Abstention as the operative design choice

Selective classification pairs a predictor with a rejection rule and analyses the coverage/risk trade-off [14]; our contribution is not that machinery but the observation that in a regulated setting the trade-off is fixed *by the instrument rather than by tuning*. A field whose statutory category is unresolved has a prescribed disposition, the §1.4 interim default, so the operating point is not a hyperparameter and the rejection region carries a citation of its own. That is why we report coverage and selective agreement separately rather than a single accuracy: the two axes answer to different authorities. Policy-aware governance work pursues the same auditability goal for data access with generative components [15]; by contrast our level-deriving path admits no model, and the abstention, not the prediction, carries the obligation.

D. Automated sensitive-data classification

Benchmarks for personal-information detection and privacy evaluation have matured rapidly [16]–[18], and sensitive-data classification and grading has been studied for financial cloud and structured-data settings [19], [20]. A parallel RegTech strand generates compliance checks from regulation and retrieves regulatory answers [21]–[23]. Saudi-specific legal analyses examine PDPL scope and effectiveness [24], [25]. Table I makes the contrast systematic: no prior work we identified combines an NDMO-grounded impact derivation, a separately derived PDPL axis, a per-decision policy citation, and an explicit abstention-and-gate architecture for Saudi banking; that combination, not compliance automation in general, is what we claim.

III. ARTIFACT AND METHOD

A. Design

The artifact follows design-science practice [26], [27], with evaluation designed for risk and effectiveness [28] and reported as a nascent design theory [29]. Four steps run per field (Fig. 1): **detect** (match the field name and its

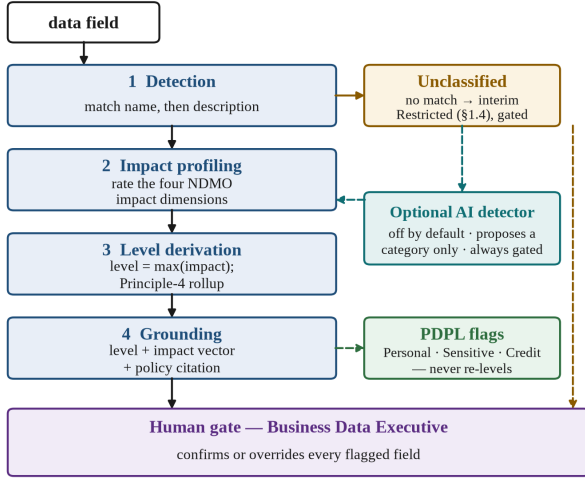


Fig. 1. The four-step pipeline: detect, profile impact, derive level (Principle-4 rollup), ground with a policy citation. PDPL flags are a separate axis that never re-levels a field; unmatched fields return Unclassified with interim Restricted handling; the optional AI detector (off in every experiment here) proposes only a category. The gate escalates on a Top Secret candidate, assurance below 0.75, the §1.4 default, an AI-detected category, or a name-based PDPL candidate.

description against the rule set), **profile** (rate the four NDMO impact dimensions), **derive** (level = max of the four ratings; Principle-4 rollup for tables), and **ground** (emit the level, the impact vector, and a policy citation); Algorithm 1 states it in full. Read as layers, the design separates four concerns: steps 1–3 are the *classification model*; step 4 and the PDPL axis are the *regulatory reasoning layer*; the gate is the *human-control mechanism*; and abstention with the §1.4 interim default is the *workflow-safety mechanism*. Each is evaluated on its own terms in Section IV.

Two properties do the safety work. First, an unmatched field is *not* assigned a level: it is returned Unclassified with interim Restricted handling under §1.4 and always gated. Second, the PDPL axis is derived independently. Flags are unioned over *every* rule that fires plus a conservative name-based candidate detector, so `credit_score_balance` retains its Credit Data candidate although the lower-level `balance` rule wins the level.

B. Matching, and where silent error enters

Matching is boundary-aware: a single-word pattern must equal a whole normalised token and a multi-word pattern must occur as a complete phrase. Every rule that fires on *either* the name or the description is collected, and the most protective derived level wins (DP5), with ties resolved by rule order so the emitted citation is stable.

This is a correction, and its reason is the paper’s central methodological finding. The previous matcher used unrestricted substring containment, returned the first matching rule, and consulted the description only when the name matched nothing. That produced false positives (`pan` fires

Algorithm 1 `CLASSIFYFIELD(name, desc)`. Dual output: one candidate level and an independently derived PDPL flag set

```

1:  $M \leftarrow \{r \in \text{RULES} : \text{FIRES}(r, \text{name}) \vee \text{FIRES}(r, \text{desc})\}$  both fields,
   always
2:  $F \leftarrow \bigcup_{r \in M} \text{PDPLFLAGS}(r) \cup \text{NAMEFLAGS}(\text{name})$  axis is
   independent of the level
3: if  $M = \emptyset$  then
4:   return  $\langle \text{UNCLASSIFIED}, \text{interim} \leftarrow \text{Restricted}, \text{flags} \leftarrow F, \text{gate} \leftarrow$ 
     true,  $\text{cite} \leftarrow \text{§1.4} \rangle$ 
5: end if
6:  $r^* \leftarrow \arg \max_{r \in M} \text{rank}(\text{LEVEL}(r))$  DP5: most protective wins
7:  $\ell \leftarrow \text{LEVEL}(r^*) = \text{map}(\max(r^*. \text{impact}))$ 
8:  $g \leftarrow (\ell = \text{Top Secret}) \vee (r^*. \text{assurance} < 0.75) \vee \text{aiDetected} \vee$ 
    $\text{NAMEFLAGS}(\text{name}) \neq \emptyset$  name-only signals gate
9: return  $\langle \text{CANDIDATE}, \ell, \text{flags} \leftarrow F, \text{gate} \leftarrow g, \text{cite} \leftarrow r^*. \text{citation},$ 
    $\text{trace} \leftarrow M \rangle$ 

```

FIRES matches whole normalised tokens or complete phrases only. Table rollup: $\text{level}(T) \leftarrow \max_{c \in T} \text{level}(c)$ (Principle 4); $\text{flags}(T) \leftarrow \bigcup_{c \in T} \text{flags}(c)$.

inside “company”; `race` inside “trace”) and, more seriously, silent under-classification: `aml_alert_balance` described as “suspicious activity investigation account balance” matched the low-level `balance` rule by name, so its AML description was never read, giving Secret evidence, a Restricted label and no gate. `adversarial_lexical.py` runs the 19 cases against both matchers and reports 19/19 versus 5/19 (12 collisions, 2 silent under-classifications) plus the 27 patterns short enough to collide (at most five characters with no separator; marked in the artifact’s pattern inventory). *The conformance corpus exposes none of this*: the substring matcher scores 100% on it (Table III); only the adversarial probe separates them.

C. Evaluation design

We report two metric families and never mix them. *Classifier* metrics cover only fields that actually received a candidate level: coverage, agreement among candidates, selective agreement among ungated candidates, and macro-F1 over candidates only. *Workflow* metrics describe what an operator experiences: gate rate, interim-handling agreement, safe-handling (correct or escalated), silent unsafe rate, and error direction. In plain terms, *coverage* is the share of fields the engine will label, *agreement* how often an issued label matches the reference, *selective agreement* the same among labels no human re-checks, *gate rate* the share a human must look at, *safe-handling* the share either correct or in front of a human, and a *silent* error a wrong label nobody was asked to review — the outcome the design exists to prevent. Counting an unmatched field’s interim handling as a prediction of “Restricted” would mix the two families and inflate both.

Sets. A 200-column *synthetic rule-conformance corpus*, so named because it was built from the vocabulary the rules encode and measures conformance, not independent validity. Two out-of-sample corpora never built or tuned on: 18 insurance and telecom fields, and 45 banking-alternate,

TABLE II

CONFORMANCE CORPUS ($n=200$), CLASSIFIER METRICS SEPARATED FROM WORKFLOW METRICS. COV. = SHARE RECEIVING A CANDIDATE LEVEL; AGR. = AGREEMENT AMONG CANDIDATES; SEL. = AGREEMENT AMONG *ungated* CANDIDATES. THE BASELINES NEVER ABSTAIN: 100% COVERAGE BY CONSTRUCTION.

Approach	Classifier			Workflow		
	Cov.	Agr.	Sel.	Gate	Safe-h.	Silent
Keyword $\rightarrow$ Public default	100	35.5	35.5	0.0	35.5	119
Keyword + default + gate	100	53.0	59.0	50.0	79.5	31
Always-Restricted	100	40.0	40.0	0.0	40.0	80
Always-escalate	100	40.0	n/a	100	100	0
Impact-driven engine	85.0	99.4	99.1	43.5	99.5	0

healthcare and government fields. A post-freeze external test of 60 column names transcribed verbatim from Apache Fineract’s published core-table schema, with reference levels fixed before the single run, classified name-only with the rule set frozen, evaluated once. Reference levels throughout are author-assigned from policy and fixed before scoring.

Each rule carries a hand-assigned, uncalibrated *assurance* value used as a gate trigger; it is not a confidence and implies no probability. Whether safety depends on those values is tested by the threshold sweep in Section IV-B.

IV. RESULTS

A. Conformance corpus and baselines

Table II gives the split. The engine issues a candidate level for 85.0% of fields and agrees with the reference on 99.4% of those, at a 43.5% gate rate, with **zero silent under-classifications**. The single ungated disagreement is an over-classification (`public_rate`, Public $\rightarrow$ Restricted) produced by DP5 resolving a two-rule conflict upward: the safe direction, costing analyst time rather than exposure.

The keyword baselines never abstain, so their 100% coverage is not comparable capacity, and they carry 18 keywords against the engine’s rule vocabulary. The defensible reading is not that the engine is “more accurate” but that it is precise where it commits and abstains elsewhere, whereas the baselines commit everywhere and are silently wrong on 12–63.5% of fields.

B. Ablations: what each design element contributes

Table III holds the vocabulary fixed and varies one element at a time.

Three results matter. *The impact vector contributes no predictive power*: storing each rule’s implied level is identical to deriving it, necessarily, since the level is a pure function of the vector. What the representation buys is a replayable rationale and the Principle-4 rollup, and the paper claims nothing more. *The safety architecture is what suppresses silent error*: removing the gate and the interim default introduces 15 silent under-classifications where there were none. *Withholding a whole rule family degrades gracefully*: dropping the PDPL-sensitive family cuts coverage to 78.0% and raises the gate

TABLE III

ABLATIONS ON THE CONFORMANCE CORPUS, VOCABULARY HELD FIXED. THE SUBSTRING ROW IS THE PRE-CORRECTION MATCHER: INDISTINGUISHABLE HERE, AND 5/19 ON THE ADVERSARIAL PROBE WHERE THE CORRECTED MATCHER SCORES 19/19.

Variant	Cov.	Agr.	macro-F1	Gate	Silent
Engine	85.0	99.4	0.995	43.5	0
Stored level, no impact vector	85.0	99.4	0.995	43.5	0
No gate, no §1.4 default	100	90.0	0.897	0.0	15
Name-only (no description)	84.0	100	1.000	44.0	0
Substring matcher (pre-fix)	85.0	100	1.000	44.0	0
Leave out PDPL-sensitive family	78.0	99.4	0.995	50.5	0

to 50.5%, with silent error still zero: the loss is absorbed by abstention, not by wrong answers.

A fourth result addresses whether safety depends on the hand-assigned assurance values. The gate compares each value to a cut $\tau = 0.75$; sweeping τ over the whole assurance range (0.60–1.00) on the conformance corpus and on Fineract moves the gate rate from 22.5% to 100% and selective agreement between 99.1% and 100%, *and produces zero silent under-classifications at every τ on both corpora*. The property is structural: the three other triggers (Top Secret candidate, §1.4 interim, name-based PDPL candidate) never consult assurance, and since shifting every value by δ equals lowering τ by δ , the sweep covers every global perturbation: the values tune reviewer workload, not safety (`sensitivity_gate.py`).

C. Out-of-sample behaviour

Table IV reports three corpora the rule set was not built on. The pattern is consistent and is the paper’s main empirical claim: **every rule that fires agrees with the author-assigned reference label, on every set, and coverage is the binding limit**. On Fineract, coverage falls to 23.3% while agreement among candidates stays at 100% and safe-handling at 100%. Vocabulary transfer, not regulatory reasoning, is what fails off-domain, and it fails into abstention and the gate rather than into silent error.

The operational cost must be stated as plainly as the benefit. The share of fields a human must still examine (abstentions plus gated candidates) is 43.5% on the conformance corpus, 31.1% held-out, 77.8% fresh, and 81.7% on Fineract. On unfamiliar vocabulary the artifact is a triage and audit aid, not an automation win: abstention keeps it safe but is not capability, and a deployment should fix targets for coverage, gate rate, and reviewer workload, measuring analyst minutes per gated field (which we have not), before the approach is judged useful.

D. Adversarial lexical probe

Nineteen cases covering lexical collisions, compound fields whose description outranks their name, and flag-independence. The corrected engine passes 19/19; the pre-correction matcher passes 5/19, failing twelve collisions

TABLE IV
OUT-OF-SAMPLE CORPORA. FINERACT IS NAME-ONLY, RULES FROZEN,
EVALUATED ONCE. INTERIM = INTERIM-HANDLING AGREEMENT AMONG
ABSTENTIONS, REPORTED SEPARATELY BECAUSE IT IS WORKFLOW
SAFETY, NOT CLASSIFICATION ACCURACY.

Corpus	Cov.	Agr.	Interim	Safe-h.	Silent
Fresh: insurance + telecom (18)	22.2	100	6/14	100	0
Held-out: bank/health/gov (45)	75.6	100	7/11	100	0
External: Apache Fineract (60)	23.3	100	24/46	100	0

and two silent under-classifications. Boundary defects differ by pattern arity, so `card_notification_code` and `api_keyboard_layout` assert that a multi-token pattern cannot fire across token interiors, the defect that survived a first fix of the single-word branch only. Read against Table III, where the matchers are indistinguishable, this is the paper’s clearest methodological point: a corpus built from a rule set’s own vocabulary cannot detect a matcher defect; only adversarial construction can. Nineteen cases characterise the two failure modes found, not the lexical risk of a 243-pattern vocabulary; generated collision sets at scale, abbreviations, misspellings, mixed Arabic–English names, and ambiguous names from real estates are required before the matcher can be called adversarially validated.

E. Independent PDPL flag evaluation

A data engineer and a data scientist, neither a legal practitioner, labelled 123 fields blind to engine output and to each other: the 60 Fineract columns plus 63 out-of-sample fields with descriptions. Their labels are practitioner readings of the statute, not legal opinions. Agreement was computed before adjudication, the engine run only afterwards. Two instrument defects found later are disclosed rather than absorbed: ten worked examples the guide called absent from the workbook were units, and five fields are contested. Excluding the ten leaves $n=113$ for agreement, where Cohen’s κ is 0.874, 0.960 and 1.000; dropping the contested five as well leaves $n=108$ for scoring, those rows having no reference label. Personal reaches 97.1% precision at 36.3% recall (10.9% on names alone, 62.2% with a description); Sensitive 100% at 85.7%, provisional, since the guide paraphrased the security-and-criminal limb too narrowly for four security-context fields. **Credit has eighteen reference positives and detects none** — precision undefined rather than zero, since it issues no positive prediction on any scored field. `pdpl_name_flags` tests the column name and never the table, so `m_loan.principal_amount` cannot fire. We report this rather than repair it: bare `loan` and `credit` hints recover only 3 of the 18, and fitting a repair to these labels would spend the only independent Credit evidence there is.

F. Threats to validity

The conformance corpus shares vocabulary with the rule set by construction and is therefore not independent validation; we name it accordingly and report out-of-sample corpora

alongside. Reference levels are author-assigned throughout, which is the study’s principal limitation and one we do not claim to have addressed: the same team wrote the policy interpretation, the rules, and the labels the rules are scored against, so the agreement figures reported here are internal consistency, not accuracy. A single annotator cannot establish that a label is correct rather than merely consistent, so no *level* result here is validated against independent judgement; Section IV-E applies that standard to the flag axis and the same study now needs repeating for levels. The planned instrument: blind re-labelling of the same frozen corpora by several independent experts in banking cybersecurity, information governance, or Saudi regulation, inter-rater agreement reported before adjudication, and the adjudicated labels replacing ours — the Section IV-E protocol applied to levels, with at least one annotator holding demonstrated Saudi PDPL or banking-compliance expertise. Zero-error cells admit an exact one-sided 95% upper limit of 1.49% ($n=200$) under an independent Bernoulli assumption, but the cases are synthetic or curated, not a random sample from a production population: conditional descriptive estimates, not bounds on a live-estate error rate. The flag study in Section IV-E is one round on 123 fields by two annotators: enough to establish that an axis never fires, not to separate close precision figures.

V. DISCUSSION

The artifact yields five transferable design principles. **DP1**: derive the level from an impact assessment, not a keyword lookup, so the rationale is replayable. **DP2**: emit a policy citation with every decision. **DP3**: keep statutory data categories on an axis separate from the security level, derived independently. **DP4**: never assign a level to an unrecognised field: abstain, apply protective interim handling, and gate. **DP5**: resolve conflicting matches to the more protective level.

The results locate the engineering problem precisely: regulatory grounding transfers (recognised categories are correctly elevated out of domain, and precision among candidates holds) while vocabulary does not, so a deployment’s cost is dominated by gate workload rather than error correction. This is a conservative, auditable workflow with explicit abstention, not a broadly validated classifier. The three-alignment test (regulatory, operational, infrastructural) is offered as a design conjecture; operational alignment is the weakest of the three here, because of that workload.

Adoption also depends on costs the evaluation does not yet price. The gate workload has to be paid in analyst time, and at estates of 10^5 metadata fields the gate rate, not engine runtime, is the binding cost. The rule set is a governed asset, not a constant: regulations, taxonomies, and naming conventions drift, so a deployment needs an owner for the rules (NDMO’s Business Data Executive is the natural home), versioned change control in which every rule edit re-runs the adversarial probe as a regression gate, update triggers tied to regulatory amendments, and institution-specific vocabulary — including mixed Arabic–English metadata — added under

that discipline to lift coverage. None of this is evaluated here; it defines what a deployment study must measure.

VI. CONCLUSION

We presented an NDMO-inspired, SAMA-aware classification and decision-support artifact for Saudi banking metadata that derives a candidate level from an impact profile, cites policy, keeps PDPL obligations on an independent axis, and abstains rather than guessing. What the evaluation establishes is a conservative, auditable workflow whose rules are precise where they fire, whose under-classifications surface to a human rather than pass silently at every gate threshold tested, and whose binding limit is coverage. What it does not establish is equally part of the result: every security-level reference label is author-assigned, so no level figure is independently validated; on external metadata 81.7% of fields still need a human; and the Credit Data limb of the PDPL axis missed all 18 independent reference positives, an entire statutory category, so the flag axis is not operationally mature. Next: blind multi-expert labelling of levels under the flag-study protocol, a redesigned Credit detector using table names, schema relationships, and descriptions, evaluated on a fresh test set, and coverage growth under versioned rule governance.

VII. ETHICS AND REPRODUCIBILITY

The evaluation uses a fictional bank reference model, synthetic de-identified metadata, and the published schema of an open-source system; no personal data were processed. This concerns the research evaluation only and does not demonstrate deployment-time PDPL compliance. Outputs are candidate labels pending human confirmation, not legal advice. Rule set, corpora and reference labels, the Fineract test with provenance, the ablation and adversarial harnesses, the flag study's guide, both blind label sets and its scorer, and a SHA-256 manifest ship as a public OSF project at osf.io/cqhdr (doi: [10.17605/OSF.IO/CQHDR](https://doi.org/10.17605/OSF.IO/CQHDR)). Eight commands regenerate every number reported above. The same engine runs in the browser at tasnif.live/ndmo-pdpl-classifier, a port checked against the artifact on every corpus row; field names entered there never leave the visitor's machine.

REFERENCES

- [1] National Data Management Office, "Data classification policy," v1.0, SDAIA, Kingdom of Saudi Arabia, 2020.
- [2] Saudi Data and Artificial Intelligence Authority, "Personal Data Protection Law," Royal Decree M/19 (2021), as amended by M/148 (2023), in force 14 Sep. 2023.
- [3] Saudi Central Bank (SAMA), "Cyber security framework," v1.0, SAMA, 2017.
- [4] National Cybersecurity Authority, "Essential cybersecurity controls (ECC-2:2024)," NCA, Kingdom of Saudi Arabia, 2024.
- [5] Saudi Data and Artificial Intelligence Authority, "Regulation on personal data transfer outside the Kingdom," SDAIA, 2024.
- [6] J. Pohle and T. Thiel, "Digital sovereignty," *Internet Policy Review*, vol. 9, no. 4, 2020. doi: [10.14763/2020.4.1532](https://doi.org/10.14763/2020.4.1532)
- [7] L. Floridi, "The fight for digital sovereignty: what it is, and why it matters, especially for the EU," *Philosophy & Technology*, vol. 33, no. 3, pp. 369–378, 2020. doi: [10.1007/s13347-020-00423-6](https://doi.org/10.1007/s13347-020-00423-6)
- [8] S. Couture and S. Toupin, "What does the notion of 'sovereignty' mean when referring to the digital?" *New Media & Society*, vol. 21, no. 10, pp. 2305–2322, 2019. doi: [10.1177/1461444819865984](https://doi.org/10.1177/1461444819865984)
- [9] P. Hummel, M. Braun, M. Tretter, and P. Dabrock, "Data sovereignty: a review," *Big Data & Society*, vol. 8, no. 1, 2021. doi: [10.1177/2053951720982012](https://doi.org/10.1177/2053951720982012)
- [10] F. von Scherenberg, M. Hellmeier, and B. Otto, "Data sovereignty in information systems," *Electronic Markets*, vol. 34, art. 15, 2024. doi: [10.1007/s12525-024-00693-4](https://doi.org/10.1007/s12525-024-00693-4)
- [11] H. H. Alhababi, "Cross-border data transfer between the GCC data protection laws and the GDPR," *Global Journal of Comparative Law*, vol. 13, no. 2, pp. 178–200, 2024. doi: [10.1163/2211906x-13020003](https://doi.org/10.1163/2211906x-13020003)
- [12] R. Abraham, J. Schneider, and J. vom Brocke, "Data governance: a conceptual framework, structured review, and research agenda," *Int. J. Information Management*, vol. 49, pp. 424–438, 2019. doi: [10.1016/j.ijinfomgt.2019.07.008](https://doi.org/10.1016/j.ijinfomgt.2019.07.008)
- [13] G. C. Bowker and S. L. Star, *Sorting Things Out: Classification and Its Consequences*. MIT Press, 1999. doi: [10.7551/mitpress/6352.001.0001](https://doi.org/10.7551/mitpress/6352.001.0001)
- [14] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *J. Machine Learning Research*, vol. 11, no. 53, pp. 1605–1641, 2010.
- [15] S. Al Mandalawi, M. A. Mohammed, H. Maclean, M. C. Cakmak, and J. R. Talburt, "Policy-aware generative AI for safe, auditable data access governance," in *Proc. 17th Int. Conf. Knowledge and System Engineering (KSE)*, 2025, pp. 1–6. doi: [10.1109/KSE68178.2025.11309632](https://doi.org/10.1109/KSE68178.2025.11309632)
- [16] H. Shen, Z. Gu, H. Hong, and W. Han, "PI-Bench: evaluating query-aware privacy protection systems," in *Proc. ACL*, 2026.
- [17] Y. Liu, Y. Jia, J. Jia, and N. Z. Gong, "Evaluating LLM-based personal information extraction and countermeasures," in *Proc. USENIX Security*, 2025.
- [18] H. Li et al., "PrivaCI-Bench: evaluating privacy with contextual integrity and legal compliance," in *Proc. ACL*, 2025.
- [19] L. Zu, W. Qi, H. Li, X. Men, Z. Lu, J. Ye, and L. Zhang, "UP-SDCG: a method of sensitive data classification for collaborative edge computing in financial cloud environment," *Future Internet*, vol. 16, no. 3, art. 102, 2024. doi: [10.3390/fi16030102](https://doi.org/10.3390/fi16030102)
- [20] Y. Li, Z. Wu, J. Li, and L. Xie, "A multi-scene automatic classification and grading method for structured sensitive data based on privacy preferences," *Future Internet*, vol. 17, no. 9, art. 384, 2025. doi: [10.3390/fi17090384](https://doi.org/10.3390/fi17090384)
- [21] S. Li et al., "Compliance-to-code: enhancing financial compliance checking via code generation," in *Proc. 32nd ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, 2026. doi: [10.1145/3770854.3785703](https://doi.org/10.1145/3770854.3785703)
- [22] T. Gokhan, K. Wang, I. Gurevych, and T. Briscoe, "Shared task RIRAG-2025: regulatory information retrieval and answer generation," in *Proc. 1st Regulatory NLP Workshop (RegNLP)*, 2025.
- [23] N. Bagherifam, S. Naghdi, V. Ahmadian, A. Fazlzadeh, and M. Baghalzadeh Shishehgarkhaneh, "Digital regulatory governance: the role of RegTech and SupTech in transforming financial oversight," *Int. J. Financial Studies*, vol. 13, no. 4, art. 217, 2025. doi: [10.3390/ijfs13040217](https://doi.org/10.3390/ijfs13040217)
- [24] M. A. Alkhedairy, "Balancing privacy and risk: a critical analysis of personal data use as governed by Saudi insurance law," *Laws*, vol. 14, no. 4, art. 47, 2025. doi: [10.3390/laws14040047](https://doi.org/10.3390/laws14040047)
- [25] H. O. H. Suliman, "Evaluating the effectiveness of Saudi Arabia's PDPL in the global digital economy," *J. Data Protection & Privacy*, vol. 8, no. 1, pp. 97–111, 2025. doi: [10.69554/elux6976](https://doi.org/10.69554/elux6976)
- [26] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004. doi: [10.2307/25148625](https://doi.org/10.2307/25148625)
- [27] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. doi: [10.2753/mis0742-1222240302](https://doi.org/10.2753/mis0742-1222240302)
- [28] J. Venable, J. Pries-Heje, and R. Baskerville, "FEDS: a framework for evaluation in design science research," *European J. Information Systems*, vol. 25, no. 1, pp. 77–89, 2016. doi: [10.1057/ejis.2014.36](https://doi.org/10.1057/ejis.2014.36)
- [29] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quarterly*, vol. 37, no. 2, pp. 337–355, 2013. doi: [10.25300/misq/2013/37.2.01](https://doi.org/10.25300/misq/2013/37.2.01)